European Parliament

Auditing the quality of datasets used in algorithmic decision-making systems

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 729.541 – July 2022

EN

Auditing the quality of datasets used in algorithmic decision-making systems

Biases are commonly considered one of the most detrimental effects of artificial intelligence (AI) use. The European Union (EU) is therefore generally committed to reducing their incidence as much as possible. However, mitigating biases is not easy, for several reasons. The types of biases in AI-based systems are many and different. Detecting them is a challenging task. Nevertheless, this is achievable if we manage to: i) increase awareness in the scientific community, technology industry, among policy-makers, and the general public; ii) implement AI with explainable components validated with appropriate benchmarks; and iii) incorporate key ethical considerations in AI implementation, ensuring that the systems maximise the wellbeing and health of the entire population.

However, this can hardly be done if our legal frameworks are not well designed for this purpose. Unfortunately, the EU directives on discrimination include some loopholes that hinder the prevention of bias. The specific regulations on data protection could play a key role in solving this issue, by appealing to the concept of 'fairness' and by providing new uses for data protection impact assessment. This option should be complemented by the measures included in the new regulations on data and AI currently under discussion. The proposed AI, data governance and data acts, for example, might become excellent tools in avoiding bias. Indeed, some of their proposed strategies, such as strengthening the mitigation of bias from the first stages of the AI tool development process might become an excellent anticipatory compliance approach. Furthermore, the creation of certificates that could guarantee the standardisation of databases is essential to ensure that AI tools employ adequate datasets. Finally, strengthening AI-system-subject transparency rights could be extremely helpful in finding the source of biased results. However, misalignment between the regulations and with the General Data Protection Regulation should be corrected.

AUTHORS

This study has been written by Iñigo de Miguel Beriain, Pilar Nicolás Jiménez (UPV/EHU), María José Rementería, Davide Cirillo, Atia Cortés, Diego Saby (Barcelona Supercomputing Center), and Guillermo Lazcoz Moratinos (CIBERER - ISCIII) at the request of the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

ACKNOWLEDGEMENTS

We would like to thank Victor Dario Aguilar Méndez for reviewing a preliminary version of this document. Needless to say, all possible mistakes are our responsibility.

ADMINISTRATOR IN CHARGE OF SUPERVISING THE PROJECT

Andrés García Higuera, Scientific Foresight Unit (STOA)

ADMINISTRATOR RESPONSIBLE

Philip Boucher, Scientific Foresight Unit (STOA)

To contact the publisher, please e-mail stoa@ep.europa.eu

LINGUISTIC VERSION

Original: EN

Manuscript completed in July 2022.

DISCLAIMER AND COPYRIGHT

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged and the European Parliament is given prior notice and sent a copy.

Brussels © European Union, 2022.

PE 729.541
ISBN: 978-92-846-9681-9
doi: 10.2861/98930
QA-08-22-267-EN-N

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)

Executive summary

Introduction

Artificial intelligence (AI) is a transformative technology in modern society. It offers superlative opportunities, but also entails substantial risks. One of these is bias, which can produce harmful results for people, including social discrimination and a significant loss of trust from society. However, this should not be considered as a novel and unsolvable problem. The existence of bias predates the creation of AI tools. All human societies are biased – AI only reproduces what we are. Therefore, **opposing this technology for this reason would simply hide discrimination and not prevent it.** Our task must be to use the means at our disposal – which are many – to mitigate biases in AI. It is likely that at some point in the future, recommendations made by an AI mechanism will contain less bias than those made by human beings. Unlike humans, AI can be reviewed and its flaws corrected on a consistent basis. Thus, at the end of the day, AI could eventually serve to build fairer, less biased societies.

There are some good reasons to consider that such a scenario is plausible. Types of bias in AI systems are many and different. Detecting them is a challenging task. Nevertheless**, there are good options to identify, avoid, and mitigate biases**. To do so, it is utterly important to understand where and how biases can be introduced. Discrimination in AI can be prevented by ensuring fairness and trustworthiness throughout all the steps of the lifecycle of AI development. This includes addressing biases when gathering and pre-processing the data, as well as during the stages of model building, training and evaluation, and finally at deployment phase and impact assessment when AI is applied to end-users in real-world settings. Recommendations to achieve this goal are: to increase awareness of all these different types of biases in the scientific community, technology industry, among policy-makers, and the general public; to implement AI with explainable components validated with appropriate benchmarks; and to incorporate key ethical considerations in AI implementation, ensuring that systems maximise the wellbeing and health of the entire population.

Bias: tracing the boundaries

It is important to retain that not all biases should be avoided. We must differentiate between what we call bias from a statistical perspective and what we usually understand as bias from the point of view of the social sciences. In statistics, the concept of bias is aseptic: it only implies that a mechanism always segregates in a particular direction. This, in some cases, is acceptable and even necessary. For instance, in a social support service, it may be necessary to provide assistance to people on a low income and not to others. From a social science perspective, however, the idea of bias (as a prejudice) is associated with discrimination and unfairness and should therefore, in principle, be avoided (affirmative action is often an exception to this rule). Sometimes this might involve an active intervention aimed at introducing corrective measures in a database. Consider, for example, that executive hiring decisions are made in databases that accurately reflect the under-representation of women in this area. In such a context, it will be necessary to bias those databases to obtain fair results.

However, this general statement should be qualified, considering our legal framework on discrimination, which only considers as legally discriminatory those differences that are created on the basis of certain categories (gender, religion, political ideology, etc.). Moreover, these provisions cannot be extended to other biases which, in principle, seem to lead to unfair results, but which are not considered directly discriminatory. Thus, for example, if an AI tool suggests a higher price for the same product to a person who lives in the countryside and not in the city, this does not violate European anti-discrimination laws.

The present study argues that, both to deal with this kind of situation and to resolve some of the loopholes implicit in the discrimination regulations themselves, it is **perfectly possible to resort to**

the specific regulations on data protection. Within this, it is the concept of 'fairness' that must play an essential role. However, this requires the **legislation make an effort to delimit what is understood as fair or unfair in the context of bias**. Obviously, it is also possible that ultimately case law will be left to determine the limits of this concept, but this would mean accepting levels of indeterminacy that are not beneficial to the development of a European data space. Nevertheless, it is necessary to point out that this could only help us to avoid bias if the databases were composed of personal data. Otherwise, the data protection regulations would not apply. This is key when considering anonymised data.

Some tools that might serve well to mitigate bias

To avoid the occurrence of unacceptable biases, an adequate system to assign responsibilities is a key issue. The measures that can achieve this objective, however, come at a cost that can sometimes be disproportionate considering the harm that would be caused by the biases. Hence, a risk-approach such as that proposed in the AI act would be adequate for the issues at stake. The **importance of the measures and whether they are internal or external - through independent third-party audits, for example - will probably depend on the degree of foreseeable risk and harm to individuals.** However, the general preference for self-certification models made in the proposed AI act raises some doubts about the efficiency of the proposal. In any case, this aspect might be reconsidered before the act's final approval, or in the future, if practice demonstrates that it is too risky.

The creation of bias awareness and mitigation tools across the lifecycle of an AI system will depend, in turn, on the implementation of appropriate certification systems, along the lines suggested by the proposed European data strategy, the data act and the AI act. These should not follow the same strategy in the case of datasets and AI mechanisms, for a simple reason: **discrimination and unfairness are always contextual**. It cannot be known whether they are suitable or not if their purpose is unknown. A dataset that is composed of women's data alone will be suitable for building a tool linked to cervical cancer, but not for an equitable distribution of hospital beds, for example. Hence, in the case of datasets, **certificates must be associated with information or metadata about the characteristics of the data they contain,** namely: the data structures, data formats, vocabularies, classification schemes, taxonomies and code lists, which should be described in a publicly available and consistent manner.

In the case of AI tools, on the other hand, it is already possible to delimit their concrete uses (indeed, the AI act proposal classifies the risk of an AI tool on the basis of its concrete goal), so certification of the corresponding AI tool should refer to these purposes. Thus, certification must consider such uses and include the general characteristics, capabilities and limitations of the system, algorithms, data, training, testing and validation processes used, as well as documentation on the relevant risk management system.

What seems certain, however, is that the effectiveness of this system of validation through certification **will depend to a large extent on the creation of standards**, both in terms of the information to be included in a dataset and the types of procedures that will ensure the absence of bias in an AI system.

Policy options

Based on the above, this study proposes four fundamental policy options regarding the mitigation of biases provoked by use of datasets in AI tools:

- Policy option 1: **No new legislation focusing on biases is required.** Multiple proposals have been made for regulatory tools that target data and/or artificial intelligence, such as the NIS2 directive, the cyber-resilience act, the data act, the data governance act, the digital markets act, the digital services act, and the artificial intelligence act (AI act).

Creating bias-focused standards before testing the ability to address the issue through these standards would be premature. It is likely, in fact, that a suitable interpretation of the regulation that already exists, and what is about to be approved, will be sufficient to address the issue of bias. Instead, focus should be on solving the misalignments between the different regulatory tools (especially the GDPR and the new regulations), which may generate legal uncertainties for all bodies concerned. Otherwise, companies may find themselves having to balance between regulations at the intersection of the AI act, GDPR and the proposed general product safety regulation.

- Policy option 2: **Preventive approach – strengthening the mitigation of bias from the first stages of the AI tool development process.** Ex-ante control is crucial to avoid unnecessary harm. It is essential that bias prevention is incorporated into the process of creating an AI mechanism from its earliest stages. This involves ensuring, among other things, that training, validation and testing data sets are sufficiently relevant, representative, error-free and complete in view of the intended purpose of the system.
- Policy option 3: **Database certification – enforcing bias-awareness at the very beginning of the life cycle in algorithmic systems.** The creation of certificates that could serve to guarantee the standardisation of databases is essential to ensure that AI tools employ adequate datasets. It is doubtful, however, whether these certifications should be compulsory. Nevertheless, in the case of datasets that feed high-risk AI tools, it seems imprudent for the developer to use datasets that do not have certification that guarantees the quality of the information provided on the data they include.
- Policy option 4: **Granting transparency rights to AI-system-subjects – opening a window to find the source of biased results.** This policy option envisages granting AI-system-subjects transparency rights to the databases on which AI systems that make decisions affecting the subjects are developed. Unfortunately, individual rights for AI-system-subjects are not currently included in the proposed artificial intelligence act. This should be modified if adequate compliance is to be ensured with non-bias policies.
- Policy option 5: **Facilitating companies' implementation of the proposed AI act.** Compliance with the regulations that are in the process of being approved will entail costs and risks for companies, especially for small and medium-sized enterprises (SMEs). The draft AI act proposes regulatory sandboxes and specific measures to support small-scale users and providers of high-risk AI systems in complying with the new rules. However, these might be insufficient and could be complemented with additional measures. For instance, public institutions should make high-quality databases available to private agents. This would substantially reduce the expenditure associated with the review of these datasets and the corresponding certification. These initiatives could, of course, be complemented by specific subsidies aimed at helping companies to adapt to the new regulatory framework.

Table of contents

List of figures

List of tables

1. Introduction

Artificial intelligence (AI) is a transformative technology in our modern society, which has inherent risks. One of these risks is biases, which can produce harmful results for people, including social discrimination and a significant loss of trust from society, if they are not identified and mitigated. Identifying and mitigating biases in the AI-based system development process is a critical element of trust that is currently underdeveloped. However, it is important to retain that the existence of bias in AI is not an isolated event, specific to this technology. Human societies tend to embrace traditionally discriminatory schemes. The development of AI could even be considered a splendid opportunity to understand these biases and combat them in all areas of society. The development and implementation of AI mechanisms should help us to properly identify the biases we create through our social interactions, and the nature of AI should allow us to correct them much more efficiently than when biases are produced by humans. For example, it is easier to detect and erase racist practices in a machine that advises on job selection than it is to do so with thousands of people engaged in similar tasks, with sometimes very different opinions and behaviour.

Keeping this in mind, the main objective of this study is to understand the biases that occur in AI-based solutions in different fields of application and to propose policy options that serve to mitigate them. However, this is not at all simple. First, because the concept of bias is quite blurred. From a statistical perspective, the International Organization for Standardization (ISO) defines bias as 'the degree to which a reference value deviates from the truth'.[1] This deviation can contribute to harmful or discriminatory results (e.g., amplification of prejudice and stigmatisation), but can also be beneficial in specific contexts (e.g., affirmative action policies and precision medicine practices). However, from a social perspective, bias is generally considered to have a detrimental or discriminatory impact. Indeed, it is usually considered that biases go against an essential value: 'fairness'. However, fairness is, again, a vague concept, which has many different definitions. For instance, its concrete meaning in the context of the General Data Protection Regulation (GDPR) is yet to be clarified. However, it seems pretty clear that, whatever definition we adhere to, to achieve fairness in an AI-based system we need to understand how the system can be biased, and where and how bias is generated. This will always be the only way to mitigate and/or control it. Unfortunately, current regulations do not provide a precise definition of fairness.

To these preliminary issues, the complexities involved in a legal framework that is undergoing a period of great change must be added. There are multiple proposals for regulatory tools that target data and/or AI, such as the NIS2 directive, the cyber resiliency act, the data act, the data governance act, the digital markets act, the digital services act, or the artificial intelligence act (AI act). Most of them are expected to become law in the near future. They will all need to be aligned with the GDPR. Analysing the issue of bias in these circumstances is very problematic, because it is likely that some of the current drafts will contain very different wordings in their final versions. Keeping this in mind, this study tries, as far as possible, not to stick too closely to the drafts of the proposals, although they are certainly quoted when their texts reflect some of the ideas that seem to us to be reasonable.

This study is divided into three main parts. In the first, we provide an overview of biases in the context of AI, and more specifically to machine-learning applications (ML). This is the subject of the first part of the study, prepared by the Barcelona Supercomputing Center (BSC) team through a non-exhaustive review of the most recent literature on this topic. It also includes a precise analysis of the different approaches proposed to mitigate biases, explaining their advantages and weaknesses. The second part of the study is devoted to the analysis of biases from a legal point of view. To do so, the current EU legislation on discrimination is first outlined. The study shows that shortcomings in this area call for the implementation of other regulatory tools to adequately address the issue of bias. In

[1] International Organization for Standardization, see: https://www.iso.org/standard/40145.html

our view, a proper implementation of the relevant data protection legislation could serve this purpose well. However, it seems necessary to us to complement the GDPR by introducing some of the measures that are already included in the new proposals, such as the creation of standards or certifications. Standards are particularly necessary, because they can provide an order – a common language that helps promote technological development in Europe. Indeed, creating European standards could help to promote the 'European way of doing things' in terms of reconciling development of efficient AI systems and respect for human rights.

Last, but not least, it is essential to highlight that this study is 1) mainly focused on AI systems based on machine learning that use unambiguous and well-defined datasets, hence excluding those using volatile and ever-changing sources of information (e.g., internet); 2) essentially focused on datasets and the role they play in the emergence of biases. Therefore, although we analyse the occurrence of biases in the whole chain of creation and implementation of AI mechanisms, the focus is mainly on the use of data for such purposes. This is for several reasons:

- According to the European Commission's Communication on AI,[2] 'Artificial intelligence refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals'.
- Hence, AI-based systems should have the ability to perceive the environment, collect and interpret data, reason on the information that has been generated, decide which is the best action, act using some actuators and, finally, possibly modify the environment with this action.[3]
- Machine learning is a subset of AI that aims to give computers the ability to learn without being explicitly programmed (Samuel, 1959). Algorithms are trained for classification or prediction purposes based on some input data.
- The quality of the outputs produced by ML-based algorithms, and therefore the accuracy and the patterns obtained, is strongly dependent on the quality of the datasets used for training. Thus, it is fundamental to guarantee that the data collected is the right one for the problem that needs to be solved and the new insights we will obtain from it.

[2] 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 on Artificial Intelligence for Europe, Brussels, 25.4.2018 COM(2018) 237 final.

[3] High-Level Expert Group. A definition of AI: main capabilities and disciplines (2019).

2. Methodology and results used

The preparation of this study has been particularly complex for several reasons. First, a topic such as biases in AI tools and datasets can hardly be addressed from only a legal point of view. This obstacle has been successfully faced by adopting a multidisciplinary approach. This study has been produced by three teams with different but clearly complementary expertise. Teams from the UPV/EHU and the Centro de Investigación Biomédica en Red (CIBERER – ISCIII) are mainly composed of lawyers and philosophers, with extensive experience in the field of personal data and AI Act. The Barcelona Supercomputing Center team works on technical analysis of bias in the context of AI. Thanks to our interaction, we have been able to produce a study that integrates a complete multidisciplinary analysis of the issues at stake. This allowed us to ensure that both the diagnoses and the policy options designed are reasonable from the perspective of their practical implementation.

The second difficulty has to do with the fact that some of the regulations concerning the use of data for the construction of AI mechanisms have not been passed yet. Hence, different versions have been released over the course of the period of our work and others are yet to come. This has forced us to make an extraordinary effort to adapt to this situation. As a result, we have analysed a panoramic legal framework that includes:

- the Directives focused on discrimination issues.
- the General Data Protection Regulation (GDPR).
- the Digital Services Act[4] and the Digital Markets Act[5].
- the AI Act[6], the Data Governance Act[7] and the Data Act[8].

Thus, our methodology comprised different tools:

- First, it was based logically on an exhaustive review of the updated available literature on biases in Artificial Intelligence (AI) and the applicable regulations to that effect. Most sources used in our analysis (especially those related to technical aspects) have been produced in the last three years.
- Second, it comprised an extensive analysis of the regulations and jurisprudence applicable to the issues at stake. This included a careful analysis of a lot of information produced from EU institutions, think tanks, national data protection institutions, the European Data Protection Board (EDPB) and the European Data Protection Supervisor (EDPS) etc. on these new legal tools.
- Third, our analysis has benefited greatly from the organisation of a series of seminars and debates organised in the context of the Panelfit project[9], led by the UPV/EHU team. These included the following:

[4] Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC.
[5] Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act).
[6] Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data.
[7] Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act)
[8] Data Act: Proposal for a Regulation on harmonised rules on fair access to and use of data
[9] Panelfit (Participatory Approaches for a new Ethical and Legal Regulatory Framework on ICT) is an EU Commission funded project aimed at facilitating the implementation of the new regulation on data protection (www.Panelfit.eu)

- An event held in Madrid, with the participation of four experts and stakeholders: José Luis Piñar, Former Director of the Spanish Agency for Data Protection (AEPD). Director of the South EU Google Data Governance Chair; Elena Gil, Doctor of Law, Attorney; Richard Benjamins, Chief AI & Data Strategist in Telefónica, co-founder of the Observatory for Social and Ethical Impact of Artificial Intelligence, ODISEIA, Madrid; Jorge Juan Ramos, Software developer, Madrid.
- An event held in Vienna, including an extremely timely conference by Christiane Wendehorst, Scientific Director of the European Law Institute (From the GDPR to the AI Act) and a Panel Discussion about the AI Act (Panel members: Charles Raab (University of Edinburgh), David Reichel (Fundamental Rights Agency), Lorena Jaume-Palasí (The Ethical Tech Society), Johann Čas (Austrian Academy of Sciences).

An online debate on biases and transparency. Participants: Dr. Gemma Galdón Clavell, CEO and Founder of Eticas. Algorithmic Auditor, Ethics Oversight, Horizon Europe Projects; Matthias Spielkamp, founder and Executive Director at Algorithm Watch; Prof. Dr. Dirk Lanzerath, Executive Manager of the German Reference Center for Ethics in the Life Sciences (DRZE) and Secretary General of the European Network of Research Ethics Committees (EUREC)

Fourth, our analysis has benefited from an active exchange of ideas of some academics and stakeholders with expertise on ethical, legal and technical issues regarding bias, such as: Dr. Alba Soriano Arnaz (University of Valencia); Aurelie Pols, Data Protection Officer, European Center for Privacy & Cybersecurity (ECPC); Jessica Wulf (Algorithm Watch, https://algorithmwatch.org/en/team/jessica-wulf/); Richard Benjamins, Chief AI & Data Strategist in Telefónica, co-funder of the Observatory for Social and Ethical Impact of Artificial Intelligence, ODISEIA, Madrid; Prof. Dr. Fruzsina Mollnar Gabor (European Group of Ethics, EGE)).

3. Synthesis of research work and findings

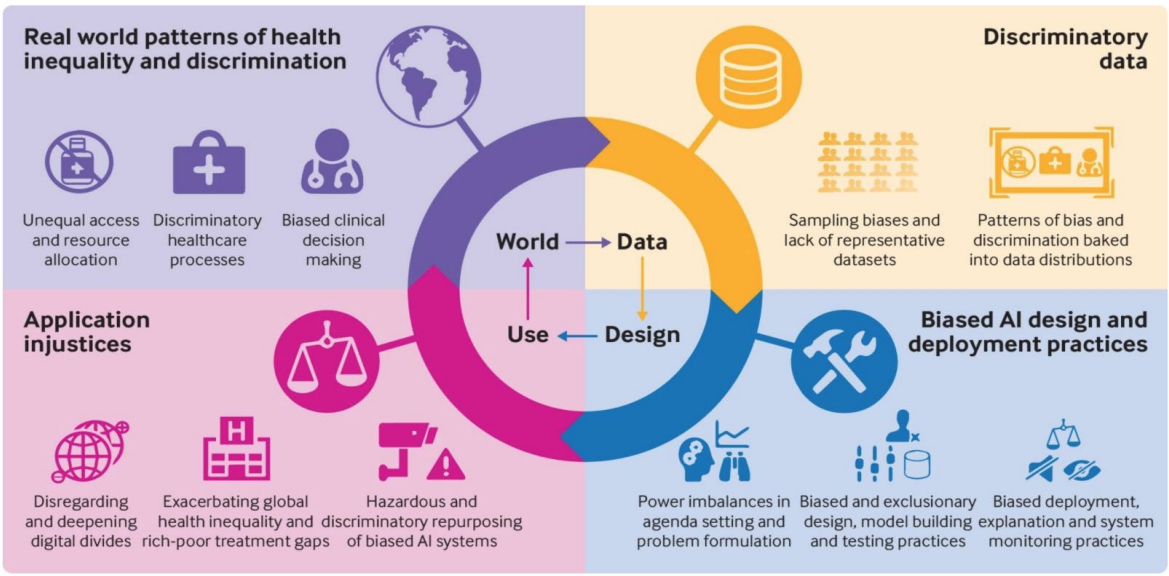
3.1. Technological analysis of biases

This first part provides an analysis of the technical aspects related to biases, offering a non-exhaustive review of different AI-based applications, and in particular of the training datasets used in Machine Learning to understand the associated ethical and social concerns. Next, a tentative taxonomy of biases is introduced along with a study of identification of such biases in the AI development processes. Finally, a set of recommendations and strategies to tackle bias mitigation is offered, as well as a collection of initiatives, tools and organisations that are currently working to put such recommendations into practice.

3.1.1. A biased representation of knowledge

Artificial Intelligence (AI) is being adopted across all kinds of industrial sectors in Europe and in different types of applications. Its use has been widely proven efficient in many scenarios, by saving time and cost as well as by augmenting and complementing human skills. However, we still face plenty of cases where AI-based solutions, and in particular machine learning applications (ML), cause harm in the form of discrimination against specific populations, underrepresented in training datasets or victims of other sorts of biases present in the AI system life cycle, as defined in the following subsections. We currently live in an algorithmic society where citizens have become data donors and they are not always aware of this. Hence, special attention needs to be paid to which data is being collected, how, for how long and for what purpose. In other words, it is important to consider **the content of the data and to what extent the information fits the purpose** (European Union Agency for Fundamental Rights, 2019). This will ensure higher quality in the datasets used to train AI-based systems, as it will enhance accuracy, validity, reliability and social inclusion. An essential step in designing AI-based solutions relies on the way this 'real-world' information is represented for a computer system to understand it, and learn from that knowledge to solve problems (Figure 1). **If knowledge representation is not complete (i.e. socially inclusive), then the AI-based system's reasoning process will be biased, and therefore provide unfair outcomes.**

Figure 1 – Inequality and discrimination in the design and use of AI in healthcare applications

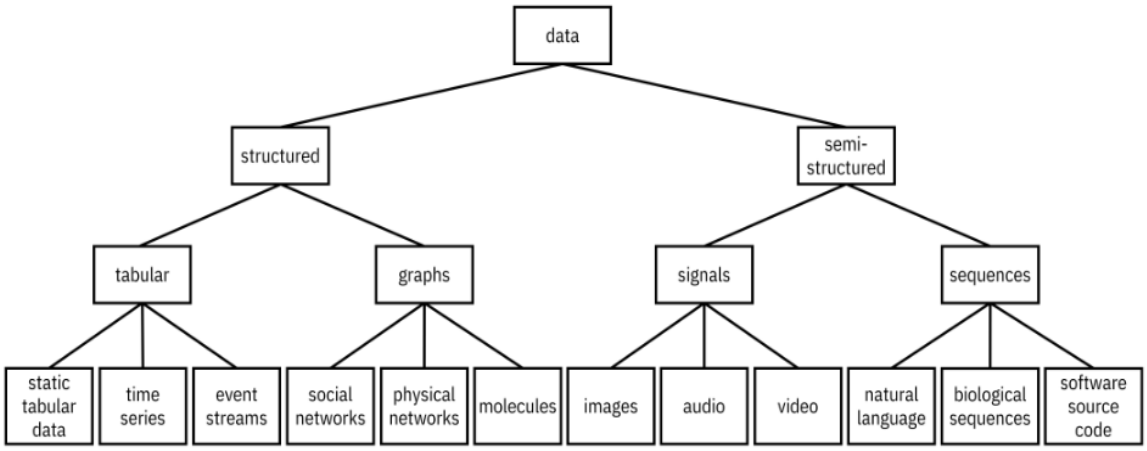


Source: Leslie et al.

The possibility of obtaining biased AI outcomes is strongly related to the characteristics of the data and the quality of the data management process, including data gathering, cleaning, annotation and processing. The technical, ethical and social challenges associated vary with the kind of data (Figure 2) and AI/ML technique used. For instance, structured data works under predefined categories of data, generally text-only, which is usually organised into relational databases. Examples of structured data include dates, contact information, lab values, demographic data and financial information. In this case, the challenge relies on the high sensitivity of consuming applications with respect to data quality[10] . This is related to any possible error in the knowledge representation approach (e.g., enough attributes, possible sparseness or heterogeneity of data, underrepresentation of social groups, etc.). These factors will affect the quality of the AI/ML-based system, hence its performance in terms of accuracy, reliability and trustworthiness in general.

In the case of **unstructured data** (e.g., images, videos, plain texts) and semi-structured data (e.g., HTML, XML, JSON files), an internal structure is not predefined in data models or schemas. Unstructured and semi-structured data can be generated by humans as is the case of images, videos, social media data and websites, media (MP3, audios, digital photos, etc.), text files, but it can also be machine-generated, generally through sensors and cameras and applied in fields like digital surveillance, traffic, weather, and space exploration amongst others. The lack of internal structure requires human intervention to organise these data through pre-processing techniques like data annotation. In this case, **knowledge representation is strongly affected by human judgement, which is historically and socially biased.** Moreover, when it comes to classifying sensitive information such as gender or race, there are no objective visual characteristics or consensus among the AI/ML community[11] on how to properly label these data. Hereafter we introduce a selection of examples of AI-based applications using unstructured data that suffered from undesirable biases, namely images and texts.

Figure 2 – Categories of data types with structured and semi-structured data formats



Source: Varshney, 2022.

For example, social network data can be stored as a graph that is a structured data format; multimedia, such as videos, can be stored as semi-structured formats that contain both information with no structure (video frames) and its meta tags (location, date, etc.).

[10] https://towardsdatascience.com/ai-should-not-leave-structured-data-behind-33474f9cd07a
[11] https://thegoodai.co/2021/04/14/gender-and-racial-bias-in-computer-vision/

Over the last years, computer vision has become a major driver of AI. It is already used in industries like manufacturing, agriculture, and the automotive or medical industries amongst others, with applications that are enhancing human skills and boosting the economy. However, its use has also raised several ethical concerns regarding bias and its negative impact on certain social groups. Computer vision is based on supervised learning techniques, generally (deep) neural networks, and is used for image segmentation, classification or recognition. They require large training datasets, which are often difficult to generate due to their cost. In 2010 ImageNet (Srinivasan and Chander, 2021) was released, containing millions of tagged images and contributing to the rapid growth and adoption of computer vision. However, **evidence in past years showed that a lack of oversight on the data management processing was leading to serious bias problems.** In particular, **there was no control on where the data came from, who was labelling it, nor how it was being labelled.** This issue was particularly delicate in the Persons' category, where the taxonomy proved to be non-inclusive with the LGTBIQ+ community for example. If we move to the level of categories, we can easily observe that these are often misogynist, racist, creating and perpetuating discriminating against women, BIPOC[12], disabled people, etc[13]. This practice has been replicated for the last decade in big tech companies where it is harder to check how these datasets are being created and ordered. However, several cases have raised awareness towards this issue, turning the attention towards the bases upon which the technology has been built and how it has been used.

For instance, in 2015 a user reported on social media that Google Photos was tagging him and his African-American friends as 'Gorillas' (Figure 3A). The main problem was two-fold: on the one hand, there was a lack of images of people of colour; on the other, there was no control on the labels associated with images of persons. In an attempt to solve the problem, Google censored some types of apes from its tagging system. Nonetheless, the issue of misrepresentation of BIPOC was still present a few years later[14]. Moreover, AlgorithmWatch[15] showed how Google's Vision AI was assigning the tag 'gun' to an image of a dark-skinned hand holding a thermometer while a similar image with a light-skinned hand was tagged as 'electronic device'. Buolamwini and Gebru's (2019) research exposed skin type and gender bias in commercial AI products, particularly facial recognition systems of big companies such as Amazon, IBM or Microsoft among others. As a result, the Gender Shades project[16] evaluated the accuracy of AI gender classification tools and, in addition, introduced an intersectional approach showing that these systems were performing particularly poorly for dark-skinned women. Another research (Scheuerman et al., 2019) showed how facial analysis technologies were consistently performing worse on transgender people and were unable to classify non-binary genders. The impact of the systematic replication and perpetuation of these biases is amplified when used by third-party applications, causing harm to the underrepresented communities, from denying access to healthcare, to being stigmatised by law-enforcement or surveillance systems, among others.

Another area of Artificial Intelligence that is quickly growing is **Natural Language Processing** (NLP), achieving high performance on multiple language-understanding tasks. NLP is used for speech recognition (e.g. virtual assistants like Siri or Alexa), automated translation (e.g. Google translate) and other applications that are already in many of our everyday lives such as email filtering, chatbots or predictive text. However, NLP models are often susceptible to learning from implicit bias in the data that replicate social stereotypes. A clear example of this behaviour is the case of Google Translate, which used to assign gender roles when working with gender-specific languages. Hence, when translating the sentences 'He/she is a doctor' and 'He/she is a nurse', from a gender-neutral

[12] Black, Indigenous and People of Colour
[13] https://excavating.ai
[14] https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/
[15] https://algorithmwatch.org/en/google-vision-racism/
[16] http://gendershades.org/overview.html

language like Turkish to a gender-specific one like English, it used the male version for the former sentence and the female for the latter. Google is working to reduce gender bias in its products, and since 2018 offers translations in male and female forms (Figure 3C). However, a study (Prates et al., 2019) showed that the prevalence of translating sentences containing job positions from neutral gender languages to gender-specific ones is higher in males, especially in scientific and engineering fields. Hence, Google translator is replicating stereotypes that exist in the real-world job distribution of female workers. NLP techniques such as word embedding can be used to build knowledge upon which the algorithm learns, however they are sensitive to producing the above mentioned situations due to a bad representation of the domain.

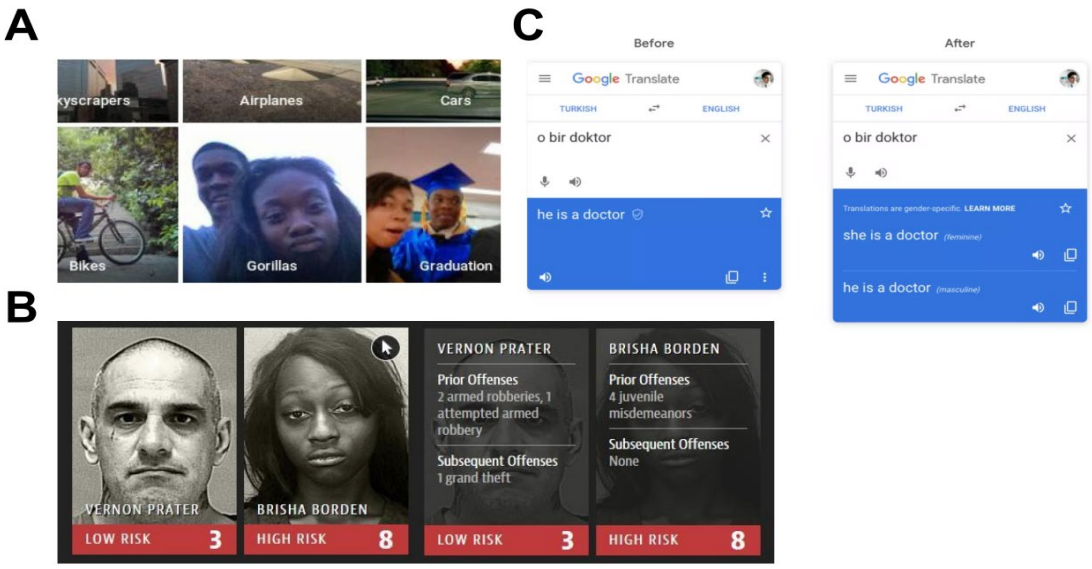
The gender bias issue in NLP starts from training large datasets that are unbalanced, but the bias is also amplified by the learning algorithm (Costa-Jussà, 2019). In 2016, Microsoft released the chatbot Tay designed to interact with users via Twitter. The objective was to improve its natural conversation skills through these interactions, but the company had to cancel it in less than a day. Tay was tricked by the social media users and started to replicate misogynistic, homophobic and racist behaviour. Similarly to Tay, OpenAI's GPT-3, Google's Meena or Facebook's Blender are designed to mimic human language. They are trained with vast amounts of data coming from the Internet, including unwanted prejudice and toxic language[17], reproducing hate speech, misogynistic and homophobic language and racist rants. This could be prevented if the algorithm was designed with some level of censorship or control towards sensitive topics, implementing governance mechanisms such as human oversight or the AI system policing itself. However, it is difficult to fully automate this task or contemplate any risky situation from the beginning and this process would need continuous review.

In addition to the bias-related issues mentioned above, it is also necessary to check the negative impact that NLP systems that advise or provide information to people, such as conversational agents, can produce. Miner (Miner et al. 2016)2016 analyses the effect on people of 4 common conversational agents: Siri, Google Now, Cortana and S Voice, asking them questions about mental health, interpersonal violence and physical health, concluding that they respond inconsistently and incompletely to the questions. This example shows how AI-based systems with which we can interact casually that can be considered simple can cause significant harm to people if ethical considerations are not appropriately addressed.

Over the last few years we have seen how, if not used correctly, Artificial Intelligence can amplify socially-constructed biases, replicating stereotypes and bad practices that produce discrimination against certain vulnerable or underrepresented groups. It is fundamental to include a multidisciplinary perspective in the design of new technologies that take into account the benefits of deploying such technologies and how these will impact on society. In this sense, in a recent paper[18] from DeepMind, authors establish and analyse in detail the large-scale Language Models potential risks, drawing on multidisciplinary literature from computer science, linguistics, and social sciences. An important step in this process is to ensure a high quality data management process in order to avoid situations as the ones presented in this section. Instead, if training datasets are created and processed following an inclusive perspective, AI/ML-based solutions could then help reduce social inequalities.

[17] https://www.technologyreview.com/2020/10/23/1011116/chatbot-gpt3-openai-facebook-google-safety-fix-racist-sexist-language-ai/
[18] https://deepmind.com/research/publications/2021/ethical-and-social-risks-of-harm-from-language-models

Figure 3 – Notable examples of real-world AI bias



A: Race bias in Google Photos (described in Section 5.1.1). B: Race bias in COMPAS algorithm (described in Section 5.1.2). C: Gender bias in Google Translate from Turkish to English and subsequent fix (described in Section arxiv 5.1.1). Sources adapted from (A) Twitter @jackyalcine; (B) Google The Keyword, James Kuczmarski[19]; (C) ProPublica)[20].

Main points

- Over the last few years, the AI community has taken a turn towards ethically responsible practices of design, development and use of the technology.
- Data management has become one of the principal requirements to align with Trustworthy AI, European values and fundamental rights. In particular, to ensure the quality of the data is to enable social inclusiveness and thus provide a better representation of the world. Nonetheless, for those AI-based applications with direct impact on human beings, it will be necessary to implement further governance mechanisms including human oversight.
- Knowledge representation is the first step in the process of generating fair training datasets and algorithms. Some aspects of this process could be policed or standardised, while others will require a commitment of the AI community towards ethical, legal, socio-economic and cultural aspects of AI.

3.1.2. Taxonomy of biases

[19] https://blog.google/products/translate/reducing-gender-bias-google-translate/

[20] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

- **Pre-existing bias** refers to a bias that already exists before the construction of the computer system. It is a bias that comes from the developers, society, or subculture. It can be introduced at every stage of the development of a model. From defining the goals of the model, collecting the data, building and training the model, to the user interaction with the model. For example, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software, which measures the risk of a person committing another crime, has been found to be negatively biased towards Afro-American people[21] (Figure 3B). This bias is entangled in the sociocultural context in which the system has been created, and was subsequently introduced into this software.

- **Technical bias** is introduced during the development of computer systems. It can be introduced from the data collection to the model training, testing and deployment. It is part of the technical aspects of the construction of the system. Examples of technical bias include unnoticed imbalance and underrepresentation of protected attributes (e.g. age, gender and race) in the data used in the model, or a misleading visualisation of the data or the results to the end users. Technical bias is the most common type of algorithmic bias. Dedicated guidelines and best practices can help avoid this type of bias. An example of bias introduced during the data sampling process happened at the beginning of the COVID-19 pandemic, when countries reported the Case Fatality Rates (CFRs) using different sampling strategies, and hence created confusing statistics. Some countries introduced bias in their CFRs measures by only sampling from seriously ill persons while other countries measured their CFRs by sampling from a wider range of the infected population (Ward, 2020).

- **Emerging bias** occurs after the system has been built. It happens when the users that interact with the system, or the data employed to build the system, become unrepresentative of its intended use anymore. An example is the app StreetBump[22] that reports potholes in the city. After its release, it was found that the users of the app were located in the wealthiest parts of the city, therefore discriminating against people with lower income[23]. One way to avoid this type of bias is to keep an updated dataset to represent different and unexpected situations.

Several alternative categorizations of bias are present in the literature. For example, the web page (Catalog of bias[24]) is dedicated to listing and defining possible types of biases. Just as Friedman and Nissenbaum (1996) created categories of bias based on the moment when this occurred in AI development, Baeza-Yates defines categories based on the origin of the bias: data, algorithms or user interaction (Baeza-Yates, 2018). Specifically, this categorization refers to three types of bias:

- **Cognitive bias**: Systematic pattern of deviation from norm or rationality in judgement. Shana Lebowitz in business insider[25] describes the 20 cognitive biases that screw up your decisions.

- **Statistical bias**: Systematic deviation from a reference value.

- **Cultural bias**: Interpretation and judgement phenomena acquired throughout our life. Even though we generally want to avoid and remove bias from AI systems there are some cases where this would not apply. Stemming from this, we can define some different attributes that bias can have.

[21] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[22] https://www.boston.gov/transportation/street-bump
[23] https://hbr.org/2013/04/the-hidden-biases-in-big-data
[24] https://catalogofbias.org/
[25] https://www.businessinsider.com/cognitive-biases-that-affect-decisions-2015-8

• **Implicit/Explicit**: Whether or not the bias has been introduced to the system on purpose and with intention.

• **Desirable/Undesirable**: Whether or not the bias is desired because it fulfils a specific objective.

• **Expected/Unexpected**: Whether or not the bias in the system was foreseen.

Whether a set of attributes can be ascribed to a bias depends on the context. For example, data collected in real-time from social media could contain biases that are implicit, undesirable and unexpected. Conversely, the prospective collection of medical information with historically low female representation ought to be 'biased' on purpose in order to avoid further imbalance in the data (Cirillo et al., 2020). Among the aforementioned attributes of bias, **the most problematic ones are those of unexpected, implicit and undesired biases as they are difficult to find and to fix.**

An additional layer of complexity is represented by real-world scenarios in which controlling for bias is complicated by the fact that **the dataset does not contain any explicit sensitive variables**, which are purposely excluded because of law proscription. One solution to test against unwanted discrimination in this scenario is to always report such variables and check the outcome of the algorithm against it before excluding any of them. Moreover, possible strategies to obtain sensitive variables to test against unwanted discrimination include submitting an informed consent, conducting a survey, inferring this information using publicly available data sources and/or organizations' internal data, and use proxy variables (Benjamins, 2019).

Table 1 shows some of the most common biases in developing AI systems. Mainly, algorithmic biases are reported.

Table 1 – Illustrative examples of bias types found in the literature

Type of bias	Definition	When it originates
Cultural bias	*Bias in history and society projected by the data and the developers onto the system.*	Model definition Model validation Model deployment
Measurement bias	*Results from the way we choose, utilise and measure a feature.*	Data design
Simpson's Paradox	*Results from an aggregation of subgroups that differs from the separation of those subgroups*	Data design
Population Bias	*Results from a difference of representation between the dataset and the target population*	Data collection
Sampling bias	*Results from a non-random sampling from subgroups*	Data collection
Labelling bias	*Mislabelling introduced in the training data. Often subject to human judgement.*	Data collection
Omission bias[26]	*Result of excluding data features considered*	Data collection & Pre-

[26] https://www.topbots.com/eliminating-ai-bias/?vgo_ee=fHiAYjypwCJourhb1nl8F%2B6GVdc0pWK%2BWh1FRA8O8OI%3D#eliminating-ai-bias

	irrelevant	processing
Excluding bias	*Results when the data scientist removes one or more features in the model*	Model development
Confusion bias	*Results when the model erroneously correlates variables*	Model development
Model bias	*Results when using an insufficient, or unbalanced, or unfair, or discriminatory, or toxic data model*	Model development
Statistical bias	*Result of the statistical correlation of the model*	Model development
Confirmation bias	*The model result confirms the developer's beliefs*	Model validation
Cause-effect bias	*Model correlation errors*	Model validation
Funding bias	*The sponsors of the project* *the results according to financial issues*	Model deployment
Temporal bias	*Results from differences in population and behaviours over time.*	Data maintenance

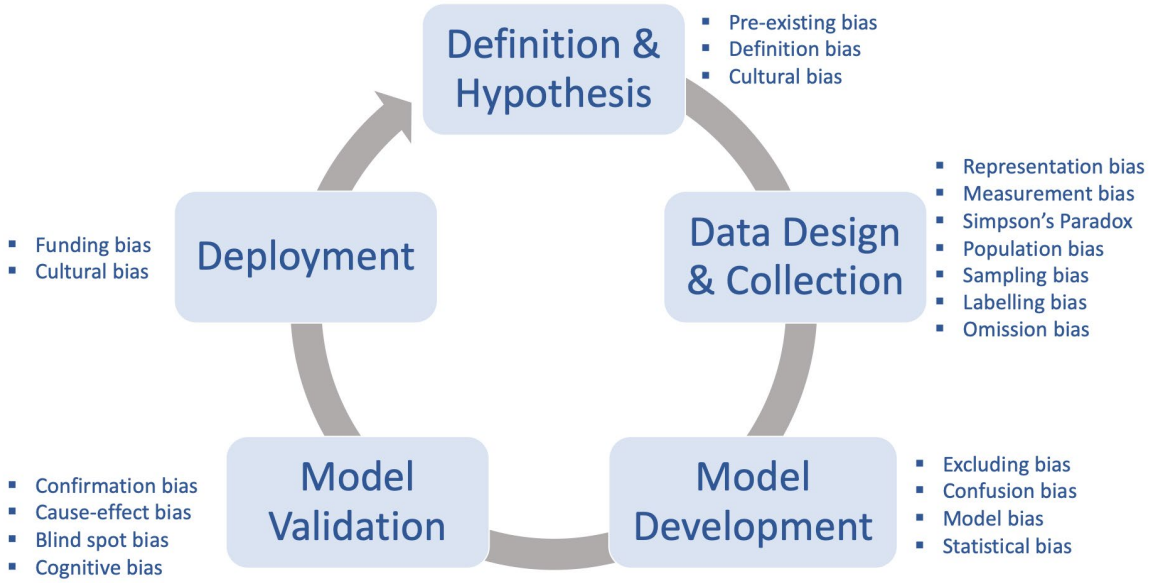
Main points

- There are many types of bias and different categorizations. Bias in AI is a challenging problem where an exhaustive analysis would be impossible. There are different classifications of bias depending on the perspective considered.
- It is important to understand the origin of bias and its nature in order to control and tackle undesired biases and avoid major consequences.
- The most problematic bias is the unexpected, implicit and undesirable type.

3.1.3. Bias occurrence in the AI development process

Understanding what and where biases can appear in the development process of an AI system is necessary to prevent biases from persisting and intensifying. Biases appear when the data set is insufficient, incomplete, or unbalanced, **but they are also inferred by correlation with other factors.** For example, in resume analysis during employee recruitment, even if gender or race are not explicitly reported in the training data set, the algorithm can infer this information from other variables, such as the neighbourhood area, level of studies, practised sports and lifestyle. In general, all this information may correlate strongly with gender or race. Figure 5 shows the main stages in the development process of an AI system and its most common biases.

Figure 4 – Emergence of bias in the AI development life cycle



Source: authors.

Problem Definition and Hypothesis: In this phase of the process, stakeholders and developers define the desired goals of the AI-based solution and project their own personal and cultural biases into the system definition. These biases exist prior to the construction of the solution and are part of the socio-cultural context of the system.

Biases at this stage might correspond to AI solutions designed, for instance, to classify people by characteristics, such as marketing applications that identify population groups based on particular features such as gender, socioeconomic status, political orientation, and religion.

An example is the binary definition of gender as only male or female. This definition excludes or mislabels data from non-binary people. Other examples of biases seeping in at the problem definition stage can be found in AI facial scanning apps (Dolgin, 2019).

Data Collection and Pre-processing: In this process step, data are collected, cleaned, and prepared to be used by the model. If developers are not aware of it, they can introduce bias in the definition and modelling of the training dataset: in the labelling of data (labelling bias), the selection of the training dataset (sampling bias), the aggregation of different existing datasets (Simpson's Paradox), in the exclusion of features (excluding bias), etc.

Examples of bias at this stage can be found in the AI-based solution for breast cancer developed by researchers at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) and Massachusetts General Hospital (MGH)[27]. This AI system predicts whether a patient will develop breast cancer from a mammogram image based on deep learning. Developers used 60,000 mammograms, of which 95% belonged to white women, despite the fact that black women are 43% more likely to develop breast cancer. Although the authors indicate that their system is valid for either black or white women, the review carried out by M. Stewart[28] showed worse results in black women and other underrepresented communities.

[27] https://www.csail.mit.edu/news/using-ai-predict-breast-cancer-and-personalize-care
[28] https://www.topbots.com/programming-fairness-in-algorithms/

Model Development: In this step, the model learns a knowledge representation of the problem to be solved based on the training dataset. When the training data is unbalanced, unfair, discriminatory, or toxic (model bias), the training process results in models that reflect these undesirable aspects. As a result, an AI system that performs well concerning its optimization goal may perform poorly concerning social damages. Such a system encodes and perpetuates harmful stereotypes and biases present in the training data.

Other reasons that can also introduce bias include the removal of features from the model that the developers subsequently consider irrelevant (excluding bias) or when the algorithm incorrectly correlates the data (confounding bias).

Amazon's recruiting engine is a real-world example. The model ranks the applicant's suitability based on the CV submitted, but it was found to be biased against women. The developers tried to solve the problem despite the insufficient data available to train or fit the model to reduce gender bias. Eventually, the problem of de-biasing this algorithm was declared unsolvable and the project was stopped (Dastin, 2018).

Model Validation. In this step, developers test and validate the AI-based solution and fine-tune the model. At this stage, cognitive and cultural biases originating from the developers themselves can enter the model. Confirmation bias occurs when a model's results support the developer's ideas or beliefs; this can lead the developer to stop fine-tuning the model because it is considered correct. Confirmation bias can also occur when the developer discards the results that do not support their ideas or beliefs. On the other hand, blind-spot bias can happen when the developer does not detect biased results. In both cases, developers will not correct the bias, and this will therefore persist, and may even become amplified by the system.

Examples of bias at this stage are illustrated by some facial recognition systems developed from IBM, Microsoft and Face++. The Gender Shades project [29] evaluated the accuracy of three commercial AI-based gender classifiers from facial images. Gender Shades created a test data set, called the Pilot Parliaments Benchmark, to demonstrate that those systems did not balance the facial recognition capabilities across genders and skin tones. Following the Gender Shades report, IBM worked to substantially increase the accuracy of its new recognition system for facial analysis, which now uses different training data and different recognition capabilities than the service Gender Shades evaluated in April 2017 [30].

Model Deployment: In this phase, the responses of the AI-based solution are reviewed by stakeholders and final users to validate them and provide feedback. It is at this point, where human cultural and cognitive biases are included.

It may happen that the response is more expensive than expected and the stakeholders suggest modifications that lower the cost (funding bias), or that the response of the system is contrary to cultural ideas and modifications are suggested (cultural bias). Other examples of bias at this stage occur when the response is biased but confirms the user's ideas (confirmation bias), or the user focuses on surviving and does not validate the correctness of the solution (survival bias), etc.

One example is Delphi [31] (Jiang et al., 2021), which is a research prototype that responds with moral reasoning about situations and actions. Delphi was trained on a large-scale collection of 1.7 million ethical demands from COMMONSENSE NORM BANK. The system responds with moral judgments in a large number of situations, including those that are socially sensitive. However, the system allows users to confirm reasoning and doing so introduces numerous cultural and cognitive biases. After a

[29] http://gendershades.org/
[30] http://gendershades.org/docs/ibm.pdf
[31] https://delphi.allenai.org/

short period of use, the model responds with advice that could be potentially offensive, problematic or harmful (Jiang et al., 2021).

Another example occurred in 2016 when Microsoft implemented a Twitter chatbot designed to interact with young people, but soon after its release the AI started to swear and make racist comments[32]. This was because the AI was learning from its interactions and people started to teach it to swear and incite some racist reactions.

Although there is not a single, universal solution to remove all bias from an AI system, high-profile consulting firms **develop best practice recommendations**[33] to minimise bias in AI-based solutions:

1. **Analyse the points where there is a risk** of prejudice or discrimination:
    - Verify that the size and metadata of the training dataset are adequate, neither excessive nor insufficient. The training data set must be representative and balanced.
    - Perform statistical analysis of the training dataset, including subpopulation analysis. Calculate model metrics for specific groups in the dataset, which can help determine if the model performance is identical for all subpopulations.
    - Pay attention to self-learning solutions. Monitor the results over time, since biases may appear as the models learn independently.
2. **Establish strategies to identify and mitigate biases** from the technical, operational and organisational points of view:
    - Technical strategy. Identify and involve tools and methodologies to identify possible sources of bias. Evaluate the effect of removing this data on the precision of the AI-based solution.
    - Operational strategy. These strategies focus on improving data collection processes. Use internal and external audit groups.
    - Organisational strategy. Establish work and visualisation environments to verify the different metrics and results of the model transparently.
3. **Identify significant use cases**, comparing system results with human results.
4. Improve the process by **including the 'human in the loop' approach in the construction** and evaluation of the model.
5. Have a diverse approach. **Development teams should be diverse and have inclusive participation**. Diversity facilitates the identification of biases throughout the development and validation of a system. People belonging to a specific underrepresented community can help identify the biases that affect them
6. Have a multidisciplinary approach. **Development teams should have interdisciplinary knowledge and ethical reflection**. Mitigating bias requires an interdisciplinary approach that includes specialists in various fields, such as domain experts, ethicists, social scientists, philosophers, and legal experts.

[32] https://www.bbc.com/news/technology-35890188

[33] https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

Main points

- Bias can be introduced at any stage of the development of an AI-based solution
- It is necessary to understand what biases are and where they can appear in the process of developing an AI-based solution in order to prevent and mitigate them.
- Humans introduce their own cultural and cognitive biases in AI-based solutions at the development and deployment phases where they interact with the data and the model.
- The best development practices recommend including the 'human in the loop' in the development process, and building diverse, interdisciplinary development teams with ethical reflection and inclusive participation

3.1.4. Fairness and strategies for bias detection, mitigation and elimination

Most of the biases that we can avoid and control without necessarily making cultural and societal changes require technical interventions that we ought to adopt when we design, build and apply an AI-based system. In this section, we focus on the strategies for detecting and mitigating undesired biases, which are the main sources of unfair decisions. Both detection and mitigation of bias in AI is not a trivial endeavour. Indeed, **it is challenging to differentiate between patterns in the data that represent factual knowledge that we want the AI-based system to learn (e.g., obesity increases colorectal cancer risk) and stereotypes that we want to avoid (e.g., fat people do not have exercise habits). Without any intervention, the algorithm will learn both the knowledge and the stereotypes.**

As previously mentioned, the concept of 'fairness' itself is very difficult to specify. Indeed, several notions of fairness exist that are not only technically defined but also entangled with concepts of social justice, specifically the concept of **privilege**, held by virtue of belonging to certain social identity groups[34]. In AI applications, a privileged group is defined by a favourable outcome (e.g., a classification or a predicted score) that provides a **systematic advantage**. Nevertheless, such privilege can be either acceptable or unacceptable depending on the context. For instance, privileged access to health care services according to clinical severity is generally acceptable. However, if the clinical severity is assessed using an AI model that is biased towards, for instance, ethnicity (Gumbsch and Borgwardt, 2021), it is generally not acceptable due to the systematic advantage given to the privileged group. Ethnicity, as well as sex, gender, race and religion, are **protected attributes** that define groups, and are considered sensitive by laws, regulations and policies. As some decisions can be deemed fair from the perspective of a group but not of an individual (Dwork et al., 2012), two main notions of fairness can be defined:

- **Group fairness**: groups defined by sensitive attributes receive similar predicted outcomes.
- **Individual fairness**: similar individuals receive similar predicted outcomes.

These notions of fairness map to three main **fairness criteria** that AI-based systems should satisfy (Barocas et al., 2019), based on the predicted outcome *R*, the target variable *Y*, and the sensitive attribute *A*: independence, separation, and sufficiency (Table 2).

[34] https://columbiabasin.libguides.com/c.php?g=1049179&p=7614300

Table 2 – Fairness or non-discrimination criteria

Fairness criterion	Formula	Definition	Equivalent terms
Independence	$R \perp A$	The protected attribute is statistically independent of the prediction.	Demographic parity; Statistical parity; Group fairness
Separation	$R \perp A \mid Y$	The protected attribute is statistically independent of the prediction given the target value.	Equalised odds; Conditional procedure accuracy; Disparate mistreatment avoidance
Sufficiency	$Y \perp A \mid R$	The protected attribute is statistically independent of the target value given the prediction.	Cleary model; Conditional use accuracy; Calibration within groups

R: prediction outcome; A: sensitive characteristics; Y: target variable. Formalism and equivalency of terms.

Source: Barocas et al., 2019.

In the case of a binary classifier, the independence criterion can be expressed as equal probability for two groups of receiving the same predicted class:

$$P(R = 1|A = a) = P(R = 1|A = b) \quad \forall r \in R \quad \forall a, b \in A$$

The independence criterion can also be expressed as a difference (a.k.a. disparate impact difference) or as a ratio (a.k.a. disparate impact ratio, relative risk ratio, adverse impact ratio) with the possibility of having the classification constraints relaxed by a slack variable $\varepsilon > 0$:

$$\frac{P\,(R = 1|A = a)}{P\,(R = 1|A = b)} \geq 1 - \varepsilon \qquad \forall r \in R \quad \forall a, b \in A$$

Another form of relaxation of the independence criterion is the **conditional statistical parity** (Corbett-Davies et al., 2017), which permits a set of legitimate attributes *L* to affect the outcome, for instance the marital status in predictions related to financial well-being:

$$P\,(R = 1|L = l, A = a) \;=\; P\,(R = 1|L = l, A = b) \qquad \forall r \;\in\; R \quad \forall a, b \;\in\; A \quad \forall l \in L$$

Despite being widely used, decisions based on an AI system that satisfies the independence criterion can lead to undesired bias, for instance, when the clinical severity is assessed at the same rate in two groups but the error of this prediction differs between them. Instead, the separation criterion demands an equality of error rates in the two groups, which can also accommodate some forms of relaxation. For instance, depending on the scenario, we might only require equality of false negative rates (**equal opportunity** (Hardt et al., 2016)) or of false positive rates (**predictive equality** (Chouldechova, 2017)). Other forms of relaxation of the separation criterion include balancing for the negative class (Kleinberg et al., 2017), and equalising correlations (Woodworth et al., 2017).

Finally, the sufficiency criterion acknowledges that a prediction can absorb the sensitive characteristic for the purpose of predicting the target, for instance by applying the same decision threshold (margin) to all groups. Nevertheless, sufficiency does not imply equality of positive predictive value (**predictive parity**, a.k.a. outcome test (Chouldechova, 2017)) or vice versa. Indeed,

if some individuals of one group are far from the margin, the decision threshold could incur a biased decision. This problem is known as infra-marginality (Simoiu et al., 2017).

It is important to stress out that, although the three fairness criteria (independence, separation, and sufficiency) have diagnostic value in highlighting how different groups experience different costs of classification, a conclusive argument cannot be drawn based on any fairness criterion alone (Corbett-Davies et al., 2017). For instance, an external intervention on any criteria could conceal an unacceptable practice and, as we have shown, each criterion is based on strong assumptions and can be equipped with different forms of relaxation.

When building a model, a fairness criterion can be achieved through three algorithmic interventions (Table 3). The choice of such techniques depends on the degree of control over the training process and the possibility of accessing the raw data, the training pipeline, or only the trained model.

Table 3 – Techniques for algorithmic interventions to achieve fairness criteria

Technique	Description	Example	Advantages	Disadvantages
Pre-processing	Adjusting the feature space before training the model	Disparate impact remover (Feldman et al.)	The feature space is transformed into a fairer representation	These techniques are agnostic of the downstream applications
In-processing	Adjusting the constraints of the optimization process at training time	Adversarial de-biasing (Lahoti et al.)	The classifier is optimised following a specific fairness criterion	These techniques are specific to each model or optimisation process
Post-processing	Adjusting the model once it is fully trained	Multi-accuracy boost (Kim et al.)	These techniques are suitable for any black box model and no re-training is needed	These techniques generally display a limited utility compared to the previous ones

Examples of notable machine learning algorithms that have been rendered fair through specific interventions include fair regression (Agarwal et al., 2019), fair decision trees (Aghaei et al., 2019) and fair support vector machines (Olfat and Aswani, 2018), amongst many others.

Moreover, toolkits have been developed to address machine learning fairness (Table 4) and several companies and organisations are committed to this area (Table 5).[35]

[35] https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/mgi-tackling-bias-in-ai-june-2019.pdf

Table 4 – Selection of tools for machine learning fairness

Tool	Company	Description
Fate	Microsoft https://www.microsoft.com/en-us/research/theme/fate/	Tools and services
Fairlearn	Microsoft https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/	Interactive visualisation dashboard and unfairness mitigation algorithms
AI Fairness 360	IBM https://github.com/Trusted-AI/AIF360	A comprehensive set of fairness metrics for datasets and machine learning models, explanations for these metrics, and algorithms to mitigate bias in datasets and models.
Crowdsourcing	Microsoft & University of Maryland	Detection of bias in natural language processing applications
Clearbox	ClearBox AIhttps://clearbox.ai/	Synthetic Data Generation
Datagen	Datagen https://datagen.tech/	Synthetic Data Generation
Synthesis	Synthesis AIhttps://synthesis.ai/	Synthetic Data Generation
Mostly	Mostly AIhttps://mostly.ai/	Synthetic Data Generation
FairML	DSpace@MIT https://github.com/adebayoj/fairml	Tool box for diagnosing bias in predictive modelling
Teach and Test	Accenture https://newsroom.accenture.com/news/accenture-launches-new-artificial-intelligence-testing-services.htm	Methodology for AI-based solutions validation
What-if tool	Google https://pair-code.github.io/what-if-tool/	Visually probe the behaviour of trained machine learning models
LIME	O´Reilly https://www.oreilly.co	Generation of explanations for the predictions of machine learning

	m/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/	classifiers
Gender Shades	MIT Media Lab http://gendershades.org/	Benchmark image dataset

Table 5 – Companies and organisations providing services for machine learning fairness

Name	Purpose
Eticas Consulting https://www.eticasconsulting.com/	Applied ethics solutions and algorithmic auditing
ITU-T Focus Group on AI for Health & WHO (Wiegand et al.)[36]	Inter-agency collaboration between the WHO & ITU to create a benchmarking framework to assess the accuracy of AI in health
Consulting and Algorithmic Auditing (ORCAA) https://orcaarisk.com/	Consulting company that helps companies and organisations manage and audit their algorithmic risks
Sama https://www.sama.com/	Training Data Platform
Appen https://appen.com/	Data sourcing, data annotation and model evaluation
Z-Inspection http://z-inspection.org/	Validation and ethics certification

Main points

- The design of strategies for bias detection, mitigation and elimination require the application of fairness criteria, namely independence, separation, and sufficiency, with specific levels of relaxation.
- Algorithmic interventions, to achieve a fairness criterion can be applied during pre-processing, in-processing, or post-processing.
- Companies and research institutions are producing several toolkits for machine learning fairness as well as best practice recommendations, but there is still a lack of standardisation.

[36] "Whitepaper for the ITU/WHO Focus Group on Artificial Intelligence for Health"

3.2. Normative analysis of biases

Once we have already analysed the technical framework regarding biases, it is time to address these issues from the perspective of the law. What are the legal tools that might serve us well to prevent such bias? This is a complicated question. To begin with, 'bias' is an almost unknown term in the European regulatory ecosystem. The multiple kinds and sources of bias make it difficult to conceptualise and regulate bias. Furthermore, the polysemy of the term can lead to confusion. 'Bias' is widely used as 'a tendency to prefer one person or thing to another, and to favour that person or thing', this is, as a synonym for 'prejudice'. However, as previously stated, in statistics -a field that is of fundamental importance for algorithmic systems- bias refers to any type of systematic error or deviation that is found with the use of statistical analysis. While bias as prejudice expresses a problematic ethical burden, bias in its statistical sense does not. In fact, the introduction of (statistical) bias in a dataset can be the solution to (prejudice) bias embedded in a database that accurately expresses the real world.

Therefore, we should start by distinguishing two different issues: the problems related to discrimination as such and the problems related to quality of the data. This is precisely what the next section is devoted to. Afterwards, we will highlight the main limitations in the current regulatory approach to address those problems at the data collection and pre-processing stages.

3.2.1. A preliminary issue: to what extent are biases relevant at the regulatory arena

Biases in algorithmic systems can generate results that are unlawful in a broad sense. Here, we understand unlawful as any result that harms or contravenes a right or a benefit protected by law. The identification of bias with prejudice has led legal analysis to focus on the problems of discrimination they can generate. Yet, this analysis has a fundamental limitation, which is that unlawful results can also stem from non-prejudice-related bias. This is why we propose to divide the relevance of bias at the normative level in the following categories: problems of discriminatory nature and problems related to quality.

Problems of a discriminatory nature

Algorithmic systems and discrimination are inseparably tied: machine learning algorithms are supposed to discriminate between data points – that is why we use them – yet some logics of discrimination, even if predictively valid, are not socially acceptable (Veale and Binns, 2017). Discriminatory problems lie not only on how these models work, but also in the limited human capacity to understand how they perform this work. **Epistemic opacity** of Machine Learning algorithms, understood as the human inability to understand the complexity of the inner workings of certain models, is a major obstacle to understanding discriminatory problems. Moreover, the **opacity** regarding the operation of these models **as a deliberate choice** (often on a regulatory basis such as trade secret or intellectual property), may be equally harmful because it does not allow other key aspects to be explored. Even if opacity in algorithmic systems can be considered a problem in itself, it also makes it difficult to uncover discrimination (Borgesius, 2020).

Understanding the correlations that lead to a particular decision may be humanly unfeasible, even if work is done on methods to facilitate the interpretability and explainability of models. However, understanding what kind of data has been used to feed a model is a political decision.

Quality problems

Bias can lead to harmful outcomes without necessarily being discriminatory. Bias may affect the level of accuracy of algorithmic results and, therefore, contribute to an unjustified harm to different goods affected by the use of an algorithmic system. However, this harm does not necessarily entail affecting the way in which goods and services are distributed in a discriminatory way.

Given the various kinds of normative problems that may arise in this context, it may be useful to label them as 'quality problems'. Some regulations protect the value of accuracy in the collection and processing of data. Derived from the fundamental right to the protection of natural persons in relation to the processing of personal data, the GDPR includes the accuracy principle for the processing of personal data in its Article 5(1)(d). Other regulations, especially those regulating the development of products to be placed on the EU market, protect values such as 'quality' or 'safety'. For example, to achieve a high level of human health protection (Article 168 TFEU), the Medical Device Regulation (MDR) sets high standards of quality and safety for medical devices by ensuring, among others, that data generated in clinical investigations are reliable and robust and by requiring a rigorous quality management system to manufacturers. This means that, under the MDR, manufacturers must not place on the market AI medical systems based on faulty data that endangers the clinical condition or the safety of patients (Stöger et al, 2021).

As we will see below, anti-discrimination legal remedies have limitations when it comes to protecting harms produced on ad hoc groups formed by algorithmic correlations. For this reason, it seems indispensable to emphasise that other rights and values -and the governance mechanisms linked to them- can protect against the 'quality problems' that bias raises.

3.2.2. Exposition of the current regulatory approach

Non-discrimination is a fundamental value at the EU level and it is part of its legal framework. In primary law, article 2 of the Treaty of the European Union states that 'The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail.' Furthermore, its article 3(3) states that the Union 'shall combat social exclusion and discrimination, and shall promote social justice and protection, equality between women and men, solidarity between generations and protection of the rights of the child.' Similar ideas can be found in the Treaty on the Functioning of the European Union, especially in its Part Two, entitled 'Non-Discrimination and Citizenship of the Union' (see article 19.1.) and the EU Charter of Fundamental Rights (article 21).

The question is **how to ensure that this anti-discrimination stance translates into an implementation of policies able to fight biases in datasets and AI tools**. In principle, there are two possibilities to consider. The first concerns the anti-discrimination law tools that emerged from the year 2000 onwards. As will be explained below, this legal framework, unfortunately, shows certain loopholes. It is therefore relevant to analyse the possibilities presented by the GDPR. In our opinion, there are many reasons to argue that it can be an effective tool to ensure an efficient defence of the values of equality and non-discrimination, even though anonymised data will still cause concerning issues. Last but not least, the new regulations that have merely been approved or are now being discussed include some important measures to avoid biases that will be carefully explored in this study.

Anti-discrimination regulations

There are some secondary legal tools on discrimination at the EU level. T**he directives approved at the beginning of this century**[37] **constitute a valuable effort towards this purpose**[38]**. They**

[37] Until 2000 non-discrimination law in the EU applied only to the context of employment and social security, and only covered the ground of sex.

[38] This normative framework includes the Gender Equality Directive (Council Directive 2004/113/EC), the Racial Equality Directive (Council Directive 2000/43/EC), Directive 2000/78/EC, establishing a general framework for equal treatment in employment and occupation, Directive 2006/54/EC of the European Parliament and of the

introduced a wide protection against both direct[39] **and indirect**[40] **discrimination.** Unfortunately, the specific legislative technique selected to regulate these issues has drawn a quite complicated framework that hinders adequate protection of the human rights involved. Some of the aforementioned rules are aimed at avoiding discrimination based on a specific factor in a wide range of human activities. Others, on the other hand, sought to avoid discrimination in a broad sense of the term, i.e., including multiple factors, in many different sectors. As a consequence, there are some loopholes that hinder adequate protection against discrimination[41]. For instance, the goods and services equality directives only cover gender and race as protected grounds and not the others. Furthermore, most of the applicable **directives include some possibilities for exceptions to the ban on discrimination especially when we talk** about indirect discrimination. In such cases, the Directives state that discrimination can be permissible if it is 'objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary'.[42] The problem is that **such open-ended exceptions can create uncertainty when determining whether or not a form of discrimination is justified.** This, in principle, should be solved through case law. However, we are still far from having sufficient precedents in this regard. Even worse, this contextually limited interpretation of the concept of discrimination has been endorsed by the Court of Justice of the EU[43].

As a result of these legislative gaps, **the 'EU law does not protect EU citizens against algorithmic profiling and targeting in this area[44], which means that certain disadvantaged groups can be lawfully excluded from the access to certain goods and services.** For example, one could imagine discrimination to arise in the offer of particularly vital goods and services such as housing, health, education, etc. Even though national law could prohibit such instances of discrimination, no harmonized prohibition exists at EU level' (Gerards and Xenidis, 2021).

Council of July 5, 2006, on the Implementation of the Principle of Equal Opportunities and Equal Treatment of Men and Women in Matters of Employment and Occupation (Recast).

[39] Article 2(2)(a) Directive 2000/43/EC; Article 2(2)(a) Directive 2000/78/EC; Article 2(a) Directive 2004/113/EC and Article 2(1) (a) Directive 2006/54/EC.

[40] Article 2(2)(b) Directive 2000/43/EC; Article 2(2)(b) Directive 2000/78/EC; Article 2(b) Directive 2004/113/EC; Article 2(1)(b)
Directive 2006/54/EC.

[41] For instance, discrimination on grounds of religion or belief, disability, age and sexual orientation is not explicitly prohibited in sectors that are not covered by Directive 2000/78/EC, such as education, social security, and access to goods and services including healthcare, housing, advertising and the media. Indeed, the Gender Goods and Services Directive expanded the scope of sex discrimination to the area of goods and services, but protection on the grounds of sex does not quite match the scope of protection under the Racial Equality Directive since the Gender Social Security Directive guarantees equal treatment in relation to social security only and not to the broader welfare system, such as social protection and access to healthcare and education. (European Union Agency for Fundamental Rights European Court of Human Rights - Council of Europe Handbook on European non-discrimination law 2011, p. 14. At: https://www.refworld.org/pdfid/4d886bf02.pdf)

[42] Council Directive 2000/43, art. 2 ¶ 2.b, 2000 O.J. (L 180) 22, 24; Council
Directive 2000/78, art. 2 2.b, 2000 O.J. (L 303) 16, 18; Council Directive 2004/113, art. 2 b,
2004 O.J. (L 373) 37, 40; Directive 2006/54, of the European Parliament and of the Council, art.
2 1.b, 2006 O.J. (L 204) 23, 26.

[43] Judgment of 18 December 2014, Fag og Arbejde (FOA) v Kommunernes Landsforening (KL) C-354/13 EU:C:2014:2463 [36]; Judgment of 17 July 2008, S. Coleman v Attridge Law and Steve Law C–303/06 EU:C:2008:415 [46] and Judgment of 11 July 2006, Sonia Chacon Navas v Eurest Colectividades SA C-13/05 EU:C:2006:456 [56].

[44] This refers to the goods and services market.

In 2008, the European Commission tried to improve this framework by proposing a Horizontal Directive[45], which was aimed at implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation outside the labour market. This Proposal set out a framework for the prohibition of discrimination on these grounds and establishes a uniform minimum level of protection within the European Union for people who have suffered such discrimination. Unfortunately, it was never approved by the corresponding bodies. Furthermore, a number of EU Member States have delayed the implementation of these directives for many years and it does not seem that this situation is going to change soon.

Main points:

- In the EU context the prohibition on discrimination seems to be limited to particular contexts and concrete factors. This scenario introduces doubts about the permissibility of the use of algorithms that introduce bias in some specific cases that need to be addressed as soon as possible.
- There are some different ways to deal with this situation. For instance, incongruences and contradictions could be eliminated by improving the current regulations, along the lines proposed in the Proposal for a Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation.

Data protection regulations. Analysis of some essential issues

The main conclusion reached in the previous section is that the EU anti-discrimination regulations do not include some concrete types of discrimination. **There are some good reasons to consider that the GDPR could be an excellent tool to address many of the loopholes and rigidities that hinder the extensive application of discrimination regulations**. First, the GDPR is aimed at protecting 'fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data' (art. 1.2.). Thus, one should keep in mind that the application of fundamental values such as non-discrimination is not only possible, but also necessary from a GDPR perspective.[46] Second, one must consider that the GDPR includes '**fairness**' in the principles that rule personal data processing (see article 5 GDPR) and also in recital 39, which declares that 'any processing of personal data should be lawful and fair'. Thus, processing cannot take place if it is not 'fair'.

We have previously introduced the concept of fairness as an essential condition to proceed with the bias mitigation schemes from a technical point of view. Its importance in the legal arena is even higher when we face issues related to biases: since discrimination is clearly unfair, the concept of fairness could play a key role in terms of biases mitigation.[47]: Indeed, **the lack of a concrete definition of 'fairness' could allow us extend the use of this concept to include cases of**

[45] Proposal for a Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation {SEC(2008) 2180} {SEC(2008) 2181}, at: https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A52008PC0426

[46] This interpretation has been endorsed by the CJEU, which has ruled "that the EU data protection legislation must be interpreted and applied in the light of the fundamental rights, enshrined in the Charter, and that data protection should play an instrumental role for the protection of the right to non-discrimination in particular" (Ivanova, Yordanka. "The Data Protection Impact Assessment as a Tool to Enforce Non-discriminatory AI." Annual Privacy Forum. Springer, Cham, 2020).

[47] See Recitals 71 and 129 of the GDPR

inequity and/or discrimination that go beyond the classical 'non-discrimination' categories included in the Directives.

There is a nice argument that supports such additional use of the concept: even though the European Court of Justice has never defined the notion of fairness in data protection law, it has used this notion of fairness in two different contexts: fair balance and transparency.[48] If we consider that fairness has to do with the reasonable expectations of data subjects, **then it should help us to avoid some of the discriminatory results that are not so easy to uncover**: a data subject would hardly allow the type of processing that would cause him/her to suffer a damage that other people do not suffer. Thus, if this is the final result, an appeal to fairness seems suitable.

There is, finally, an additional reason to support data protection regulations: it should be emphasized that the approach based on these regulations has an enormous advantage over any other. Since a fundamental principle in the GDPR is data protection by design, it is obvious that **the data controller will be obliged to introduce measures to control bias before processing begins.** Those measures will have to be reviewed afterwards, even after the deployment of the tool. This makes it much easier to prevent the non-discrimination check from occurring only when the damage has already been done.

What are the concepts and tools that shall render this principle applicable in practice? The GDPR includes an essential tool that should serve well (if adequately performed) to face the risk posed by biases: the **data protection impact assessment** (DPIA), which constitutes an excellent opportunity to ensure an adequate protection of the human rights at a preliminary stage of any processing. A DPIA is a **continuous process** that guides and supervises the implementation of a processing activity so that it complies with all data protection requirements and the impact on natural persons is minimized.[49] It is based on a risk-based approach that works extraordinarily well with the AI Act that is about to be approved in the EU arena (we will introduce it in the next section).

The introduction of elements corresponding to the assessment of bias in a DPIA is important, because this contributes to the proper implementation of the idea of fairness by design just mentioned. AI in general and its ability to introduce biases in particular is difficult to control *a posteriori*. Recourse to the courts is neither the only nor the best mechanism to prevent biases from occurring. Hence, a mechanism such as the DPIA, which makes it possible to assess the possibility of pernicious consequences of the use of an AI mechanism from its early stages of creation, is particularly promising for achieving the intended goal. **It is, however, necessary to point out that DPIAs might get into conflict with the type of impact assessments foreseen in the AI Act, creating a considerable legal chaos.**

To sum up, data protection regulation could become an efficient tool in order to fight against bias. However, it suffers from a condition that limits its real effectiveness: that is, **it only focuses on the processing of personal data**. It is, therefore, perfectly possible to train and validate AI systems using non-personal data and the GDPR will not be applicable to those datasets. This may have irreparable consequences. While anonymization protects the identity of individuals, the data is imbued, nonetheless, with personal characteristics that are equally biased. To deal with these consequences, **AI providers should conduct impact evaluations in training phases, even if no personal data is involved.** To do otherwise merely entails delaying to a later stage those phases that would have been better implemented from the beginning. **An explicit provision in the AI Act could contribute to clarify this issue.**

---

[48] CJEU Opinion 1/15 on Draft agreement between Canada and the European Union — Transfer of Passenger Name Record data from the European Union to Canada, ECLI:EU:C:2017:592; Case C-524/06 Heinz Huber v Bundesrepublik Deutschland, ECLI:EU:C:2008:724

[49] PANELFIT Guidelines. DPIA.

Main points:

- The GDPR could be an excellent tool to fight against bias through its concept of "fairness" and the use of DPIAs as essential tools to protect data subjects against bias, if adequately performed.
- However, it shows some weaknesses. The main problem of data protection law is its lack of coverage of databases that do not contain personal data.
- AI impact assessments could help us solve this deficiency. However, it would be necessary to consider how to conciliate these tools with the DPIAs.

The new regulatory tools. A general introduction

Beyond the discrimination regulations and the GDPR, there is another set of rules that are also of great importance when approaching the issue of bias. This set of rules comprises some regulations that are about to be passed, such as the Digital Services Act and the Digital Markets Act, which are aimed at creating a safer digital space where the fundamental rights of users are protected, and establishing a level playing field for businesses. Both of them include measures devoted to fighting against discrimination. The **Digital Services Act** introduces an obligation for large online platforms to **introduce measures that ensure that the design of the algorithmic systems used by them does not create discrimination among the users of the platforms.**[50] These measures include the need to conduct risk assessments and design their risk appropriate mitigation measures[51]. The Act also states that very large online platforms should be accountable, through independent auditing, for their compliance with the obligations laid down by the Regulation and, where relevant, for any complementary commitments undertaken pursuant to codes of conduct and crises protocols[52]. Needless to say, these provisions shall be efficient in order to fight against bias.

The Digital Markets Act does not include such explicit references to the measures that need to be incorporated to fight bias, but it does state that a gatekeeper[53] **shall apply fair and non-discriminatory general conditions of access for business users** to its software application store[54]. It also introduces some clauses aimed at prohibiting confidentiality clauses in agreements or other written terms that hinder the exercise of the right of business users to raise concerns about unfair behaviour by gatekeepers with any relevant administration or other public authorities[55].

However, the regulations that are of most interest for analysing how to avoid biases that cause discrimination in both datasets and AI mechanisms are primarily the Data Governance Act, the Data Act and the Artificial Intelligence Act (AI Act onwards)[56]. The most important clauses regarding data are probably those included in article 10 of the AI Act:

> *3. Training, validation and testing data sets shall be relevant, representative, free of errors and complete. They shall have the appropriate statistical properties, including, where applicable, as*

[50] Recital 57 and article 26.1 (b). Digital Services Act
[51] Recital 58. Digital Services Act
[52] Recital 59. Digital Services Act
[53] That is, a provider of core platform services.
[54] Article 6.1 (k) Digital Markets Act
[55] Recital 39 Digital Services Act
[56] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Act.

regards the persons or groups of persons on which the high-risk AI system is intended to be used. These characteristics of the data sets may be met at the level of individual data sets or a combination thereof.

4. Training, validation and testing data sets shall take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, behavioural or functional setting within which the high risk AI system is intended to be used.

The **main issue** that we have to face is that the proposed Acts are missing a clear connection with the data protection law. Some concepts that play a key role in the GDPR, such as data minimization and data protection by design are not so apparently important in the Acts. The AI Act does not mention them explicitly whereas the Data Act only includes some references in articles 8 and 34. In the next sections, some more misalignments that may generate legal uncertainties for all concerned bodies will be showed.

3.2.3. Essential legal tools to fight biases

In previous pages, we have analysed the main normative approaches to bias. In this section, we examine the main measures and tools that will serve to mitigate them. To this purpose, it is first necessary to establish **certification procedures** for AI systems. **These procedures must sometimes be dynamic,** given that some AI tools learn from the data gathered. Indeed, in the case of AI, 'risk management system shall consist of a continuous iterative process run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic updating'[57]. Datasets, on the other hand, will probably need to include an adequate information about some of their most important features if they are expected to be used in AI systems development processes. Since such datasets are often altered, the information provided should also be updated.

In turn, the efficiency of these certification processes depends on the creation of **standards** that 1) provide reliable knowledge of the datasets (including essential information on how they have been created, what types of data they contain, etc.)[58], and 2) provide homogeneous guidelines to determine the absence of bias in the predictions made by the AI mechanism. This should be reinforced with a certification system and the implementation of accountability measures that ensure that the parties involved implement an adequate mitigation of the risks that arise. All these ideas are explored in deep in the next pages.

Standardisation as a way to avoid biases

The **creation of standards applicable to datasets and AI mechanisms is a fundamental pillar in the regulation of these assets[59].** Given the diversity that characterizes this sector, it would be very difficult to validate datasets or AI systems –and thus, provide universally recognized certificates- without using standards. However, it must be emphasized that **standardization is in its early stages,** both in datasets and in the AI arena. There are no universally agreed standards for data quality assessments for machine learning tools at the moment, even though there are some interesting proposals in this area (Gebru et al, 2021).

This will probably change in the next few years, since considerable efforts are being implemented to improve this scenario. In the case of **the currently discussed regulations**, some efforts worth highlighting are related to the promotion of standardization. Indeed, article 29.4. of the Data Act states that the Commission may, in accordance with Article 10 of Regulation (EU) No 1025/2012,

[57] AI Act, article 9.2
[58] See AI Act, article 10.2
[59] See REGULATION (EU) No 1025/2012 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 25 October 2012 on European standardisation, on the regulation of standards at the EU level

request one or more European standardisation organisations to draft harmonised standards that satisfy the essential requirements under paragraph 1 of Article 29.1 of the Data Act**. The requirements that these standards shall fulfil include information about the dataset content, use restrictions, licences, data collection methodology, data quality and uncertainty.** Furthermore, article 28 (b) of this proposed regulation details that such datasets shall include information about: '*the data structures, data formats, vocabularies, classification schemes, taxonomies and code lists shall be described in a publicly available and consistent manner*'.

All this information shall be sufficiently described to allow the recipient to find, access and use the data. This is well aligned with the proposals made, by instance, by the EU Agency for Fundamental Rights, which defines what constitutes a minimum guidance for understanding the quality of data on the basis of the following questions (Focus, 2019):

- 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?

Thus, the Data Act is making a significant effort to introduce standards in the datasets scenario. These provisions work well with the European strategy for data[60], which constantly appeals to the need to create standards applicable to datasets. In any case, one must always keep in mind that standardisation in datasets would have to allow for **flexibility** in order to be able to include the variety of possible data formats and collections used in AI applications. This is essential, because data formats often differ substantially and, moreover, they are not usually collected for a specific purpose. Thus, it makes good sense that the AI Act specifies that these requirements 'can have a generic nature or concern specific sectors' (Art. 28.1). Last, but not least, it is worth mentioning that the Data Act states that where the standards do not exist for the service type concerned, '*the provider of data processing services shall, at the request of the customer, export all data generated or co-generated, including the relevant data formats and data structures, in a structured, commonly used and machine-readable format*' (Art. 26.4). This might be of particular interest at the present circumstances.

In the case of the AI systems, the AI Act states that 'High-risk AI systems which are in conformity with harmonised standards or parts thereof the references of which have been published in the Official Journal of the European Union shall be presumed to be in conformity with the requirements set out in Chapter 2 of this Title, to the extent those standards cover those requirements.' (article 40). However, these **standards do not exist at present at the EU level**. Some organizations, such

[60] Communication from the commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions a European Strategy for Data. Brussels, 19.2.2020 COM(2020) 66 final.

as the IEEE[61] or the ISO are trying to produce available tools,[62] but they are still in their preliminary stages.[63] Moreover, the standards of these organizations may be insufficient to meet the requirements of European regulations, which are often more demanding than those of other legal frameworks**.** Indeed, the EU legislature will probably ask the European Standardization Organizations (ESOs) to produce harmonised standards. Therefore, it seems reasonable to expect organizations such as ETSI, CEN and CENELEC (which are the three European Standardization Organizations that have been officially recognized by the European Union and by the European Free Trade Association (EFTA) as being responsible for developing and defining voluntary standards at European level) to develop their own standards, specifically adapted to the EU legal framework. According to the available information, they have already started with some preliminary works. For instance, the 2022 edition of the Putting Science into Standards (PSIS) workshop, jointly organised by CEN and CENELEC and the Joint Research Centre of the European Commission (JRC), will be dedicated to Data Quality Requirements for inclusive, nonbiased, Ethical AI.[64]

It would be advisable for **the construction of all standards to be addressed through a process of co-creation, which would allow the inclusion of suggestions from stakeholders and non-governmental organisation (NGO) representatives in the final product.** It is important that citizens feel that their interests have been taken into account and that the standards guarantee an adequate level of protection. Measures must also be taken to ensure that the mechanisms adopted are of sufficient quality, both from a strictly technical and regulatory point of view.

[61] Namely, the Standards P7003 - Algorithmic Bias Considerations (https://standards.ieee.org/ieee/7003/6980/) The IEEE is also creating a more general standard, the IEEE CertifAIEd Mark. This is aimed at conveying an organization's capability to fulfil applicable transparency, accountability, reduction of algorithmic bias and privacy requirements stipulated in the appropriate criteria to foster trust and facilitate the adoption and use of AI products and services. See: https://engagestandards.ieee.org/ieeecertifaied.html

[62] Indeed, some years ago IEC and ISO created a joint committee (ISO/IEC JTC 1/SC 42) that is leading standardization activities for artificial intelligence. It has created working groups on: Foundational standards, trustworthiness, use cases and applications, data, computational approaches and computational characteristics of Artificial Intelligence and AI Applications, and governance implications of AI. Currently 45 countries participate actively in this work and there is extensive involvement from EU and European countries, with 20 countries participating and holding a total of 18 leadership positions such as convenorships and editors (CEN-CENELEC response to the EC White Paper on AI, version 2020-06, page 2, at: https://www.cencenelec.eu/media/CEN-CENELEC/AreasOfWork/CEN-CENELEC_Topics/Artificial%20Intelligence/Quicklinks%20General/Documentation%20and%20Materials/cen-clc_ai_fg_white-paper-response_final-version_june-2020.pdf). A key tools, the One of the documents on works, ISO/IEC AWI TS 12791 (entitled "Treatment of unwanted bias in classification and regression machine learning tasks"), is aimed at providing provides mitigation techniques that can be applied throughout the AI system life cycle in order to treat unwanted bias (https://www.iso.org/standard/84110.html?browse=tc)

[63] https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0

[64] CEN-CENELEC. WORKING PROGRAMME, 2022, p. 118, at: https://www.cencenelec.eu/media/CEN-CENELEC/News/Publications/2022/cen-cenelec_work_programme2022.pdf

Main points:

- The creation of standards applicable to datasets and AI mechanisms is a fundamental pillar in the regulation of these assets. However, standardisation is in its early stages, both in datasets and in the AI arena.
- Standards related to datasets shall include information about the dataset content, use restrictions, licences, data collection methodology, data quality and uncertainty.
- Standardisation of datasets used by AI tools would have to allow for flexibility in order to be able to include the variety of possible data formats and collections used in AI applications.

Certification

A certification can be defined as the attestation that a product, process, person or organization meets specific criteria.[65] Certifications are usually aimed at **reducing information asymmetries (**Cihon et al., 2021) **and their use is** usually considered a clear cost saving. Somehow, certification is linked strictly to standardization, even though it is a different concept: certifications are meant to give evidence that standards are fulfilled.

The concept of certification is already used in multiple sectors and is starting to be introduced in the AI sector. There are already some programs in place, both at the European Commission level (Cihon et al, 2021) as well as in the Institute of Electrical and Electronics Engineers (IEEE),[66] or some specific countries (Malta Digital Innovation Authority, 2019; Finkel, 2018). **The AI Act includes this approach by imposing a certification system -the Conformity Assessment-** that covers the mandatory requirements applicable to high-risk AI systems, based on European harmonized standards under Regulation (EU) No 1025/2012 and common specifications to be established by the Commission.

This system, however, **mainly relies on self-certification, even though it is supervised by a** structure of entities (Notifying Authorities/Notified Bodies/Commission)**.** Indeed, Internal Assessment procedures are permitted in a substantial number of cases in the AI framework. This has been criticised by the EDPB and the EDPS, which consider that third-party conformity assessment must be carried out on a general basis for high-risk AI (EDPB-EDPS, 2021). This seems to be an appropriate suggestion, which works well with some relevant opinions in the academia (Veale and Borgesius, 2021).

Certifications could also apply to **datasets,** of course, but **the Data Act and the Data Governance Act proposal do not specifically impose them.** However, some considerations regarding standards have already been included in their clauses as previously mentioned, and a voluntary certification model will probably arise. That would be an excellent idea, since certification could help AI developers ensure that the datasets that they use to train or validate the algorithms are not biased. Of course, the problem of so-called 'ethics washing', i.e. the possibility of companies incorporating weak requirements when providing their own certificates, will always exist. Thus, if it is a third party who performs the certification this would surely increase confidence, since AI developers and providers should not have to rely on the information provided by whoever created

[65] Conformity assessment — Vocabulary and general principles, ISO/IEC 17000:2020.
[66] IEEE, "The ethics certification program for autonomous and intelligent systems (ECPAIS)." [Online]. Available: https://standards.ieee.org/industry-connections/ecpais.html, Accessed on: Aug. 3, 2020.

the databases or perform deep analysis of their own. However, the intervention of third parties involves some inconveniences, namely, higher cost and delays in the AI tool preparation process.

In any case, **intervention by the States imposing certification on datasets by third parties might be unnecessary,** especially if AI providers decide exclusively to use datasets that contain some form of certification attesting to the quality of their data. Market inertia will eventually impose this modus operandi. If this is not the case, we should probably have to reconsider this issue.

However, there is another issue that needs to be adequately addressed. As pointed out by the EDPB and the EDPS, the certification mechanism created by the AI Act *'is different from the certification system aimed at ensuring compliance with data protection rules and principles, outlined in Articles 42 and 43 of the GDPR. It is however not clear how certificates issued by notified bodies in accordance with the Proposal may interface with data protection certifications, seals and marks provided for by the GDPR, unlike what it is provided for other types of certifications (…) As far as high-risk AI systems are based on the processing of personal data or process personal data to fulfil their task, these misalignments may generate legal uncertainties for all concerned bodies, since they may lead to situations in which AI systems, certified under the Proposal and marked with a CE marking of conformity, once placed on the market or put into service, might be used in a way which is not compliant with the rules and principles of data protection'* (EDPB-EDPS, 2021). Therefore, **the relationship between certificates issued under the said Regulation and data protection certifications, seals and marks needs to be clarified.**

Main points:

- The AI Act is imposing a certification system -the Conformity Assessment- that covers the mandatory requirements applicable to high-risk AI systems. This could be an excellent tool to fight bias.
- However, there are some issues that should be considered regarding certification and AI tools: 1) the certification schemes are still a work in progress; 2) self-assessment might not be the best approach to avoid some issues; the relationship between certificates issued under the AI Act and the GDPR needs to be clarified

Compliance and monitoring

Monitoring of high-risk AI tools is certainly needed if we want to mitigate bias. As already mentioned, the process of creating and deploying an AI mechanism is complex. Only if we are able to determine at what point a bias has been introduced will it be possible to derive responsibility for its consequences.

However, monitoring AI tools can sometimes be extremely complex for several reasons. The **first** is that the target population might be very variable: Thus, the evidences gathered cannot be universalised. If you use the same AI tool in Nebraska and in Madrid, it might happen that results will be different in terms of bias. On the other hand, some AI systems are based on **dynamic processes**. Unlike other mechanisms, such as those in the pharmaceutical industry, for example, the product (the dataset or the AI tool, for instance) **changes over time,** since they incorporate new data and/or learn from the new data. Thus, what is true at one point in time may not be true at another. This means that its **control procedures must be continuous**. This dynamism implies that controls over datasets or mechanisms must be exercised essentially in two ways:

- First, AI providers must establish a periodic monitoring system.

- Second, they must take into account the occurrence of circumstances that may substantially affect the data. For example, if there is a wave of refugee women applying for social assistance, their data may introduce trends that differ from historical ones. Hence, there is a need for a new data control.

These requirements have been adequately endorsed by articles 9, 42.4 and 61 of the AI Act. Article 9, devoted to risk management, that stipulates that **the AI provider must perform adequate testing of the high-risk AI systems at any point in time throughout the development process, and, in any event, prior to placing it on the market or putting it into service.** On the other hand, article 42.4 states that 'High-risk AI systems shall undergo a new conformity assessment procedure whenever they are substantially modified, regardless of whether the modified system is intended to be further distributed or continues to be used by the current user'. Finally, article 61 states that AI providers 'shall establish and document a post-market monitoring system in a manner that is proportionate to the nature of the artificial intelligence technologies and the risks of the high-risk AI system (…) The post-market monitoring system shall actively and systematically collect, document and analyse relevant data provided by users or collected through other sources on the performance of high-risk AI systems throughout their lifetime, and allow the provider to evaluate the continuous compliance of AI systems'

On the other hand, it is worth mentioning that the new regulatory framework on AI and data governance is promoting an *ex ante* control of AI mechanisms that includes controlling the quality of the datasets that make up its training data. As already mentioned, **an AI provider must ensure that these datasets meet reasonable quality criteria (Article 10. 1 and 2 AI Act)** and, to this end, it has a lawful basis that allows it to process data of sensitive categories (Art. 10.5, AI Act). The **need to run continuous iterative processes throughout the entire lifecycle of a high-risk AI system, requiring regular systematic updating (art. 10 AI Act) is, of course, also an excellent tool to guarantee an adequate compliance with the requirements settled by the regulation in all stages of the lifecycle of the AI tool.**

However, the governance framework introduced by the AI Act is definitely **complex**. It includes supervision by third parties, with competences shared by the Member States and the Commission as a last resort to ensure compliance. Member States will have to design a national supervisory authority, a notifying authority and a market surveillance authority. The national authority will be competent for designating and monitoring notified bodies (conformity assessment bodies). These notified bodies will perform third-party conformity assessments. To some extent, this monitoring system and the role to be played by the corresponding authorities might overlap with the role that Data Protection Authorities (DPAs) play according to the GDPR. This could be somehow avoided if DPAs were also designated as the national supervisory authorities (EDPB-EDPS, 2021).

Last, but not least, it is necessary to mention that this apparently impressive control system **does not include adequate mechanisms for isolated individuals**, let alone the organizations responsible for protecting their rights, to adequately supervise the functioning of the IA mechanisms. There are no regulatory provisions requiring IA providers to disaggregate information on the results obtained through their use, much less provide documentation of the entire process to these organizations. As stated in (Tambiama 2021) : 'individuals affected by AI systems and civil rights organisations have no right to complain to market surveillance authorities or to sue a provider or user for failure to comply with the requirements. Similarly, Veale and Zuiderveen Borgesius warn that, while some provisions of the draft legislation aim to impose obligations on AI systems users, there is no mechanism for complaint or judicial redress available to them.' This, in a way, deprives us of an element that could be very important to improve the control of these systems.

Main points

- Monitoring of high-risk AI tools is certainly needed to mitigate bias. However, monitoring might become extremely complex due to several reasons. This means that we must create adequate tools able to deal with such complexity. Dynamic monitoring should be carefully considered, as the AI Act is proposing.
- The governance framework introduced by the AI Act is definitely complex. It includes supervision by third parties, with competences shared by the Member States and the Commission as a last resort to ensure compliance. However, it does not provide individual citizens and NGOs protecting human rights with adequate tools to complain to market surveillance authorities or to sue a provider or user for failure to comply with the requirements. This probably has to be reconsidered.

4. Conclusions

Based on what has been presented in the preceding sections, the following conclusions can be reached:

1. The fight against biases – understood as systematic deviations from a reference value, norm, or rationality of judgement of an individual, a group, or an algorithm, producing unfair consequences – is important and complex. There are multiple reasons to argue that the eradication of bias is impossible, so strategies should focus on reducing their incidence and mitigating their effects. This should not be interpreted, in any case, as a problem specific to datasets or AI. A human operator can introduce biases that are much more accentuated and more difficult to eradicate. Therefore, any criticism of the biases derived from the use of AI systems must contemplate that their alternative – the human element – may incorporate the same, or worse, biases.
2. Biases are inherent to human beings, their culture and their history. There are multiple taxonomies of bias, and identifying and classifying them all is complex. AI-based solutions incorporate new biases and tend to magnify existing human biases. To identify and mitigate bias in AI-based solutions, it is necessary to understand and be aware that biases are introduced at all stages of the process of AI development by the training dataset, the algorithm, and the humans involved.
3. An essential step to mitigate biases is to create or use high quality domain-specific training datasets to guarantee a fair knowledge representation of the 'real-world' to the AI-based system. Mechanisms of oversight and accountability should be implemented to continuously assess the quality and integrity of the data.
4. Techniques exist that correct biases in AI systems via pre-processing, in-processing, or post-processing, to achieve a higher fairness in those systems. Currently several companies are developing toolkits to help in this process, although there is still a lack of standardisation in the sector.
5. Best practices recommend including the 'human in the loop' during the development process and building diverse, interdisciplinary development teams with ethical reflection and inclusive participation.
6. It is not easy to draw a regulatory framework that is able to deal with bias, since bias is a complex concept that is not synonymous to discrimination, at least from a legal point of view. In the EU context, the prohibition of discrimination is limited to particular contexts and concrete factors. This scenario introduces doubts about allowing the use of algorithms that introduce bias in some specific cases, and needs to be addressed as soon as possible. There are different ways to deal with this situation. For instance, incongruences and contradictions could be eliminated by improving the current regulations, along the lines contained in the proposal for a Council directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation. However, alternative legal tools able to cope with discrimination issues could also be considered.
7. The GDPR could be an excellent tool to fight against bias through its concept of 'fairness'. However, it also shows some weaknesses. If the main problem of anti-discrimination law is its limited application to the multiple forms of unfair treatment produced by algorithmic systems, the main problem of data protection law is its lack of coverage of databases that do not contain personal data. Anonymisation techniques are appropriate to protect privacy, however, they do not protect against the reproduction of biases when using such data. This is aggravated by the fact that the GDPR principles are not applicable to anonymised data.

8. The creation of standards and certificates applicable to datasets and AI mechanisms is a fundamental pillar in the regulation of these assets. However, they are both in their early stages, both in datasets and in the AI arena. Standards related to datasets should include information about the dataset content, use restrictions, licences, data collection methodology, data quality and uncertainty. On the other hand, standardisation and certifications related to datasets and AI tools must allow for flexibility in order to be able to include the variety of possible data formats and collections used in AI applications.
9. Monitoring of high-risk AI tools is certainly needed if we want to mitigate bias. However, monitoring might become extremely complex, for several reasons. This means that we must create adequate tools able to deal with such complexity. Dynamic monitoring must be carefully considered, as proposed in the draft AI act. The governance framework introduced by the AI act proposal is definitely complex. It includes supervision by third parties, with competences shared by the Member States and the European Commission as a last resort to ensure compliance. However, it does not provide individual citizens and NGOs protecting human rights with adequate tools to complain to market surveillance authorities or to sue a provider or user for failure to comply with the requirements. This should probably be reconsidered.

5. Policy options and assessment

In accordance with the analysis carried out in the previous pages of this study, some policy options that may serve to improve the current situation are suggested in this section. Some of them are complementary and others are mutually exclusive. In this last case, reasons for supporting one or the other are provided, so that policy-makers can gain a better understanding of the present situation.

5.1. Policy option 1: Do not create new regulations specifically focusing on biases. Instead, focus on misalignments between the different regulatory tools

A first and tempting option to address the issue of bias is to **develop specific regulations on this matter.** If this were done at EU level through the creation of a specific tool, such as a directive or a regulation, **we would achieve greater unification of criteria among all Member States when responding to this complex issue.** We would also **gain in legal certainty and specificity** of the applicable legal framework. However, it is **currently not a good idea to introduce a specific regulation on biases in datasets or on biases as a result of the use of AI mechanisms in general**. There are multiple reasons to support this hypothesis:

- First, there is already an **overabundance** of data protection rules in the EU context. Adding a new one, even if specific to bias, would probably be unnecessary. Over-regulation should be avoided wherever possible.
- Furthermore, a **general, common regulation for biases in datasets and for AI tools would be very difficult to design**. Both data sources and AI systems are diverse. There is quantitative and qualitative data, with a wide range of disparity between them. A database consisting of transcripts of conversations (natural language) is not the same as one consisting of numbers, or containing images. Furthermore, datasets and AI mechanisms can be used for different purposes. The measures we can take to control and avoid bias are likely to be very different in each scenario.
- Moreover, it is **difficult to justify that existing regulations are insufficient to solve many of the problems posed by biases**. It is true that the regulations on non-discrimination are somewhat rigid. However, it could be possible to modify the existing directives or to approve the proposed horizontal directive, which would add clarity to the current situation. Furthermore, data protection regulations offer interesting alternatives. The inclusion of the concept of 'fairness' as an essential principle in the GDPR opens the door to the possibility of introducing some specific recommendations on the processing of personal data. Thus, for example, it would be possible for the EDPB to develop an opinion on 'fairness and biases', establishing some criteria on the type of procedures necessary to justify that data processing for the adoption of some types of specific decisions is 'fair'.
- The **draft regulations** currently under discussion introduce some interesting categories and concepts that can draw an adequate approach to the issue of bias. The promotion of standards and certification is a laudable initiative. **The possibilities of these rules could be explored before considering new ones, especially at a time when biases do not yet seem to be an unbearable threat.**

All this, in short, makes it much more advisable **not to produce a specific regulation, but to take advantage of the opportunities presented by that which already exists.** In any case, slight modifications to the regulations that are now in the approval phase would be advisable. This is particularly important to promote best practices, such as the use of standards and the

implementation of certification systems, and to solve misalignment between the different regulatory tools (especially, the GDPR and the new regulations), which may generate legal uncertainties for all bodies concerned. Otherwise, companies may find themselves in breach of regulations at the intersection of the AI act, GDPR and the general product safety regulation (Malgieri and Tiani, 2021)

5.2. Policy option 2: Preventive approach – strengthening bias mitigation from data collection

The application of best practices for bias mitigation should be a policy priority **from the start of data collection onwards**. This means that data collection should comply with the FAIR principles: findable, accessible, interoperable and re-usable. The fact that FAIR principles are respected at the time of data collection will also serve as an extremely efficient measure both for introducing standards for better data governance and for auditing the databases in which they are integrated. In addition, a proper implementation of FAIR principles would allow a much more efficient integration of separately constructed databases, so that a **better control of biases at their source can be ensured**. This means, of course, including adequate information about the characteristics of the data they contain, namely: the data structures, data formats, vocabularies, classification schemes, taxonomies and code lists, which should be described in a publicly available and consistent manner.

These policies fit very well with the **principle of data protection by design**, which is essential under GDPR, but has not been conveniently emphasised in the rules now being discussed on AI or data governance. The introduction of policies aimed at promoting or enforcing FAIR principles and quality controls from the earliest stages of data processing and the construction of the datasets that will feed AI systems is essential to avoid damage occurring at later stages of the development or implementation of these tools. The introduction of standards-based models and certification on datasets can be of great help in ensuring success in this early control of dataset quality. These measures should apply both to personal data and anonymised data. Anonymisation or the use of synthetic data cannot be an excuse to evade accountability for mitigating bias. Non-personal data can be used to make decisions that affect people. Therefore, we should extend the prevention introduced for personal data processing to anonymised, non-personal data processing if such processing involves high risk in terms of the artificial intelligence act.

Finally, it would be worth considering the need for **equally effective measures on data collection in both the public and private sector**. The former may be considered a priority in many respects, but the creation of a fragmented data market, with large differences between the public and private sector, may have a negative impact on bias mitigation.

5.3. Policy option 3: Promote database certification – enforcing bias-awareness at the very beginning of the lifecycle in algorithmic systems

This policy option explores the possibility of introducing certificates for datasets with the objective of avoiding biases. This comprises two main possible actions. First, certifying that the dataset developer has applied the **best available practices** to avoid the presence of significant biases. Second, certifying that the **dataset developer provides accurate relevant information about the dataset that may prevent other stakeholders from developing or using the dataset and the algorithmic system in a biased way**. Of course, such dataset certificates **could be complemented by others referring to the artificial tool that has been shaped thanks to these data.**

However, as argued before, it should be retained that certification can hardly ensure that biases are totally excluded since:

(1) the available techniques for bias mitigation at the pre-processing stage have limited effectiveness; and

(2) the application of these techniques does not imply that the use and deployment of the system fed with these data cannot lead to biased results.

Thus, any kind of certification at this stage **should be conducted in consideration of these limitations and duly communicated to stakeholders in the development, deployment and use of the systems trained with these certified databases.**

In any of the scenarios described below, it would be necessary to describe who has the authority to conduct the certification. There are three main models of certification, which require more or less external intervention in the process: (1) Certification by public authorities; (2) Certification by authorised third parties; and (3) Self-certification.

However, the idea of introducing certificates allows for different approaches. Some options advocate certificates obtained exclusively from public entities, while others allow self-certification. It is also possible to introduce different requirements depending on the level of risk involved in a given treatment. On the other hand, one can opt for either a mandatory or a voluntary system, meaning that the data providers could adopt such certification as a kind of quality check that serves to improve the offer, or, alternatively, the data providers could be required to attach an obligatory certificate.

Moreover, the combination of different kinds of regulatory certification systems considering the different risk levels is perfectly possible and has been already proposed: e.g. only by public authorities for high-risk systems. Alternatively, it could even be possible to combine both policy options: e.g. mandatory certification for high-risk systems and voluntary certification for the rest. In this section, the issues considering these two alternative approaches (mandatory or voluntary certification for high-risk AI system databases) are analysed.

Policy option 3.1: Mandatory certification for databases that will feed high-risk AI systems – making data providers accountable

The first possible option is the introduction of a mandatory certification framework for databases feeding high-risk AI systems. In this case, **certification of databases is mandatory whenever they are to be used for the purpose of developing a high-risk AI system.** Any database to be used for this purpose must be certified either by the database provider, or by the AI system provider where the former has not done so or where they are the same provider.

This policy option would require the **separation of pre-certification and post-certification responsibilities in the proposed artificial intelligence act,** particularly with regard to the responsibilities for data governance, currently set out in Article 10 of the AI act. It therefore also requires the recognition of the database provider, separate from the AI system provider, even though the same company may occupy these roles in many cases.

Policy option 3.2: Voluntary certification for high-risk AI system databases – a tool allowing AI system providers to comply with the regulations and to demonstrate compliance.

This policy option offers **voluntary certification** as a possibility for AI system providers to demonstrate compliance with data governance requirements for the development of high-risk AI systems.

Voluntary certification of datasets **could be, by itself, an appropriate tool in complying with regulation and demonstrating compliance**. The introduction of such certification on the market can provide AI system providers with confidence when acquiring these databases. Dataset providers that achieve this certification will gain a **competitive advantage** over dataset providers that do not.

Both systems include advantages and disadvantages. Mandatory certification involves higher costs. However, it increases the reliability of the datasets, especially if certificates are provided by third parties.

Last, but not least, it should be considered that, although the focus here has been on databases for the development of high-risk AI systems, these policy options would be **equally applicable to the regulation of any other type of AI system, also on a sectorial basis**.

5.4. Policy option 4: Granting transparency rights to AI-system-subjects – opening a window to find the source of biased results

This policy option envisages granting AI-system-subjects transparency rights over the databases on which AI systems are developed that make decisions affecting them.

To this purpose, the rights to access and information provided by the GDPR include 'meaningful information about the logic involved' in automated decision-making and profiling. However, this 'meaningful information' does not extend to information about training datasets that are key to the logic involved in the automated processing. As a result, data subjects' ability to check whether the processing of their data is lawful, fair or accurate in accordance with the GDPR is limited, even though safeguards should serve to preserve their rights. Likewise, proving discrimination without information about the data used for training an algorithmic system is just as complicated.

Therefore, under this policy option AI-system-subjects should be able to ask for meaningful information about training datasets. This information should, at least, include the metadata (or datasheets) about training datasets generated with certification that could serve this purpose. The information should not adversely affect the rights or freedoms of others, including data protection rights, trade secrets or intellectual property. However, the result of this balance should not be a refusal to provide relevant information to the AI-system-subjects.

Unfortunately, individual rights for AI-system-subjects are not currently included in the proposed artificial intelligence act (Veale and Borgesius, 2021). This should be changed if the risk that the rules on prohibited and high-risk practices become infective in practise is to be avoided. For this purpose, it would be necessary to ensure that such **information rights for decisions that significantly affect AI-system-subjects are also extended to hybrid decisions, or recommender and decision-support systems, and not only to decisions based solely on automated processing** (contrary to Articles 13(2)(f), 14(2)(g) and 15(1)(h) in the GDPR). It would not be advisable to implement this perspective of individual rights without other policy options that reinforce other ways of holding data and AI system providers accountable. The information provided by transparency rights alone is rather inadequate to govern algorithmic systems and their data flows (De Hert and Lazcoz, 2022).

5.5. Policy option 5: Facilitating the implementation of the AI act

Finally, it should be considered that compliance with those regulations currently awaiting approval will **entail costs and risks for companies**. First, certification processes involve charges and fees. As an example, 'considering an external audit of €300 per hour, **the costs for yearly audits by a notified body could go up to roughly €9 000 per year** (involving two people working for two

days)**.'**[67] Furthermore, companies will have to wait for certification before launching their products on the EU market, which involves additional costs. This might be challenging, especially for SMEs or companies operating in international markets. Were this the case, **their competitiveness would probably suffer severe damage.** Indeed, the Impact Assessment Accompanying the Proposal for the AI ACT[68] acknowledges that those SMEs that supply high-risk AI systems would 'in principle be more affected than large companies for several reasons'.[69]

Thus, if the new regulations are aimed at boosting the EU's technological and industrial capacity, as well as AI uptake across the economy, they should consider some measures that could make things easier for industry and, in particular, for SMEs. AI regulatory sandboxes[70] are excellent examples of such measures. Similarly, the obligation to consider SMEs' interests when setting fees related to conformity assessment would surely help them reduce their costs. Furthermore, the measures included in Article 55.1 of the draft AI act (such as providing SMEs with priority access to the AI regulatory sandboxes) are excellent steps in the right direction. However, **these measures will not serve to erase all additional costs caused by the new regulations**. As the Impact Assessment acknowledges, 'whether the additional costs can at the margin discourage some SMEs from entering into certain markets for high-risk AI applications will depend on the competitive environment for the specific application and its technical specificities'.

Therefore, it is necessary to **consider other initiatives that can be used to reduce the costs faced by EU companies,** especially (but not exclusively) SMEs. Indeed, the AI act states that 'the framework will envisage specific measures supporting innovation, including regulatory sandboxes and specific measures supporting small-scale users and providers of high-risk AI systems to comply with the new rules'. In our opinion, one of the most appropriate measures would be for public institutions to make high-quality databases available to private agents. This would substantially reduce the expenditure associated with the review of these datasets and the corresponding certifications. Initiatives such as the European health data space 'will facilitate non-discriminatory access to health data and the training of artificial intelligence algorithms on those datasets, in a privacy-preserving, secure, timely, transparent and trustworthy manner, and with an appropriate institutional governance. Relevant competent authorities, including sectoral ones, providing or supporting the access to data may also support the provision of high-quality data for the training, validation and testing of AI systems'.[71] These initiatives could, of course, be complemented by specific subsidies aimed at helping companies to adapt to the new regulatory framework. The most appropriate approach would probably be a type of aid that would alleviate their substantial one-off cost for market entry.

[67] Renda, A. et al., Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe. FINAL REPORT (D5), April 2021, page 151.

[68] Commission Staff Working Document Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. SWD/2021/84 final

[69] Page 70.

[70] See article 53, AI Act.

[71] AI Act, whereas 45.

References

Agarwal, Alekh, et al. 'Fair Regression: Quantitative Definitions and Reduction-Based Algorithms.' Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019, pp. 120–29. proceedings.mlr.press, https://proceedings.mlr.press/v97/agarwal19d.html.

Aghaei, Sina, et al. 'Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making.' Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI Press, 2019, pp. 1418–26. ACM Digital Library, https://doi.org/10.1609/aaai.v33i01.33011418.

Baeza-Yates, Ricardo. 'Bias on the Web.' Communications of the ACM, vol. 61, no. 6, May 2018, pp. 54–61. DOI.org (Crossref), https://doi.org/10.1145/3209581.

Barocas, Solon, et al. Fairness and Machine Learning. fairmlbook.org, 2019, http://www.fairmlbook.org.

Beauchamp, Tom L., and James F. Childress. Principles of Biomedical Ethics. Oxford University Press, 2001.

Belavusau, U., & Henrard, K. (2019). A Bird's Eye View on EU Anti-Discrimination Law: The Impact of the 2000 Equality Directives. *German Law Journal, 20*(5), 614-636. doi:10.1017/glj.2019.53

Benjamins R. (2019). 'Is your AI system discriminating without knowing it?: The paradox between fairness and privacy'. Available at: https://business.blogthinkbig.com/is-your-ai-system-discriminating-without-knowing-it-the-paradox-between-fairness-and-privacy/

Berk, Richard, et al. 'Fairness in Criminal Justice Risk Assessments: The State of the Art.' Sociological Methods & Research, vol. 50, no. 1, Feb. 2021, pp. 3–44. SAGE Journals, https://doi.org/10.1177/0049124118782533.

Bickel, P. J., et al. 'Sex Bias in Graduate Admissions: Data from Berkeley: Measuring Bias Is Harder than Is Usually Assumed, and the Evidence Is Sometimes Contrary to Expectation.' Science, vol. 187, no. 4175, Feb. 1975, pp. 398–404. DOI.org (Crossref), https://doi.org/10.1126/science.187.4175.398.

Chouldechova, Alexandra. 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.' Big Data, vol. 5, no. 2, June 2017, pp. 153–63. liebertpub.com (Atypon), https://doi.org/10.1089/big.2016.0047.

Cirillo, Davide, et al. 'Sex and Gender Differences and Biases in Artificial Intelligence for Biomedicine and Healthcare.' Npj Digital Medicine, vol. 3, no. 1, 1, June 2020, pp. 1–11. www.nature.com, https://doi.org/10.1038/s41746-020-0288-5.

Corbett-Davies, Sam, et al. 'Algorithmic Decision Making and the Cost of Fairness.' Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, 2017, pp. 797–806. ACM Digital Library, https://doi.org/10.1145/3097983.3098095.

Costa-jussà, Marta R. 'An Analysis of Gender Bias Studies in Natural Language Processing.' Nature Machine Intelligence, vol. 1, no. 11, 11, Nov. 2019, pp. 495–96. www.nature.com, https://doi.org/10.1038/s42256-019-0105-5.

De Hert, P. and Lazcoz, G. (2022) 'When GDPR-principles blind each other. Accountability, not transparency, at the heart of algorithmic governance', European Data Protection Law Review, 8(5), pp.1-10.
Dolgin, Elie. 'AI Face-Scanning App Spots Signs of Rare Genetic Disorders.' Nature, Jan. 2019. www.nature.com, https://doi.org/10.1038/d41586-019-00027-x.

Dwork, Cynthia, et al. 'Fairness through Awareness.' Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Association for Computing Machinery, 2012, pp. 214–26. ACM Digital Library, https://doi.org/10.1145/2090236.2090255.

EDPB-EDPS Joint Opinion 5/2021on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)18 June 2021, at: https://edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf, accessed 23/03/2022

European Union Agency for Fundamental Rights. Data Quality and Artificial Intelligence :Mitigating Bias and Error to Protect Fundamental Rights. Publications Office, 2019. DOI.org (CSL JSON), https://data.europa.eu/doi/10.2811/615718.

Feldman, Michael, et al. 'Certifying and Removing Disparate Impact.' ArXiv:1412.3756 [Cs, Stat], July 2015. arXiv.org, http://arxiv.org/abs/1412.3756.

Floridi, Luciano, and Josh Cowls. 'A Unified Framework of Five Principles for AI in Society.' Harvard Data Science Review, vol. 1, no. 1, July 2019. hdsr.mitpress.mit.edu, https://doi.org/10.1162/99608f92.8cd550d1.

Focus, F. R. A. (2019), 'Data quality and artificial intelligence—Mitigating bias and error to protect fundamental rights.' *Eur. Union Agency Fundam. Rights, Vienna, Austria, Rep. TK-01-19-330-EN-N* p. 15

Friedman, Batya, and Helen Nissenbaum. 'Bias in Computer Systems.' ACM Transactions on Information Systems, vol. 14, no. 3, July 1996, pp. 330–47. July 1996, https://doi.org/10.1145/230538.230561.

Gebru, Timnit, et al. 'Datasheets for datasets.' Communications of the ACM 64.12 (2021): 86-92.

Gerards, J., Xenidis, R., Algorithmic discrimination in Europe: challenges and opportunities for gender equality and non-discrimination law, European Commission, Directorate-General for Justice and Consumers, Publications Office, 2021, https://data.europa.eu/doi/10.2838/77444

Gianclaudio Malgieri and Vincenzo Tiani, How the EU Council is rewriting the AI Act, REPORT - 6 December 2021, December 2021, at: https://brusselsprivacyhub.eu/publications/how-the-eu-council-is-rewriting-the-ai-act, last accessed 24/03/2022.

Gumbsch, Thomas, and Karsten Borgwardt. 'Ethnicity-Based Bias in Clinical Severity Scores.' The Lancet Digital Health, vol. 3, no. 4, Apr. 2021, pp. e209–10. www.thelancet.com, https://doi.org/10.1016/S2589-7500(21)00044-3.

Hardt, Moritz, et al. 'Equality of Opportunity in Supervised Learning.' Advances in Neural Information Processing Systems, edited by D. Lee et al., vol. 29, Curran Associates, Inc., 2016, https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Jiang, Liwei, et al. 'Delphi: Towards Machine Ethics and Norms.' ArXiv:2110.07574 [Cs], Oct. 2021. arXiv.org, http://arxiv.org/abs/2110.07574.

Kleinberg, Jon, et al. 'Inherent Trade-Offs in the Fair Determination of Risk Scores.' 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), edited by Christos H. Papadimitriou, vol. 67, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, p. 43:1-43:23. Dagstuhl Research Online Publication Server, https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

Kusner, Matt, et al. 'Counterfactual Fairness.' Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., 2017, pp. 4069–79.

Leslie, David, et al. 'Does 'AI' Stand for Augmenting Inequality in the Era of Covid-19 Healthcare?' BMJ, vol. 372, Mar. 2021, p. n304. www.bmj.com, https://doi.org/10.1136/bmj.n304.

Matias, Nathan, and Lucas Wright. Impact Assessment of Human-Algorithm Feedback Loops. Just Tech, Social Science Research Council, 1 Mar. 2022. DOI.org (Crossref), https://doi.org/10.35650/JT.3028.d.2022.

Mehrabi, Ninareh, et al. 'A Survey on Bias and Fairness in Machine Learning.' ACM Computing Surveys, vol. 54, no. 6, July 2021, p. 115:1-115:35. July 2022, https://doi.org/10.1145/3457607.

Miner, Adam S., et al. 'Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health.' JAMA Internal Medicine, vol. 176, no. 5, May 2016, pp. 619–25, https://doi.org/10.1001/jamainternmed.2016.0400.

Olfat, Mahbod, and Anil Aswani. 'Spectral Algorithms for Computing Fair Support Vector Machines.' Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR, 2018, pp. 1933–42. proceedings.mlr.press, https://proceedings.mlr.press/v84/olfat18a.html.

Prates, Marcelo O. R., et al. 'Assessing Gender Bias in Machine Translation: A Case Study with Google Translate.' Neural Computing and Applications, vol. 32, no. 10, May 2020, pp. 6363–81. Springer Link, https://doi.org/10.1007/s00521-019-04144-6.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

Scheuerman, Morgan Klaus, et al. 'How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services.' Proceedings of the ACM on Human-Computer Interaction, vol. 3, no. CSCW, Nov. 2019, p. 144:1-144:33. November 2019, https://doi.org/10.1145/3359246.

Schwartz, Reva, et al. A Proposal for Identifying and Managing Bias in Artificial Intelligence. National Institute of Standards and Technology, 22 June 2021. DOI.org (Crossref), https://doi.org/10.6028/NIST.SP.1270-draft.

Simoiu, Camelia, et al. 'The Problem of Infra-Marginality in Outcome Tests for Discrimination.' The Annals of Applied Statistics, vol. 11, no. 3, Sept. 2017, pp. 1193–216. Project Euclid, https://doi.org/10.1214/17-AOAS1058.

Srinivasan, Ramya, and Ajay Chander. 'Biases in AI Systems: A Survey for Practitioners.' Queue, vol. 19, no. 2, Apr. 2021, p. Pages 10:45-Pages 10:64. March-April 2021, https://doi.org/10.1145/3466132.3466134.

Stöger, K. et al. (2021) 'Legal aspects of data cleansing in medical AI', Computer Law & Security Review, 42, p. 105587. doi: https://doi.org/10.1016/j.clsr.2021.105587.

Tambiama Madiega Artificial intelligence act, 2021, at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdfj, accessed: 24/03/2022

Varshney, Kush R. Trustworthy Machine Learning. 2022, http://www.trustworthymachinelearning.com/.

Veale, M. and Binns, R. (2017) 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data', Big Data & Society, 4(2), pp. 1–17. doi: 10.1177/2053951717743530

Ward, Dan. Sampling Bias: Explaining Wide Variations in COVID-19 Case Fatality Rates. 2020. ResearchGate, https://doi.org/10.13140/RG.2.2.24953.62564/1.

Wiegand, Thomas, et al. 'WHO and ITU Establish Benchmarking Process for Artificial Intelligence in Health.' The Lancet, vol. 394, no. 10192, July 2019, pp. 9–11. www.thelancet.com, https://doi.org/10.1016/S0140-6736(19)30762-7.

Woodworth, Blake, et al. 'Learning Non-Discriminatory Predictors.' Proceedings of the 2017 Conference on Learning Theory, PMLR, 2017, pp. 1920–53. proceedings.mlr.press, https://proceedings.mlr.press/v65/woodworth17a.html.

Zuiderveen Borgesius, F. J. (2020) 'Strengthening legal protection against discrimination by algorithms and artificial intelligence', The International Journal of Human Rights. Routledge, 24(10), pp. 1–22. doi: 10.1080/13642987.2020.1743976.

OA-08-22-267-EN-N

Biases are commonly considered one of the most detrimental effects of artificial intelligence (AI) use. The EU is therefore committed to reducing their incidence as much as possible. However, the existence of biases pre-dates the creation of AI tools. All human societies are biased – AI only reproduces what we are. Therefore, opposing this technology for this reason would simply hide discrimination and not prevent it. It is up to human supervision to use all available means – which are many – to mitigate its biases. It is likely that at some point in the future, recommendations made by an AI mechanism will contain less bias than those made by human beings. Unlike humans, AI can be reviewed and its flaws corrected on a consistent basis. Ultimately, AI could serve to build fairer, less biased societies.

This study begins by providing an overview of biases in the context of artificial intelligence, and more specifically to machine-learning applications. The second part is devoted to the analysis of biases from a legal point of view. The analysis shows that shortcomings in this area call for the implementation of additional regulatory tools to adequately address the issue of bias. Finally, this study puts forward several policy options in response to the challenges identified.

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

ISBN 978-92-846-9681-9 | doi: 10.2861/98930 | QA-08-22-267-EN-N