

Artificial Intelligence and Algorithms in Risk Assessment

*Addressing Bias,
Discrimination and other Legal
and Ethical Issues*

A Handbook

© European Labour Authority, 2024

Reproduction is authorised provided the source is acknowledged.

For any use or reproduction of photos or other material that is not under the copyright of the European Labour Authority, permission must be sought directly from the copyright holders.

Neither the European Labour Authority nor any person acting on behalf of the European Labour Authority is responsible for the use which might be made of the following information.

This document has been produced and updated by Véronique Bruggeman, Marco Paron Trivellato, Maxime Moulac (Milieu Consulting SRL), Prof. Raphaële Xenidis and Prof. Benjamin van Giffen as authors. This task has been carried out exclusively by the authors in the context of a contract between the European Labour Authority and Milieu Consulting SRL, awarded following a tender procedure. The document has been prepared for the European Labour Authority; however, it reflects the views of the authors only. The information contained in this report does not reflect the views or the official position of the European Labour Authority. This is an updated version from July 2024.

The handbook is based on a training session, facilitated by Milieu Consulting, that took place on 26 May 2023 and was headed by Prof. Raphaële Xenidis and Prof. Benjamin van Giffen.

Contents

List of tables	4
List of figures.....	4
Abbreviations	5
Introduction	6
1.0 Algorithms, automation and AI.....	7
1.1 Defining key concepts.....	7
1.2 How do they work?	8
1.3 The CRISP-DM model	9
1.4 How can AI and algorithms discriminate?	11
2.0 Machine learning bias	14
2.1 Overview of machine learning biases	14
2.2 Other examples of AI bias	17
3.0 The legal framework	19
3.1 Anti-discrimination law: The fundamental right to non-discrimination in Europe.....	19
3.2 Applying anti-discrimination law.....	21
3.2.1 When is bias unlawful discrimination?.....	21
3.2.2 Fairness and bias versus equality and discrimination	25
3.2.3 Gaps in and limitations of the legal scope	25
3.3 European AI Sectoral Regulation	28
3.3.1 EU AI Act	28
3.3.2 AI Liability Directive (proposed).....	33
3.3.3 Council of Europe Framework Convention on AI	34
3.4 Interaction with data protection law: taking stock of European developments.....	34
3.4.1 EU General Data Protection Regulation.....	34
3.4.2 Council of Europe Convention 108+	36
4.0 Mitigation framework.....	37

5.0 Key ethical requirements. Beyond the law, what ethical requirements can support non-discriminatory AI?	44
5.1 Ethics Guidelines for Trustworthy AI	45
5.2 The five converging ethics principles	47
5.3 Applying ethics principles	50
List of References	52

List of tables

Table 1: Overview of ML biases, with examples of Real-world scenario 1 (literally taken from van Giffen et al. 2022)	15
Table 2: EU primary anti-discrimination law	19
Table 3: EU secondary anti-discrimination law	20
Table 4: Direct versus indirect discrimination in the framework of AI/ML	23
Table 5: Overview of 24 mitigation methods for addressing biases within the CRISP-DM process phase	38
Table 6: Why and how to use the seven key mitigation methods (extended from van Giffen et al. 2022)	39
Table 7: Explanation of the five ethics principles	48

List of figures

Figure 1: Artificial Intelligence vs. Machine Learning vs. Deep Learning vs. Data Science	7
Figure 2: Conceptual approach: using machine learning for predictions in real-world applications	8
Figure 3: Process phases of the CRISP-DM model	10
Figure 4: When is algorithmic bias unlawful discrimination?	21
Figure 5: The risk-based approach in the EU AI Act	28
Figure 6: AI Fairness checklist	43
Figure 7: Geographical distribution of ethical AI guidelines	45
Figure 8: The three components of trustworthy AI	46
Figure 9: The five converging ethics principles	47

Abbreviations

AI	Artificial Intelligence
AI HLEG	High-Level Expert Group on Artificial Intelligence
CAHAI	Council of Europe's Ad Hoc Committee on Artificial Intelligence
CJEU	Court of Justice of the European Union
CRISP-DM	Cross-Industry-Standard-Process-for-Data-Mining
DL	Deep Learning
ECHR	European Convention on Human Rights
EU	European Union
FRAIA	Fundamental Rights and Algorithms Impact Assessment
GDPR	General Data Protection Regulation
ML	Machine Learning
TEU	Treaty on European Union
TFEU	Treaty on the Functioning of the European Union

Introduction

Automation, rule-based models and Artificial Intelligence (AI) systems are already used for risk assessment in many Member States, contributing to focused inspections and tackling fraud in the domain of labour mobility and social security. However, like humans, algorithms, and in particular machine learning (ML) algorithms, are vulnerable to biases that can make their predictions unfair and/or discriminatory – as the Dutch child benefit scandal has indeed already proven.

As part of ELA's support for national competent authorities and experts from the Member States in the domain of risk assessment, it is important to understand the biases and other legal and ethical issues involved in developing and using algorithms, automation or AI for risk assessment in the field of labour mobility and social security (for data analysis, risk assessment, data matching and data mining, etc.).

This handbook has been prepared as a follow-up to a training session, organised on 26 May 2023, that focused on sharing knowledge and experience regarding practical, legal and ethical issues concerning the use of machine-learning tools embedded in AI. Additionally, an overview of the applicable EU and ECHR legal framework pertaining to equal treatment and non-discrimination was provided. The aim of the handbook is to provide support to the development of new knowledge and competences in the field through the following key objectives:

- 1) Understand the types of bias involved both when developing risk assessment tools in the field of labour mobility and social security, and when utilising them;
- 2) Understand the legal, practical and ethical issues concerning the use of algorithms, automation (including rule-based models) or Artificial Intelligence (AI) for risk assessment;
- 3) Provide an overview of equal treatment and non-discrimination legislation applicable at EU and ECHR level and relevant for the use of artificial intelligence or other algorithmic solutions, and discuss the consequences of non-compliance with the applicable legal framework;
- 4) Provide overview knowledge about methods to avoid and mitigate the biases and to eliminate discrimination in the use of algorithmic, automated or AI processes by analysts and legal professionals;
- 5) Illustrate the theory with practice-oriented case studies and examples.

This practical handbook is structured as follows. **Section 1** presents a brief introduction on the basic functioning of, and key terminology related to algorithms, automation and artificial intelligence. Algorithms can easily be biased or develop bias over time, which then can lead to discrimination. Biases can also amplify discrimination because of feedback loops and redundant encoding. The concept of 'bias' and the different types of bias will be elaborated in **Section 2**, while the complexity of bias in algorithmic decision-making will be shown through concrete examples. More specifically, bias is analysed in the context of discrimination (as a legal and normative concept). Discrimination can be linked to prejudices and structural inequalities enshrined in data but may also be the result of biased algorithmic models, interpretations or deployment. Discrimination can also take different forms. In this context, understanding the risks and the types of legal challenges they create is key to ensuring equality and combating discrimination. **Section 3** will therefore provide an overview of how the current gender equality and non-discrimination legislative framework in place in the EU (and internationally, within the Council of Europe) captures and redresses algorithmic discrimination. **Sections 4 and 5** will further focus on mitigation methods for addressing the aforementioned biases and on the key ethical requirements that need to be ensured. This includes the provision of examples of good practice (legal and non-legal solutions) for legal compliance with gender equality law and general non-discrimination law.

1.0 Algorithms, automation and AI

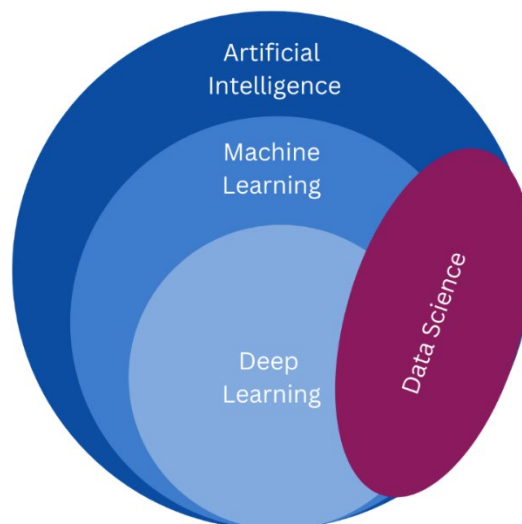
1.1 Defining key concepts

This section defines an initial set of concepts that are needed to navigate the areas of algorithms, automation, and Artificial Intelligence (AI). The first two key definitions in this context are those relating to automation and algorithm.

Automation is defined as *the creation and application of technologies to produce and deliver goods and services with minimal human intervention*. The implementation of automation technologies, techniques and processes improves the efficiency, reliability, and/or speed of many tasks that were previously performed by humans. An **algorithm**, on the other hand, is *a procedure used for solving a problem or performing a computation*. Algorithms act as an exact list of instructions that conduct specified actions step by step in either hardware- or software-based routines.

In the context of automation and algorithms, four different domains are relevant. **Artificial Intelligence (AI)** is *a branch of computer science that deals with the automation of intelligent behaviour*. It is a very wide definition, and very often used as an umbrella definition including machine learning and deep learning.

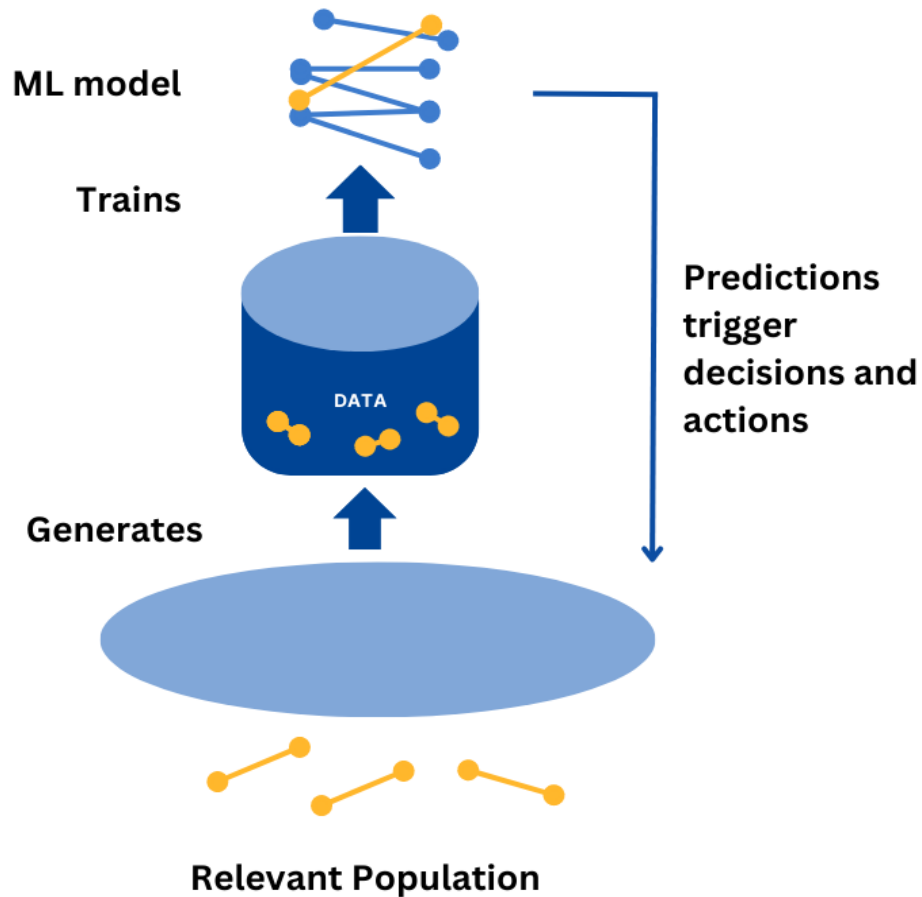
Figure 1: Artificial Intelligence vs. Machine Learning vs. Deep Learning vs. Data Science



Machine Learning (ML) describes *the processes in which algorithms learn patterns from data*. Machines have sample data that they can learn to propose new solutions. Machines learn from examples and, after a learning phase, can generalise and propose (new) solutions (make new predictions). Machine Learning is a process that introduces an aspect of innovation, moving beyond a deterministic learning process but instead using data to learn new deterministic and probabilistic systems. This is of course a source of new potential, but also a threat for discrimination. **Deep Learning (DL)** instead represents *special procedures of machine learning, in which neural networks are trained with data*. **Data science** is the field of study that combines this domain expertise.

1.2 How do they work?

Figure 2: Conceptual approach: using machine learning for predictions in real-world applications



Source: van Giffen, B., Herhausen, D., & Fahse, T. (May 2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, Vol. 144, pages 93-106, available at: <https://www.sciencedirect.com/science/article/pii/S0148296322000881>.

The figure above summarises the typical application logic of ML in marketing. Data points are generated or extracted from the relevant population which are then used to train a predefined ML model that, once completed, can be used to make predictions which trigger marketing decisions and actions.

Example 1: Netflix and ML

For example, Netflix generates data from the viewing behaviour of all its customers, uses this data to train a recommendation algorithm, whose predictions then trigger individual movie and series recommendations for all its customers.

Example 2: AI / ML applications for risk assessment

There are several potential examples of AI and ML applications for risk assessment. The first concerns the profiling of jobseekers using machine learning to predict risk of unemployment. The so-called job seekers profiling is based on a logic of optimisation of resources, whereby candidates are profiled based on a data set that could potentially be affected by biases that are present in our society, penalising vulnerable groups who have more difficulties in finding employment.

Another one could be deployed to predicting the likelihood of labour contract violations.

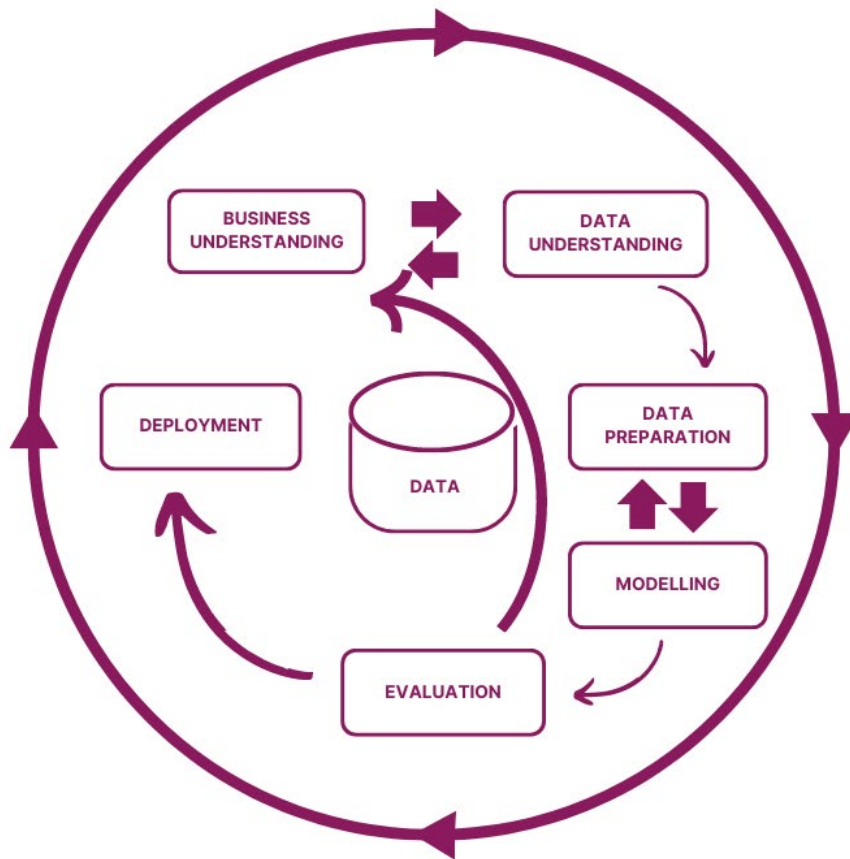
Potential challenges relate to the non-reiteration of positions of disadvantage or discrimination of people on the basis of race, gender, ethnicity, etc. as well as the identification of under-represented groups.

1.3 The CRISP-DM model

The CRISP-DM model is a standardised process for the development of AI/ML applications.¹ It is a standard process in which an understanding of the data is created to solve the problem and to make a prediction for the machine learning process.

¹ The abbreviation stands for Cross-Industry-Standard-Process-for-Data-Mining. Published in 1999 to standardize data mining processes across industries, the CRISP-DM model has since become the most common process model for data mining, data analytics, and data science projects in various industries. See: Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R., "CRISP-DM 1.0 step-by-step data mining guide," 2000.

Figure 3: Process phases of the CRISP-DM model



Source: Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R., "CRISP-DM 1.0 step-by-step data mining guide," 2000.

Figure 3 displays the six phases of the CRISP-DM process model that can be used to plan, organise, and implement an ML project:

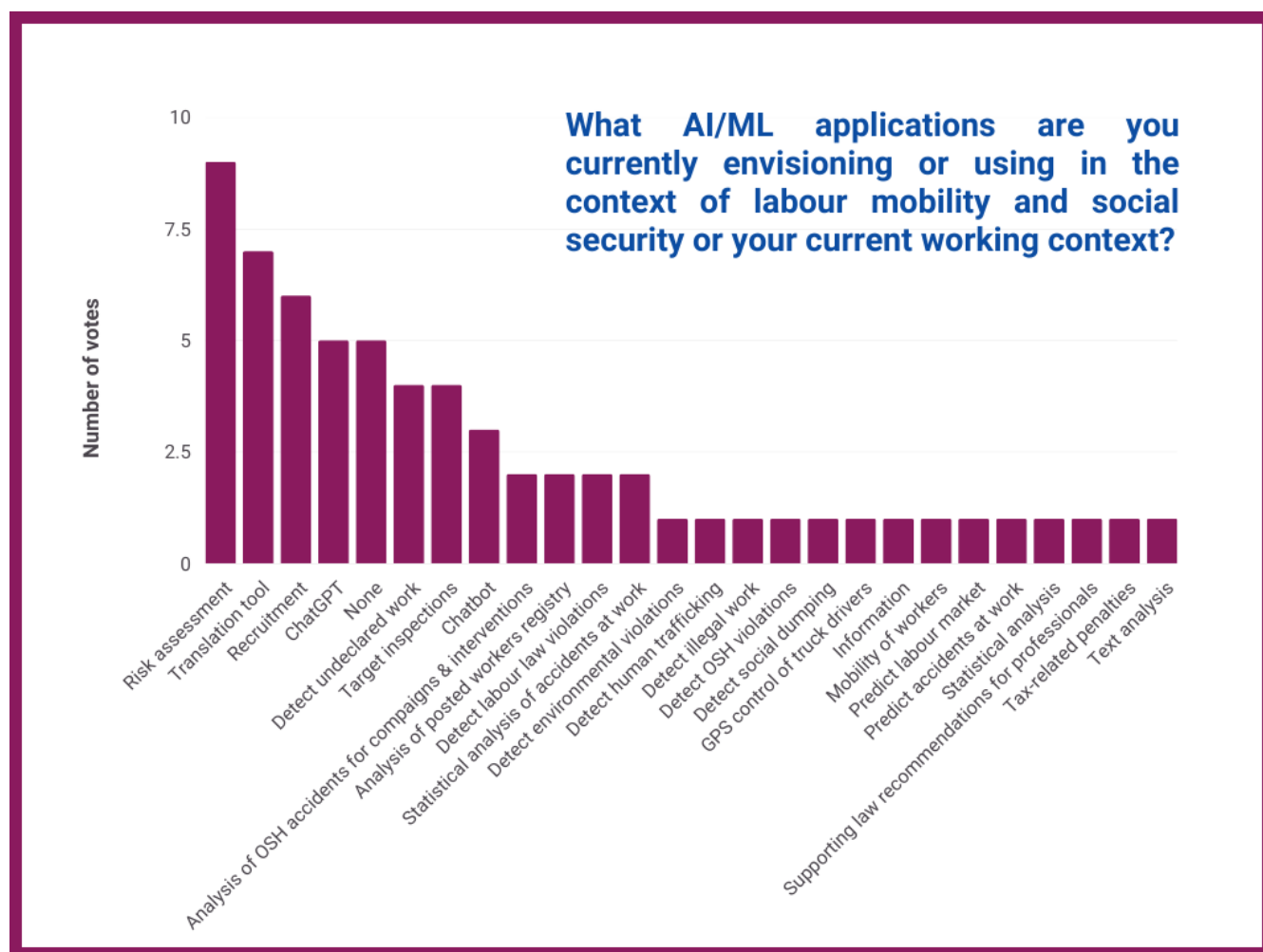
- 1) The **business understanding** phase focuses on understanding the problem and defining the project goals.
- 2) The **data understanding** phase starts with initial data collection and proceeds with evaluating the data needed to solve the problem.
- 3) The **data preparation** phase prepares the data (i.e. constructs the final dataset from the initial raw data) for use in modelling.
- 4) Several ML techniques to solve the problem are then selected and developed in the **modelling** phase.
- 5) The **evaluation** phase evaluates the performance of the models and selects the best one.
- 6) The **deployment** phase implements the selected model in production.

The performance of the model needs to be monitored continuously so that it can be refined as needed.



Several modules are tried to optimise stochastically, as the prediction performance needs to be evaluated, and thus an understanding of the problem is required.

Participant input 1: What AI/ML applications are you currently envisioning or using in the context of labour mobility and social security?



1.4 How can AI and algorithms discriminate?

Examples of algorithmic bias giving rise to discrimination regularly make media headlines: face recognition applications underperforming for female faces from racialised groups, CV-screening tools excluding female applicants, risk assessment software systematically flagging citizens with migration background, etc. Part of the problem of **'algorithmic discrimination'** is that decision-making machine learning algorithms rely on data about the past to make future predictions. Such predictions could amplify the main types of discrimination that have existed in human decision-making in the past (and thus in the datasets used to train algorithms), such as race or gender stereotypes about who is best suited to certain types of work. This mechanism is referred to as **'garbage in garbage out'**, which means that if the data is biased, it is very likely that the system's output will exhibit biases too. Yet data is not the only source of algorithmic discrimination. Biased data collection, curation, labelling, modelling and problem framing, and biased interpretations of algorithmic output could also lead to algorithmic discrimination. Importantly, human and machine bias interact in the socio-technical systems which algorithms are parts of.

Biases may then continue in the implementation of the system through so-called **feedback loops**. For example, a system used to predict which areas of a city are particularly at risk of criminality could rely on past crime data and this might embed prejudice against ethnic minorities or disadvantaged groups. In turn, the system predictions could lead to increased police deployments and controls in that area. Over-surveillance would then confirm the predicted higher crime rates, which leads to reinforcement loops in that area, thereby further fuelling prejudice against racialised and socio-economically disadvantaged population groups. Importantly, it is not enough to make a system 'blind' to certain characteristics such as race or gender because these characteristics are encoded in multiple other data points, for example in the syntax and language used in CVs and application letters, names, career breaks, tastes, etc. The problem of 'redundant encoding' often gives rise to proxy discrimination against minority and disadvantaged groups.

Example 3: Examples of harm arising from algorithmic bias

Job searching platform

A first example concerns a matching platform where job seekers can input search criteria and consult job offers. Results can be different based on gender language in the search query. For example, the job searching platform offers more results when the search term is “chercheur” (the masculine term for “researcher” in French) compared to the search term “chercheuse” (the feminine term for “researcher” in French). This anecdotal example demonstrates the kind of impact that a biased distribution of valuable information can have on the real world.

AMS algorithm developed by Austria

A second example concerns a job seeker profiling system developed in Austria. The system divides job seekers into three categories: Group A - High prospects to find employment in the short term; Group B - Mediocre prospects, meaning job seekers not part of groups A or C; Group C - Low prospects to find employment in the long-term. Austria aimed to streamline resource allocation, make job search assistance more efficient by targeting Group B – given that group A was perceived as not problematic, and that group C was considered as entailing a too high cost compared to the benefits. The system, however, incorporated and affected a negative weight to data such as gender and migration background to 'reflect the harsh reality of the labour market'.² This can create a negative loop whereby discrimination patterns in society are encoded in the distribution of valuable resources such as training or support for labour market integration.

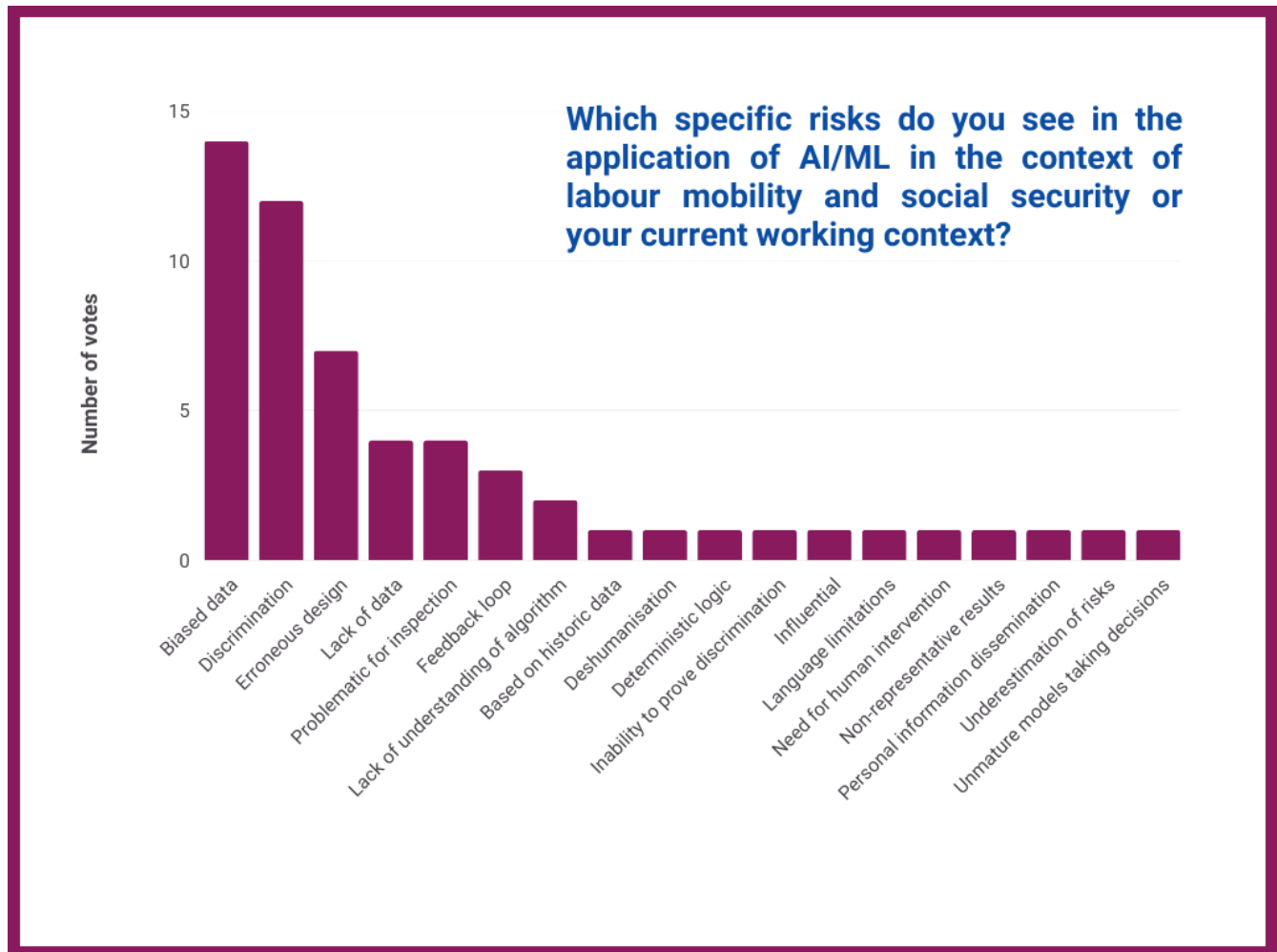
Fraud detection systems

The Dutch tax authorities introduced in 2013 an algorithmic system aiming to detect potentially fraudulent applications for child benefits. The system encoded non-Dutch citizenship as a higher risk factor received a higher risk score. For this reason, non-Dutch parents flagged by the system had their benefits suspended and were subjected to investigations and benefit recovery policies. This led to financial problems for affected families as well as cases of mental health problems and stress. This example shows how the system design strengthened existing institutional prejudices about the link between race, ethnicity and crime.³

² Allhutter, D., Cech, F., Fischer, F., Grill, G., & Mager, A. (2020). Algorithmic profiling of job seekers in Austria: how austerity politics are made effective. *Frontiers in Big Data*, 5.

³ “Dutch scandal serves as a warning for Europe over risks of using algorithms”, available at: <https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/> . See also: <https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/> and https://www.europarl.europa.eu/doceo/document/O-9-2022-000028_EN.html

Participant input 2: Which specific risks do you see in the application of AI/ML in the context of labour mobility and social security or your current working context?



It is important to note that not all kinds of bias are considered discriminatory from a legal point of view and therefore legally prohibited. European fundamental rights law, anti-discrimination law, consumer protection law, data protection and AI sectoral regulation offer tools to enforce the protection against (algorithmic) discrimination.

2.0 Machine learning bias

2.1 Overview of machine learning biases

Machine learning bias describes an unintended or potentially harmful property of data or a model that results in a systematic deviation of algorithmic results. **Bias** can then be defined as unwanted effects or results which are evoked through a series of choices and practices in the machine learning developing process.

There are different channels and stages in which biases can be channelled into a system. Biases can already arise in data collection: there is a lot of room for biases to enter the system, e.g. because the data collection is not sufficiently representative, or because the data curation presents gender stereotypes. Biases can also arise when designing a module or problem. The questions the project team asks can generate biases. If a system is developed to select candidates with the highest leadership rates, by now it is known that this category favours men over women, because the quality of leadership usually prefers men over women. Several researchers also pointed out that with the same algorithm, the interpretation of the results could also be differentiated. The output could in fact be in favour of dominant groups and not in favour of racialised groups.

Leading technology outlets consider AI and machine learning bias as a fundamental, difficult and unresolved problem. The research question therefore is: *what types of bias emerge in machine learning projects and how can they be mitigated?*

Prof. van Giffen and his colleagues conducted a systematic, problem-centred literature review which integrates existing knowledge about ML biases and mitigation strategies into the CRISP-DM model.⁴ They ended up in coding biases into eight distinct categories and used a real-life case study to provide relevant examples of each type of bias.

Real-world scenario 1. Two bakeries (excerpt from van Giffen et al. 2022)

The company of interest is a nationwide bakery chain, with a central production, and multiple bakeries in city centres, train stations and villages. The company uses ML models for decision-making regarding demand forecasting, promotions and campaigning, new product development, and their loyalty program.

The case study relates to two bakeries within this bakery chain, that operate in a different context: one is situated in a train station of a big city and the other one is placed in a small city centre. The customers are consequently different: one has villagers, while the other in the train station is mainly visited by travellers. The two bakeries thus have a very different purchasing pattern. One centralised production facility supplies both bakeries.

The case study is not focusing on racial or ethnical discrimination, but just on the adverse economic effects of bias, and on the mitigation measures that could be put in place. *Please bear in mind that this example has been simplified for illustrative purposes.*

⁴ For the full methodology and results, see: van Giffen et al., "Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods", Journal of Business Research, Vol. 144, May 2022, Pages 93-106, available at: <https://www.sciencedirect.com/science/article/pii/S0148296322000881>

Table 1: Overview of ML biases, with examples of Real-world scenario 1 (literally taken from van Giffen et al. 2022)

Bias	Definition	Example
Social bias	Available data reflects existing bias in human society prior to the creation of the model. This is mainly “Garbage in garbage out”, i.e. already existing bias replicated within the system and thus reinforced.	<p><u>A customer reward initiative that ignores loyal customers</u></p> <p>The bakery implemented a customer reward initiative that ignores loyal customers. A social bias exists when the available data reflect existing biases in the perception of loyalty. In this case, they want to reward loyal customers with vouchers. Voucher recipients are identified by average spending and frequency of visits. Students go there very often, because their school is close to the bakery. But they do not spend much, so the customers do not appear as relevant and are not recognised in the data. And they don't get a voucher. This replicates an existing prejudice, because we recognise the family man as a very loyal customer, and not the students who have less money, but who go to the bakery very often. We neglect that group of people who do not spend much. This is a social prejudice that we replicate in the model.</p> <p>=> Social bias occurs when available data mirrors existing biases among customer in loyalty perceptions</p>
Measurement bias	Chosen features and labels are imperfect proxies for the real variables of interest. So, the wrong measurement is in place.	<p><u>Capturing the effects of good weather</u></p> <p>Baking cakes for a bakery is expensive. But it is even more expensive if these cakes are not sold. The idea is to prepare a model that helps us predict the right amount of cakes. We use for example, the volume of previous sales, a 30-day moving average, precipitation, and temperature. However, precipitation is not a good proxy for good weather, imagine if people wanted to stay home with a slice of cake.</p> <p>=> Rainfall was an unsuitable proxy for good weather to predict the demand for cakes because the key drivers is not the amount of rain, but the temperature.</p>
Representation bias	The input data is not representative for the real world which leads to systematic errors in model predictions.	<p><u>Using apples to predict oranges</u></p> <p>Let's roll out our “successful” prediction model to other bakeries. The sales data of the city centre locations is extracted for training a model that is deployed in the train station locations. The data that is generated in the train station is not used in the prediction.</p> <p>=> Representation bias emerges if the probability distribution of the development population differs from the true underlying distribution.</p>

Bias	Definition	Example
Label bias	Labelled data systematically deviates from the underlying truth. ⁵	<p><u>Old wine in new bottles</u></p> <p>The bakery has a product relaunch: the 'Wheat Bread' is now called the new 'Wheat fitness bread', i.e. the same bread with a different name. If the bakery staff does not standardise the name of the bread (of the old and new versions) in the data set, it means that the historical data are not consistent with the newly generated data. This causes a problem in the way the data is labelled, leading to the creation of an unreliable data set. => Label bias arises when training data is assigned to wrong class labels. In this case, due to the change in product label, the newly generated data does not match with the historical data that serves as input to the forecast model.</p>
Algorithmic bias	Inappropriate technical considerations during modelling lead to systemic deviation of the outcome.	<p><u>As simple as possible, but not simpler</u></p> <p>If too simple an algorithm is developed that is unable to capture the factors that determine variations and demands, the algorithm will not be flexible enough to successfully predict the distribution of the real world.</p> <p>=> Algorithmic bias is introduced during the modelling phase and results from inappropriate technical considerations.</p>
Evaluation bias	A non-representative testing population or inappropriate performance metrics are used to evaluate the model.	<p><u>Benchmarking with caution</u></p> <p>Normally 80 per cent of the data is used to learn the model and once the model is learned you take the remaining 20 per cent of the data and try to benchmark it. Evaluation bias can occur if the population of the benchmark dataset is not representative of the usage population. The algorithmic model is trained and optimised on the proprietary bakery data but is evaluated on a (non-relevant) publicly available benchmark dataset.</p>
Deployment bias	The model is used, interpreted and deployed in a different context than it was built for.	<p><u>Stick to the knitting</u></p> <p>Coffee is correlated with high (but not random) average expenditure. For customers who buy coffee, therefore, the model is likely to recommend handing out a voucher. Deployment bias occurs when the ML model is inappropriately used or interpreted (even if no other bias is present) due to e.g. human intervention: the manager assumes a causal relationship between coffee and average expenditure and therefore uses the model to</p>

⁵ It is noted that the term "label" as used here is not necessarily the same as the term "target" variable class label, as typically used in predictive modelling language.

Bias	Definition	Example
		inappropriately justify the distribution of free coffee to customers to stimulate spending.
Feedback bias	The outcome of the model influences the training data such that a small bias can be reinforced by a feedback loop.	<p><u>Mind the power of the algorithm</u></p> <p>A competing bakery is basically incentivising its customers to rate five stars on Google Maps. The competing bakery says: "If you rate five stars, a get a free coffee". This causes more customers to go to that bakery and more customers to rate 5 stars. The stimulation of 5-star ratings during customer visits manipulates the ranking of the competitor, which again triggers more customer visits and hence create a reinforcing feedback loop. It is not possible to close the gap among the two bakeries. The same applies when policing. When you police a neighbourhood, there are more arrests. One has the idea that the crime rate in that neighbourhood is very high. You overestimate the effect of the amount of crime in that area.</p> <p>=> Feedback bias can emerge when the output of the ML model influences features that are used as new inputs. An initially small bias is potentially reinforced through a feedback loop.</p>

2.2 Other examples of AI bias

Apple and Goldman Sachs

In 2019, Apple and Goldman Sachs launched the "Apple Card". The credit limit was developed using machine learning methods. The feature "gender" had been given a powerful prediction power for creditworthiness. It was then found that higher credit limits were granted to men than to women despite the latter having higher credit scores.⁶ We have an algorithm that is efficiently allocating credit limits, but its predictions are socially not desirable. In this case, the following AI biases have likely caused the flawed credit limit allocation:

- **Social bias:** Available data of creditworthiness might reflect such a bias in human society, that has not been identified in the creation of the model. For example, a company might use the data that it has available opportunistically without being aware how its subsequent use might reinforce social biases that are reflected in the data.
- **Evaluation bias:** The input data is not representative for the real world which leads to systematic errors in model predictions. In this case, the data set of Goldman Sachs (predominantly males) might not have been adequately representative of the relevant target population for its banking product.

⁶ Gupta, AH (2019), 'Are Algorithms Sexist?', The New York Times (15 November), available at: www.nytimes.com/2019/11/15/us/apple-card-goldman-sachs.html

- **Feedback bias:** The outcome of the model influences the training data such that a small bias can be reinforced by a feedback loop.

Amazon's algorithmic hiring prototype

Amazon employed a machine learning algorithm to filter applicants with tragic consequences given that the recruitment algorithm filtered out female applicants.⁷ This was caused by the fact that the AI was trained with CVs of the last 10 years that were easily available, but that predominantly consisted of male. In this case, data used to train the algorithm is outdated and distorts the predictions of the algorithm. In this case, the following AI biases have likely caused the flawed hiring process:

- **Social bias:** Because Amazon has hired mostly man, the data from CVs represented mostly male applications, which is not in line with the socially desired distribution of equality. The available data reflects existing bias in human society prior to the creation of the model.
- **Representation bias:** The input data is not representative for the real world (workforce outside of Amazon) which can further lead to systematic errors in model predictions.
- **Feedback bias:** The outcome of the model can also deteriorate over time when a bias with an originally small effect is reinforced by a feedback loop.



Despite the undoubted benefits of AI / ML, AI bias occurs easily, mostly unintentionally, and it is most often difficult to spot and distinguish.

⁷ Dastin, J (2018), 'Amazon scraps secret AI recruiting tool that showed bias against women' (10 October), available at: www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

3.0 The legal framework

The section aims to analyse the non-discrimination data protection and AI-specific legal framework at European level. After highlighting how and why algorithmic discrimination is an issue of EU non-discrimination law, this chapter assesses to what extent the legal framework in place is fit for purpose, and where the gaps and challenges lie.

3.1 Anti-discrimination law: The fundamental right to non-discrimination in Europe



This handbook does not provide legal advice or comprehensive legal guidance. National legal frameworks and instruments, which should be taken into account when developing a system, fall outside the scope of this handbook. Please be aware that national law may differ from EU provisions. In the case of discrimination for example, EU law sets minimum requirements and Member States are free to adopt more protective legislation as long as they comply with the EU treaties.

Treaties are the starting point for considering measures under EU law. As part of **primary law**, they set the framework for the EU's actions in specific fields of competence. The body of law that derives from the principles and objectives of the treaties is known as **secondary law**. The EU's legislation includes regulations, directives, decisions, recommendations and opinions.

With regard to non-discrimination, the following key EU primary and secondary law provisions are of the utmost importance:

Table 2: EU primary anti-discrimination law

EU Primary Law	
Article 19 Treaty on the Functioning of the European Union (TFEU)	Allows the Council to adopt legislation to combat discrimination in relation to six characteristics, namely sex, racial or ethnic origin, religion or belief, disability, age, sexual orientation.
Article 157 TFEU	Guarantees equality between men and women at work and in pay.
Article 21 of the Charter of Fundamental Rights	<p>Non-discrimination</p> <p>1. <i>Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.</i></p> <p>2. <i>Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited.</i></p>
Article 23 of the Charter of Fundamental Rights	<p>Equality between women and men</p> <p><i>Equality between women and men must be ensured in all areas, including employment, work and pay.</i></p>

EU Primary Law	
	<i>The principle of equality shall not prevent the maintenance or adoption of measures providing for specific advantages in favour of the under-represented sex.</i>

Table 3: EU secondary anti-discrimination law

EU Secondary Law – i.e. minimum requirements in Directives ⁸		
	Ground	Scope of application
Council Directive 2000/43/EC implementing the principle of equal treatment between persons irrespective of racial or ethnic origin	race or ethnic origin	employment, goods and services, education, social protection, including social security and healthcare, and social benefits
Council Directive 2000/78/EC establishing a general framework for equal treatment in employment and occupation	age, religion or belief, disability, sexual orientation	employment
Council Directive 2004/113/EC implementing the principle of equal treatment between men and women in the access to and supply of goods and services	sex	goods and services
Directive 2006/54/EC on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation	sex	employment

At the Council of Europe level, **Article 14** of the European Convention on Human Rights (**ECHR**) that applies to all EU 27 Member States and to 19 other countries, affirms that: *The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.*

The enforcement of the ban on discrimination in EU law and the ECHR is overseen respectively by the Court of Justice of the European Union (CJEU) in Luxembourg and the European Court on Human Rights (ECtHR) in Strasbourg.

⁸ A directive is a legal act adopted by the EU institutions directed to all Member States and it is binding as to the result to be achieved. It is up to the single Member States to determine the form and methods to transpose in its legal framework law. The national authorities must notify the European Commission of the measures taken. See: <https://eur-lex.europa.eu/EN/legal-content/glossary/directive.html>

3.2 Applying anti-discrimination law

3.2.1 When is bias unlawful discrimination?

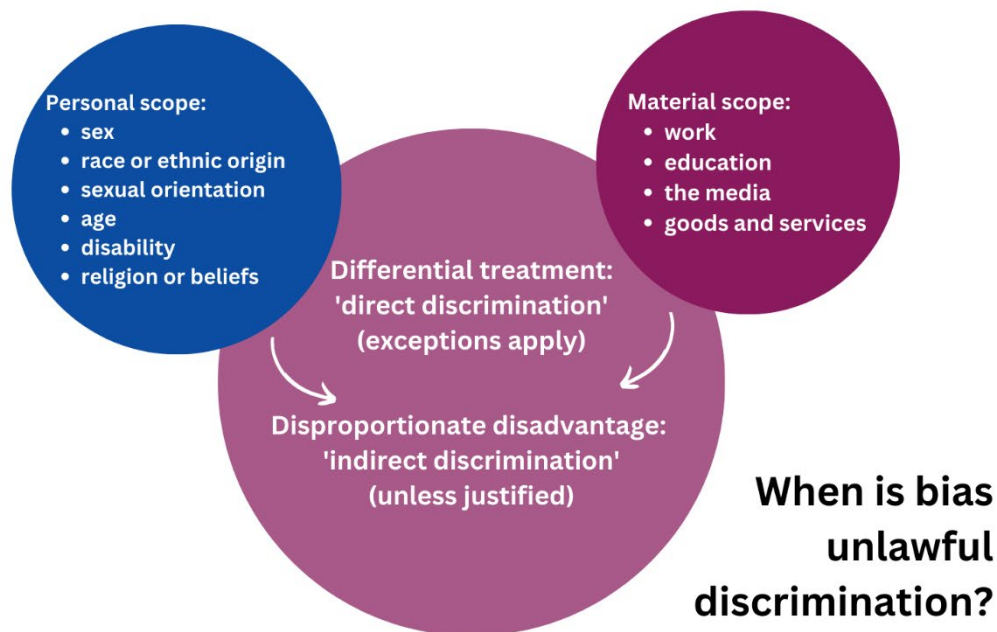


Figure 4: When is algorithmic bias unlawful discrimination?

There are three conditions for algorithmic bias to fall within the scope of **unlawful discrimination**. The first condition is linked to the personal scope of anti-discrimination law. Algorithmic bias must either harm a protected group or result in the unfavourable treatment of people based on a protected ground, i.e. sex, racial or ethnic origin, sexual orientation, age, disability and religion or belief. Some EU Member States have gone beyond this personal scope and ban discrimination on a broader basis. The second condition is that algorithmic bias falls within the material scope of anti-discrimination law. Algorithmic bias will only be legally prohibited if it falls within the scope of application of the law, i.e. the fields of work and employment, education, the media and the sale and consumption of goods and services. The third condition in order for bias to constitute unlawful discrimination is that it must fall under one or two definitions: either it must result in a form of differential treatment, or it must create a disproportionate disadvantage for a protected group. Note that the directives are addressed to Member States that must transpose their provisions into national law, which then applies to individuals, private entities such as companies, and public bodies such as state authorities. However, the CJEU has interpreted the prohibition of discrimination as directly applicable in disputes between private parties, including individuals and companies.⁹

- **Direct discrimination** is defined in EU law as a situation in which ‘one person is treated less favourably than another is [...] in a comparable situation on any of the [protected] grounds’ [defined in the relevant directives].¹⁰ Direct discrimination focuses on unfavourable treatment or differential treatment and captures situations in

⁹ See e.g. C-414/16 - Egenberger (Judgment of the Court (Grand Chamber) of 17 April 2018, Vera Egenberger v Evangelisches Werk für Diakonie und Entwicklung e.V., ECLI:EU:C:2018:257).

¹⁰ See, e.g., 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.

which a decision is made taking into consideration a protected ground, to the disadvantage of the person or group of persons related to that protected ground.

Example 4: A concrete example of direct discrimination

The police of an EU Member State open a call for recruitment of new staff. The vacancy is only open to male candidates, due to their physical characteristics. This is a form of direct discrimination, because it directly mentions the protected ground of sex.

- **Indirect discrimination** refers to situations ‘*where an apparently neutral provision, criterion or practice would put [members of a protected category] at a particular disadvantage compared with other persons, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary*’.¹¹ Instead of focusing on the unfavourable treatment of given groups and individuals because of a given protected ground, the notion of indirect discrimination places the focus on the disadvantageous effects of any given – apparently neutral – practice or measure.

Example 5: A concrete example of indirect discrimination

The police of an EU Member State open a call for recruitment of new staff. The vacancy requires that the candidates for a post as police officer are at least 170 cm tall. In this case, it is not direct discrimination because protected grounds are not explicitly mentioned, but in practice it creates a disproportionate disadvantage for women. As a result of the call, fewer women than men will be able to apply.

The distinction between direct and indirect discrimination has important legal consequences. There is in principle no justification for direct discrimination, save for certain exceptions, such as genuine and determining occupational requirements.¹² By contrast, the notion of indirect discrimination triggers an open-ended regime of justifications. This means that a defendant can invoke justifications to be put to the consideration of a court. The CJEU applies a so-called **proportionality test**, the aim of which is to find out whether the existence of a disproportionate disadvantage can be justified by a measure serving a legitimate aim, appropriate to fulfil that aim, and strictly necessary in the sense that no other less intrusive measure could have been taken to fulfil the same purpose.¹³ One particular challenge in this context is for applicants and respondents to bring evidence to support or rebut discrimination claims. Several problems arise: machine learning models evolve when exposed to new data, some of them are so complex that they represent so-called ‘black boxes’, and trade secrets and intellectual property rights can restrict applicants’ access to decision-making processes. In general, access to intelligible, meaningful and actionable information in the context of algorithmic discrimination claims might be difficult to obtain both for applicants and judges, but this might also be the case for defendants themselves.

¹¹ See, e.g., 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.

¹² Such exceptions include, for instance, genuine and determining occupational requirements, which can be invoked as laid out in Article 4 Directive 2000/43/EC; Article 4 Directive 2000/78/EC; and Article 14(2) of Directive 2006/54/EC. Article 6 of Directive 2000/78/EC also contains a number of exceptions to direct age discrimination.

¹³ Purely economic justifications are excluded from the scope of acceptable justifications in principle.

Participant input 3: Is AI more likely to produce direct or indirect discrimination?

Is AI more likely to produce direct or indirect discrimination?

According to 83% of respondents, AI is more likely to produce indirect discrimination, whereas only 17% believe it produces direct discrimination.

Commentators have highlighted the risks of indirect discrimination posed by algorithmic decision-making and risk assessment, in particular through **proxy discrimination** where protected characteristics are not directly used as variables but where correlated datapoints encode discrimination nonetheless.¹⁴ Recent research, however, demonstrates that algorithmic bias can amount to direct discrimination, in particular when they lead to decisions that exclude an entire protected group from a valuable opportunity such as a job position.¹⁵ The two central notions in gender equality and non-discrimination law, namely direct and indirect discrimination, have a different capacity to adequately capture the challenges posed by machine learning models. Yet, when considering whether the use of given variables in algorithmic systems, it is important to consider the reason for using them and the aim of the system. For example, actively using gender as a decision variable might amount to discrimination if the system is used to restrict access to valuable opportunities whereas it might amount to positive action if used to allocate support or temporary benefits. The case of age illustrates this point as well: used in a system deployed to help predict sickness, it can be a useful variable for diagnosis purposes whereas using age as a predictor of unemployment in a system used to restrict access to labour market integration support programmes might be discriminatory.

Table 4: Direct versus indirect discrimination in the framework of AI/ML

Direct discrimination vs indirect discrimination in the framework of AI/ML ¹⁶		
	Strengths	Weaknesses
Direct discrimination	<ul style="list-style-type: none"> ► It is not necessary to prove an intention to discriminate to show direct discrimination under EU law, meaning that direct discrimination also covers situations where the developers or users of an algorithm did not intend to design a discriminatory model, but the deployed algorithm treats individuals and groups sharing certain protected categories less favourably. ► Direct discrimination covers situations in which a person is treated unfavourably because he or she belongs to a vulnerable group, without sharing its characteristics (discrimination by ascription and association). 	<ul style="list-style-type: none"> ► The processing of data and its categorisation by algorithms may not be comprehensible to the human brain. For example, the variables and categories on which an algorithm is based may have no meaning for humans, in the case for instance of mere mathematical probabilities. Thus, there might be a difficulty in understanding whether they can be considered as (direct proxies for) protected characteristics or not. ► The black box nature of certain algorithms could represent a challenge when it comes to proving direct discrimination in

¹⁴ See e.g. Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, Vol. 55, Issue 4, pages 1143 – 1185.

¹⁵ Adams-Prassl, J., Binns, R. and Kelly-Lyth, A. (2022). Directly Discriminatory Algorithms. *Modern Law Review*, Vol. 86, Issue 1, pages 144-175.

¹⁶ European Commission, Directorate-General for Justice and Consumers, Gerards, J., Xenidis, R., (2021). Algorithmic discrimination in Europe: challenges and opportunities for gender equality and non-discrimination law, Publications Office, pages 67-73, available at: <https://data.europa.eu/doi/10.2838/544956>.

Direct discrimination vs indirect discrimination in the framework of AI/ML¹⁶

	<ul style="list-style-type: none"> ▶ Direct discrimination covers situations of proxy discrimination where proxies are 'inextricably linked' with a protected ground (e.g. pregnancy and sex).¹⁷ ▶ The increasing awareness of relevant legal obligations could lead to a reduction of direct discrimination patterns. ▶ Direct discrimination could decrease in the context of algorithms, as the direct inclusion of protected categories in the decision-making process could produce lower predictive accuracy. For this reason, developers aware of these risks might remove protected categories from the pool of available variables for algorithmic decision making in order to avoid direct discrimination. Yet, it might also make sense to include these characteristics to actively trace or combat direct discrimination (positive action). 	<p>a trial, due to the need to establish a comparator under EU law. If the lack of transparency or intelligibility (black box) of the functioning of an algorithm prevents the gathering of evidence on how the algorithm has treated (or would have treated) an individual, then the direct discrimination may be entirely precluded.</p>
Indirect discrimination	<ul style="list-style-type: none"> ▶ Regardless of the intention of the developers, or the company or public administration using the AI/ML model, if the AI/ML disproportionately disadvantages a protected group the situation falls under the definition of indirect discrimination. ▶ The indirect discrimination concept focuses mainly on the effects of any decision, measure or policy in terms of disadvantage experienced by protected groups. ▶ The concept of indirect discrimination, as interpreted by the CJEU, can adequately address situations of discrimination by proxy, i.e. even in situations where the group or individual being harmed does not possess the protected characteristic in question. 	<ul style="list-style-type: none"> ▶ Indirect discrimination will be difficult to prove for individual applicants without the support of monitoring bodies or organisations. ▶ Access to group-based data on the potentially discriminatory effects of algorithmic systems on different groups will condition the ability to bring proof and establish meaningful comparisons in the context of court proceedings. ▶ The proportionality test that accompanies the assessment of cases of indirect discrimination is open-ended. Courts might encounter difficulties when assessing whether the parameters of an algorithmic system are 'proportionate' i.e. whether they reach the right balance in fairness/accuracy trade-offs or use the right technical definition of fairness.¹⁸

It is noted that proxy discrimination has sometimes been treated by the CJEU as a case of direct discrimination while it has been considered indirect discrimination in other cases. For instance, pregnancy discrimination in the

¹⁷ See e.g. C-177/88 - Dekker v Stichting Vormingscentrum voor Jong Volwassenen (Judgment of the Court of 8 November 1990, Elisabeth Johanna Pacifica Dekker v Stichting Vormingscentrum voor Jong Volwassenen (VJV-Centrum) Plus, ECLI:EU:C:1990:383).

¹⁸ Many technical definitions of fairness co-exist and some are incompatible.

Dekker case¹⁹ was considered by the CJEU as displaying an 'inseparable link' to gender discrimination and was therefore treated as direct discrimination. But in another case (*Jyske Finans*), the differential treatment of an EU citizen based on his birthplace was not considered a case of direct (nor indirect) discrimination on grounds of racial or ethnic origin.²⁰ It is therefore difficult to identify the legal framework applicable to algorithmic proxy discrimination, as the approach of the CJEU has not always been consistent in the past.

3.2.2 Fairness and bias versus equality and discrimination



The notions of 'bias' and 'fairness' are grounded in statistics, computer science and ethics and have specific meanings that are not necessarily well-suited to capturing the specific problems that arise in relation to the law.

'Bias' has a much wider meaning than 'discrimination' as it is not only concerned with unfair errors but with all kinds of 'systematic' errors, which can include those of a statistical, cognitive, societal, structural or institutional nature.

When invoked in the particular context of 'fairness', 'algorithmic bias' refers to a particular type of error that 'places privileged groups at a systematic advantage and unprivileged groups at a systematic disadvantage'. Even though there is commonality with the legal definition of 'discrimination', the term 'algorithmic bias' is more encompassing than the legal term 'algorithmic discrimination' as it refers to any kind of disadvantage that could be viewed as ethically or morally wrong. From a legal point of view, 'algorithmic discrimination', on the other hand, only pertains to the unjustified unfavourable treatment of, or disadvantage experienced by, specific categories of population protected by the law either explicitly (e.g. protected grounds) or implicitly (e.g. general or open-textured non-discrimination clauses).

3.2.3 Gaps in and limitations of the legal scope

Hierarchy of protection

Table 3 shows clearly that discrimination in relation to racial or ethnic origin is prohibited by the Racial Equality Directive 2000/43/EC in employment matters, social protection, including social security and healthcare, social advantages, education and the access to and supply of goods and services. The material scope of this Directive is thus far-reaching and extends even beyond that of the gender acquis, since it also includes education.

Sex discrimination is prohibited in the realm of employment as well as in the access to goods and services. The content of media and advertising and education are outside of the material scope of Directive 2004/113/EC. In light of the growing use of AI in the fields concerned, these exceptions might lead to important weaknesses in

¹⁹ Case C-177/88, [12] and [17]. The case concerned the decision of an employer not to hire a female applicant because she was pregnant. The Court indicated that 'only women can be refused employment on grounds of pregnancy and such a refusal therefore constitutes direct discrimination on grounds of sex'. It also explained that 'whether the refusal to employ a woman constitutes direct or indirect discrimination depends on the reason for that refusal. If that reason is to be found in the fact that the person concerned is pregnant, then the decision is directly linked to the sex of the candidate'.

²⁰ C-668/15 - *Jyske Finans* (Judgment of the Court (First Chamber) of 6 April 2017, *Jyske Finans A/S v Ligebehandlingsnævnet*, acting on behalf of Ismar Huskic, ECLI:EU:C:2017:278), [20], [33]-[37]. The case concerned the request of additional proof of identity by a credit institution for loan applicants born outside the EU, the Nordic countries, Switzerland and Liechtenstein. While the credit institution argued that this was required under existing rules on money laundering, the applicant claimed that it was discriminatory on grounds of ethnic origin. The Court of Justice indicated that 'the practice of a credit institution which requires a customer whose driving licence indicates a country of birth other than a Member State of the European Union or the EFTA to produce additional identification' is 'neither directly nor indirectly connected with the ethnic origin of the person concerned' and therefore does not give rise to either direct or indirect discrimination.

terms of the ability of EU law to redress algorithmic discrimination against women, trans, intersex and gender non-conforming persons.

Algorithms can easily be used in media and advertising services, and gender-based algorithmic discrimination is plentiful in these areas. This often leads to harmful stereotyping.²¹ It has been shown, for instance, that online search results tend to reflect the gender segregation that characterises the real labour market: in the absence of bias mitigation measures, mostly female pictures are shown when searching for a “nurse” while mostly male pictures appear when searching for a “doctor”. These problems can be addressed at national level, but only in those Member States that have adopted legal frameworks on the matter that can go beyond the letter of EU law.

The grounds of religion or belief, disability, age and sexual orientation are protected under another instrument, Directive 2000/78/EC, which, unlike the Racial Equality Directive, only applies to employment matters. As a result, under EU secondary law, discrimination on grounds of religion or belief, disability, age and sexual orientation is not prohibited in relation to education, social security, and access to goods and services including healthcare, housing, advertising and the media. This problem is well known among discrimination lawyers and has been referred to as constituting an undue ‘hierarchy’ of grounds in EU non-discrimination law.

This ‘hierarchy of grounds’ that characterises EU non-discrimination legislation is highly problematic. Indeed, algorithmic discrimination is likely to arise in areas where only race and gender equality are protected, and in particular in the market for goods and services. Although ML/AI discrimination is very likely to happen in the market of good and services, EU law does not protect EU citizens against algorithmic discrimination in this area, meaning that certain groups can be lawfully excluded from the access of certain goods and services, charged higher prices or be targeted by discriminatory advertising on online platforms.

ML algorithms are also increasingly used in the field of education. The lack of EU legal guarantees against discrimination on the grounds of gender, age, disability, sexual orientation and religion is therefore problematic in this field. In the area of AI/ML, it leads to a reiteration of a negative loop in the under-representation of women and minority groups in the curricula related to IT, software development and sciences, which leads to under-representation and discrimination at a later stage in the labour market.²²

Intersectional discrimination

Another important aspect to consider is the limitations of the EU legal framework in relation to intersectional discrimination. According to the Gender Shades project²³, facial recognition software of large commercial platforms is biased against several groups, but especially against dark-skinned women. This is a case of discrimination on two intersecting protected grounds, gender and race, called intersectional discrimination. Inherent in the notion of intersectional discrimination is the fact that the discriminatory harm might not exist in relation to a sole protected ground taken in isolation, but rather only in relation to a combination of protected grounds.

European non-discrimination laws do not fully recognise intersectional discrimination, as illustrated by the decision of the CJEU in the *Parris case*²⁴. In its decision in *Parris*, the CJEU on the one hand recognised the existence of multiple discrimination, stating that ‘*discrimination may indeed be based on several of the grounds*’ protected under EU law, but on the other hand it rejected a finding of intersectional discrimination, declaring that ‘*there is [...] no new category of discrimination resulting from the combination of more than one of those grounds...*’ where

²¹ Stereotyping can be harmful for various reasons, for instance when undermining dignity, preventing access to certain goods, services or social recognition, maintaining gender segregation by prescribing certain roles and maintaining given expectations, etc. See e.g. Timmer, A. (2016). Gender Stereotyping in the Case Law of the EU Court of Justice. *European Equality Law Review*, Issue 1, p. 38-9.

²² Gerards & Xenidis (2021).

²³ Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Sorelle A. Friedler, Christo Wilson, editors, *Conference on Fairness, Accountability and Transparency, FAT 2018*, 23-24 February 2018, New York, NY, USA. Volume 81 of *Proceedings of Machine Learning Research*, pages 77-91, PMLR, 2018.

²⁴ C-443/15 - Parris (Judgment of the Court (First Chamber) of 24 November 2016, David L. Parris v Trinity College Dublin and Others, ECLI:EU:C:2016:897).

discrimination based on each protected ground taken in isolation cannot be proven. However, intersectional discrimination may be covered by those Member States that have decided to go beyond EU law.

The lack of redress for intersectional discrimination in EU law – despite the recognition of the issue of ‘multiple discrimination’ in Directives 2000/78/EC and 2000/43/EC – is particularly problematic in light of the increasing risks of intersectional discrimination linked to the granular profiling abilities of algorithmic systems fed by pervasive data mining and data brokering: it will be rare for an algorithmic system to discriminate only on the basis of a protected ground, since it will usually base its output on a multitude of different factors and variables that are all statistically correlated. The focus on a few protected grounds and the lack of proper legal recognition of intersectional discrimination in the current EU and national legislation means that such instances of ‘combined’ or highly differentiated discrimination cannot be effectively redressed.

Further, as there is no available data on intersectional discrimination, this specific type of discrimination is difficult to test and detect. Intersectional discrimination can create feedback loops, leading to exclusion and invisibility of vulnerable groups. When an AI/ML system is tested, it is important to test it also in the intersection between the different grounds.

Emergent patterns of discrimination?

Algorithmic discrimination challenges the current boundaries of EU non-discrimination law. Even though Article 21 of the Charter of Fundamental Rights establishes a non-exhaustive and open-ended list of discrimination grounds by prohibiting discrimination ‘based on any ground such as’ the characteristics listed, the CJEU curtailed the potential of Article 21 as a basis for introducing more flexibility in the personal scope of EU secondary equality and non-discrimination law.²⁵ The exhaustive nature of the list of protected grounds in EU law and the limits put by the CJEU to their expansive interpretation raise problems in relation to proxy discrimination, an issue that is particularly acute in respect of algorithms. Therefore, an issue arises with the emergence of new patterns of discrimination, such as social origin. EU secondary law does not protect all groups which are at risk of social sorting or algorithmic exclusion from discrimination. While in EU primary law, the open-ended clause of Article 21 of the Charter of Fundamental Rights protects social origin as a ground of discrimination, the CJEU has adopted in the *FOA* case a very restrictive approach on new emergent patterns of discrimination by excluding an extension protection by analogy.²⁶



Improving the quality of an AI/ML model can be used as a legal argument in the context of e.g. a proportionality test under European discrimination law. However, the improvement of an already pre-existing practice does not in any way guarantee that the deployment of an AI/ML model practice will be accepted by the court as having some information value. It cannot be concluded whether such improvement will suffice to meet the necessity threshold applied by e.g. the CJEU. Such a decision is highly context-dependent and simply cannot be predicted. In other terms, it is not clear whether a court might consider ‘the relative improvement of the situation’ (comparing a situation where age cut-off points are used with a situation with the use of more evolved ML techniques) as meeting the necessity criteria if the system produces ‘disproportionate disadvantage’ against a protected group. EU discrimination law does not explicitly give consideration to things like ‘relative improvement’ compared to a pre-existing scheme. It is hereby to be taken into account that, whether the discrimination is ‘inadvertent’ does not matter under EU discrimination law, i.e. no intention is required to qualify for discrimination.

²⁵ See e.g. C-354/13 - *FOA* (Judgment of the Court (Fourth Chamber), 18 December 2014, *Fag og Arbejde (FOA) v Kommunernes Landsforening (KL)*, ECLI:EU:C:2014:2463).

²⁶ Case C-354/13.

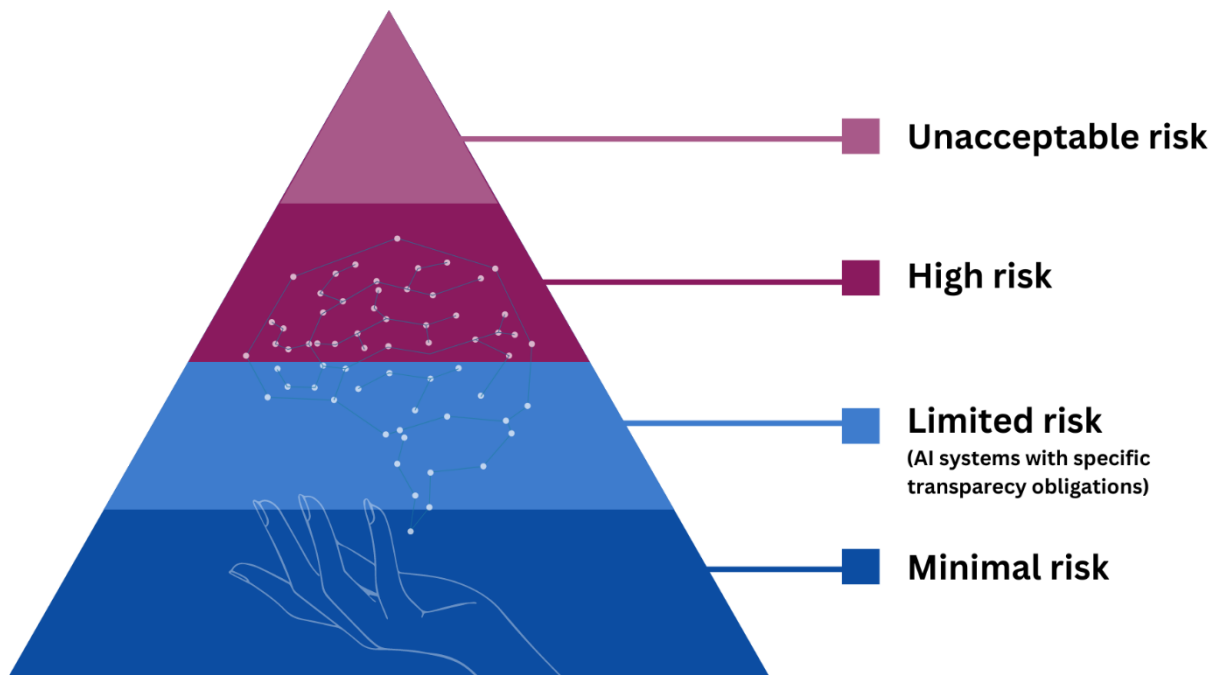
3.3 European AI Sectoral Regulation

Since March 2018, the European Union has put Artificial Intelligence on the political agenda. The [2021 review of the Coordinated Plan on AI](#) outlines a vision to accelerate, act, and align priorities with the current European and global AI landscape and bring AI strategy into action. The [European AI Strategy](#) aims at making the EU a world-class hub for AI and ensuring that AI is human-centric and trustworthy. The EU co-legislators adopted in May 2024 the EU AI Act that will contribute to building trustworthy AI, while the Council of Europe focus on ensuring that human rights, democracy and the rule of law are protected and promoted in the digital environment.

3.3.1 EU AI Act

The recently adopted **Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence** (hereinafter EU AI Act) is the EU's first-ever legal framework on AI and aims at addressing the risks of AI while positioning Europe. The EU AI Act categorises the risks of specific uses of AI into four different levels: unacceptable risk, high risk, limited risk, and minimal risk.

Figure 5: The risk-based approach in the EU AI Act



Source: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

For the **high-risk category**, the Regulation sets binding provisions for systems that are particularly at risk of endangering fundamental rights. This category, detailed in Annex III, includes for example, systems that are deployed in the field of essential public services and benefits, for instance those 'intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for essential public assistance benefits and services, including healthcare services, as well as to grant, reduce, revoke, or reclaim such benefits and services'. The high-risk category also covers essential private services such as where AI systems are 'intended to be used to evaluate the creditworthiness of natural persons or establish their credit score' ('with the exception of AI systems used for the purpose of detecting financial fraud'). The AI systems used by law enforcement agencies, on their behalf or in support of their action, for example 'for the profiling of natural persons [...] in the course of the detection, investigation or prosecution of criminal offences' or 'for assessing the risk of a natural person offending or re-offending not solely on the basis of the profiling of natural persons [...], or to assess

personality traits and characteristics or past criminal behaviour of natural persons or groups' also fall within the high-risk category. In the field of employment and workers' management, high-risk AI systems include those 'intended to be used for the recruitment or selection of natural persons, in particular to place targeted job advertisements, to analyse and filter job applications, and to evaluate candidates' and those 'intended to be used to make decisions affecting terms of work-related relationships, the promotion or termination of work-related contractual relationships, to allocate tasks based on individual behaviour or personal traits or characteristics or to monitor and evaluate the performance and behaviour of persons in such relationships'. Other fields covered include AI systems used in biometrics, migration, asylum and border control management, education and vocational training, or the administration of justice.

The section below cites the **most relevant articles and recitals of the EU AI Act** in relation to the protection of fundamental rights, discrimination and bias. Each legal provision is presented in full for the interested readers and **summarised and simplified in the blue boxes** to get a sense of the EU AI Act's contents.

The EU AI Act addresses the protection of fundamental rights, discrimination (mainly in its preamble) and algorithmic bias:

- **Article 1 and Recital 1:** 'The purpose of this Regulation is to improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, **fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union** (the 'Charter'), including democracy, the rule of law and environmental protection, to **protect against the harmful effects of AI systems** in the Union'.

Recital 2: 'This Regulation should be applied in accordance with the **values of the Union enshrined as in the Charter**, facilitating the protection of natural persons, undertakings, democracy, the rule of law and environmental protection, [...].

Recital 7: '[...] Those rules should be **consistent with the Charter, non-discriminatory** and in line with the Union's international trade commitments. [...].'



The protection of fundamental rights and the protection against the harmful effects of AI systems are main objectives of the EU AI Act.

Recital 27: 'While the risk-based approach is the basis for a proportionate and effective set of binding rules, it is important to recall the **2019 Ethics guidelines for trustworthy AI developed by the independent AI HLEG** appointed by the Commission. In those guidelines, the AI HLEG developed **seven non-binding ethical principles** for AI which are intended to help ensure that AI is trustworthy and ethically sound. The seven principles include human agency and oversight; technical robustness and safety; privacy and data governance; transparency; **diversity, non-discrimination and fairness**; societal and environmental well-being and accountability. [...] Diversity, non-discrimination and fairness means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while **avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law**. [...].'



The Ethics guidelines for trustworthy AI call for:

- **human agency and oversight,**
- **technical robustness and safety,**
- **privacy and data governance,**
- **transparency,**

- **non-discrimination and fairness,**
- **societal and environmental well-being, and**
- **accountability.**

► **Recital 45:** ‘Practices that are prohibited by Union law, including data protection law, non-discrimination law, consumer protection law, and competition law, should not be affected by this Regulation’.



The EU AI Act does not affect the applicability of pre-existing non-discrimination law (detailed above) nor data protection law (detailed below).

Recital 48: ‘The extent of the adverse **impact caused by the AI system on the fundamental rights** protected by the Charter is of particular relevance when **classifying an AI system as high risk**. Those rights include the right to human dignity, respect for private and family life, protection of personal data, freedom of expression and information, freedom of assembly and of association, the **right to non-discrimination**, the right to education, consumer protection, **workers’ rights**, the **rights of persons with disabilities**, **gender equality**, intellectual property rights, the **right to an effective remedy and to a fair trial**, the right of defence and the **presumption of innocence**, and the **right to good administration**. [...]’.



The potential adverse impact caused by the AI system on the fundamental rights protected by the Charter of Fundamental Rights, in particular non-discrimination and equality between men and women, impact the categorisation as high-risk AI system.

► **Recital 58:** Another area in which the use of AI systems deserves special consideration is the **access to and enjoyment of certain essential private and public services and benefits** necessary for people to fully participate in society or to improve one’s standard of living. In particular, natural persons applying for or receiving **essential public assistance benefits and services** from public authorities namely healthcare services, **social security benefits**, **social services providing protection in cases such as maternity, illness, industrial accidents, dependency or old age and loss of employment and social and housing assistance**, are typically dependent on those benefits and services and in a vulnerable position in relation to the responsible authorities. If AI systems are used for determining whether such benefits and services should be granted, denied, reduced, revoked or reclaimed by authorities, including whether beneficiaries are legitimately entitled to such benefits or services, those systems may have a significant impact on persons’ livelihood and may infringe their fundamental rights, such as the **right to social protection, non-discrimination, human dignity or an effective remedy** and should therefore be classified as **high-risk**. Nonetheless, this Regulation should not hamper the development and use of innovative approaches in the public administration, which would stand to benefit from a wider use of compliant and safe AI systems, provided that those systems do not entail a high risk to legal and natural persons.



- **AI systems used to determine the conditions of access to and enjoyment of certain essential public services and social benefits constitute high-risk AI systems.**

- **Recitals 56, 57** also specify that AI systems used in other core sectors such as education and training, and employment can ‘violate [...] the right not to be discriminated against’ and ‘perpetuate historical patterns of discrimination’.

- **Recital 59**: ‘Given their role and responsibility, actions by **law enforcement authorities** involving certain uses of AI systems are characterised by a significant degree of power imbalance and may lead to surveillance, arrest or deprivation of a natural person’s liberty as well as other adverse impacts on fundamental rights guaranteed in the Charter. In particular, if the AI system is not trained with high-quality data, does not meet adequate requirements in terms of its performance, its accuracy or robustness, or is not properly designed and tested before being put on the market or otherwise put into service, **it may single out people in a discriminatory or otherwise incorrect or unjust manner**. Furthermore, the exercise of important procedural fundamental rights, such as the right to an effective remedy and to a fair trial as well as the right of defence and the presumption of innocence, could be hampered, in particular, where such AI systems are not sufficiently transparent, explainable and documented. It is therefore appropriate to **classify as high-risk**, insofar as their use is permitted under relevant Union and national law, **a number of AI systems intended to be used in the law enforcement** context where accuracy, reliability and transparency is particularly important to avoid adverse impacts, retain public trust and ensure accountability and effective redress.



- **The use of AI systems by national authorities which are considered law enforcement authorities (for certain purposes) are classified as high-risk systems.**

- **Recital 60**: AI systems used in **migration**, asylum and border control management affect persons who are often in particularly vulnerable position and who are dependent on the outcome of the actions of the competent public authorities. The accuracy, non-discriminatory nature and transparency of the AI systems used in those contexts are therefore particularly important to guarantee respect for the fundamental rights of the affected persons, in particular their rights to free movement, non-discrimination, protection of private life and personal data, international protection and good administration. It is therefore appropriate to **classify as high-risk**, insofar as their use is permitted under relevant Union and national law, AI systems intended to be used by or on behalf of competent public authorities or by Union institutions, bodies, offices or agencies charged with tasks in the fields of migration, asylum and border control management as polygraphs and similar tools, for assessing certain risks posed by natural persons entering the territory of a Member State or applying for visa or asylum, for assisting competent public authorities for the examination, including related assessment of the reliability of evidence, of applications for asylum, **visa and residence permits** and associated complaints with regard to the objective to establish the eligibility of the natural persons applying for a status, for the purpose of detecting, recognising or identifying natural persons in the context of migration, asylum and border control management, with the exception of verification of travel documents.



- **AI systems used in the field of migration (including application for visas) are classified as high-risk systems.**

Article 10 of the EU AI Act is particularly relevant when it comes to non-discrimination

Article 10 - Data and data governance

1. High-risk AI systems which make use of techniques involving the training of AI models with data shall be developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in paragraphs 2 to 5 whenever such data sets are used.
2. Training, validation and testing data sets shall be subject to **data governance and management practices** appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular:
 - (a) the relevant **design choices**;
 - (b) **data collection processes** and the **origin of data**, and in the case of **personal data**, the **original purpose** of the data collection;
 - (c) relevant **data-preparation processing operations**, such as annotation, labelling, cleaning, updating, enrichment and aggregation;
 - (d) the **formulation of assumptions**, in particular with respect to the information that the data are supposed to measure and represent;
 - (e) an **assessment of the availability, quantity and suitability of the data** sets that are needed;
 - (f) **examination in view of possible biases** that are likely to affect the health and safety of persons, have a negative impact on **fundamental rights** or lead to **discrimination** prohibited under Union law, especially where data outputs influence inputs for future operations;
 - (g) **appropriate measures to detect, prevent and mitigate possible biases** identified according to point (f);
 - (h) the identification of **relevant data gaps or shortcomings** that prevent compliance with this Regulation, and how those gaps and shortcomings can be addressed.

In line with this Article, Recital 67 lays down the principles applicable to data quality:

- Recital 67: **High-quality data and access to high-quality data** plays a vital role in providing structure and in ensuring the performance of many AI systems, especially when techniques involving the training of models are used, with a view to ensure that the high-risk AI system performs as intended and safely and it **does not become a source of discrimination prohibited by Union law**. High-quality data sets for training, validation and testing require the implementation of appropriate data governance and management practices. **Data sets for training, validation and testing, including the labels, should be relevant, sufficiently representative, and to the best extent possible free of errors and complete in view of the intended purpose of the system**. In order to facilitate compliance with Union data protection law, such as Regulation (EU) 2016/679, data governance and management practices **should include, in the case of personal data, transparency about the original purpose of the data collection**. The data sets should also have the **appropriate statistical properties**, including as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used, with **specific attention to the mitigation of possible biases** in the data sets, that are **likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law**, especially where data outputs influence inputs for future operations (**feedback loops**). Biases can for example be inherent in underlying data sets, especially when **historical data** is being used, or generated when the systems are implemented in real world settings. Results provided by AI systems could be influenced by such **inherent biases** that are inclined to gradually increase and thereby **perpetuate and amplify existing discrimination**, in particular for persons belonging to

certain **vulnerable groups, including racial or ethnic groups**. The requirement for the data sets to be to the best extent possible complete and free of errors should not affect the use of privacy- preserving techniques in the context of the development and testing of AI systems. In particular, **data sets should take into account**, to the extent required by their intended purpose, the **features, characteristics or elements that are particular to the specific geographical, contextual, behavioural or functional setting which the AI system is intended to be used**. The requirements related to data governance can be complied with by having recourse to third parties that offer certified compliance services including verification of data governance, data set integrity, and data training, validation and testing practices, as far as compliance with the data requirements of this Regulation are ensured.



- ▶ **High-quality data is an essential principle for the design and implementation of AI systems. Datasets should be:**
 - **relevant**
 - **sufficiently representative**
 - **free of errors and**
 - **complete.**
- ▶ **Datasets should also have the appropriate statistical properties.**
- ▶ **Specific attention must be given to the mitigation of potential biases (e.g. via feedback loops, reinforcement of social bias in underlying data).**
- ▶ **Datasets should be tailored to specific geographical, contextual, behavioural of functional setting in which the AI system is implemented.**

3.3.2 AI Liability Directive (proposed)

As seen above, the EU has adopted a legal framework for artificial intelligence which aims to address the risks generated by specific uses of AI through a set of rules focusing on the respect of fundamental rights and safety. At the same time, the Commission intends to make sure that persons harmed by artificial intelligence systems enjoy the same level of protection as persons harmed by other technologies. Therefore, a Proposal for an Artificial Intelligence Liability Directive was delivered in September 2022.

The purpose of the AI Liability Directive proposal is to improve the functioning of the internal market by laying down uniform rules for certain aspects of non-contractual civil liability for damage caused with the involvement of AI systems.

Two provisions are important for us, as they facilitate enforcement of anti-discrimination rules:

▶ **Article 3: Disclosure of evidence and rebuttable presumption of non-compliance**

- ▷ Disclosure of evidence: access to evidence in the context of information asymmetries (Art 3(1)): *Member States shall ensure that national courts are empowered, either upon the request of a potential claimant who has previously asked a provider, a person subject to the obligations of a provider [...] or a user to disclose relevant evidence at its disposal about a specific high-risk AI system that is suspected of having caused damage, but was refused, or a claimant, to order the disclosure of such evidence from those persons. In support of that request, the potential claimant must present facts and evidence sufficient to support the plausibility of a claim for damages'*
- ▷ No 'blanket requests' + 'in support of that request, the potential claimant must present facts and evidence sufficient to support the plausibility of a claim for damages'

- ▷ Presumption of non-compliance: rebuttable presumption of breach of duty of care (Art. 3(5)): *Where a defendant fails to comply with an order by a national court in a claim for damages to disclose or to preserve evidence at its disposal [...], a national court shall presume the defendant's non-compliance with a relevant duty of care, [...], that the evidence requested was intended to prove for the purposes of the relevant claim for damages. The defendant shall have the right to rebut that presumption.*

► **Article 4: Rebuttable presumption of a causal link in the case of fault**

- ▷ *'[...] national courts shall presume, for the purposes of applying liability rules to a claim for damages, the causal link between the fault of the defendant and the output produced by the AI system or the failure of the AI system to produce an output'*

3.3.3 Council of Europe Framework Convention on AI

In December 2021, the Council of Europe's Ad Hoc Committee on Artificial Intelligence (CAHAI) published "Possible elements of a legal framework on artificial intelligence based on the Council of Europe's standards on human rights, democracy and the rule of law". The Possible Elements Report established the need for an international, legally binding treaty focused on AI. The Report laid the groundwork for the successive Committee on Artificial Intelligence. This Committee will build on the CAHAI's recommendations and elaborate an "appropriate legal instrument", likely to lead to a transversal legally binding document by 2023. The Council of Europe's Treaty on AI is still in development.²⁷

3.4 Interaction with data protection law: taking stock of European developments

3.4.1 EU General Data Protection Regulation

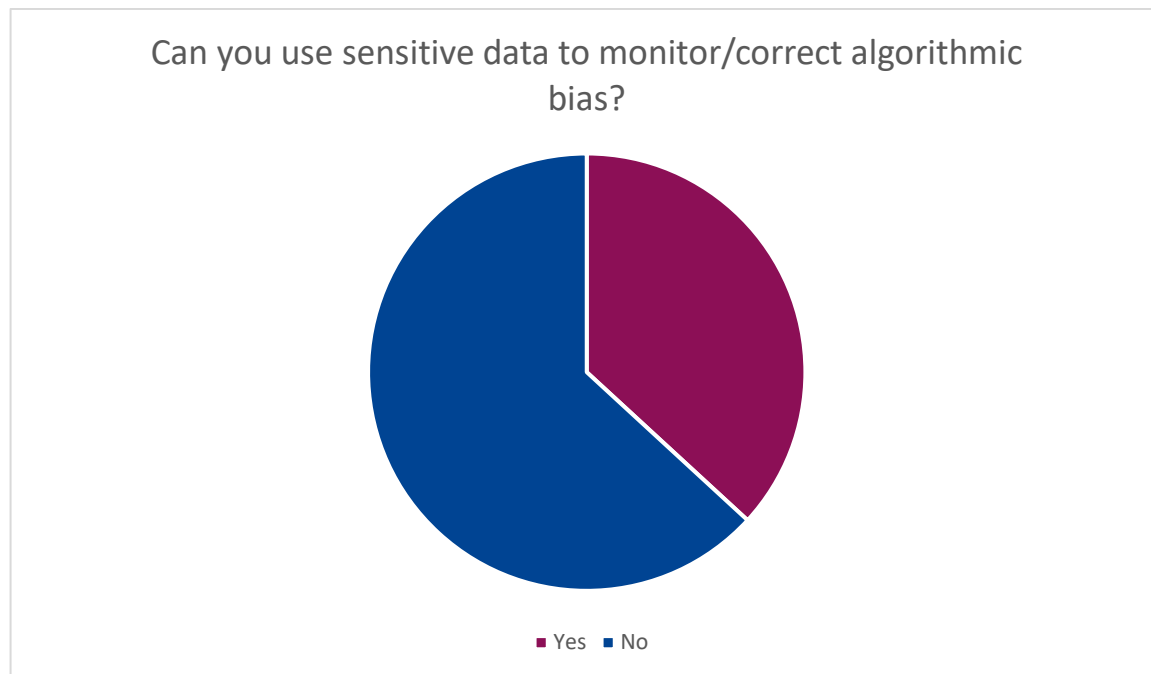
There is a strong link between the role of data protection law and legal obligations related to privacy in the prevention of algorithmic discrimination. The rationale is that certain categories of data – for instance race, religion, sexual orientation, etc. – are particularly sensitive because they can easily lead to unlawful discrimination if processed without particular precautions. This is reflected in the EU's General Data Protection Regulation (GDPR)²⁸, which identifies 'special categories of personal data' or 'sensitive data'. However, the list of categories of data the processing of which could give risk to discrimination does not neatly fit with the list of protected grounds under EU gender equality and non-discrimination law. Importantly, the issue of gender equality or sex discrimination is altogether absent from the GDPR and neither gender nor sex are mentioned as sensitive categories of personal data. Racial or ethnic origin, religion or belief and sexual orientation are explicitly mentioned both in relation to discrimination in Recital 71 and in relation to the prohibition of processing such data, but the Recital does not refer to 'sex' or grounds such as 'age' and 'disability'. Similarly, **Article 9(1) GDPR** prohibits the 'processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation'.

²⁷ <https://www.coe.int/en/web/artificial-intelligence/work-in-progress>

²⁸ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, p. 1–88, available at: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

The EU AI Act does not provide the legal ground for processing of personal data, including special categories of personal data, where relevant, unless it is specifically otherwise specified.

Participant input 4: Can you use sensitive data to monitor or correct algorithmic bias?



➔ Does data protection law allow using sensitive data for the purpose of detecting and correcting algorithmic bias?

- **The answer is yes.** But this is subject to appropriate safeguards for the fundamental rights and freedoms of natural persons, including technical limitations on the re-use and use of state-of-the-art security and privacy-preserving measures, such as pseudonymisation, or encryption where anonymisation may significantly affect the purpose pursued.

Recital 70: *'In order to protect the right of others from the discrimination that might result from the bias in AI systems, the providers should, exceptionally, to the extent that it is strictly necessary for the purpose of ensuring bias detection and correction in relation to the high-risk AI systems, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons and following the application of all applicable conditions laid down under this Regulation in addition to the conditions laid down in Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680, be able to process also special categories of personal data, as a matter of substantial public interest within the meaning of Article 9(2), point (g) of Regulation (EU) 2016/679 and Article 10(2), point (g) of Regulation (EU) 2018/1725.'*

Article 9 of the GDPR defines the processing of special categories of personal data and prohibits such processing. Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation **shall be prohibited**.

However, Art. 10(5) of the EU AI Act provides that, *to the extent that it is strictly necessary for the purpose of ensuring bias detection and correction in relation to the high-risk AI systems [...], the providers of such systems may exceptionally process special categories of personal data [referred to in Article 9(1) of the GDPR], subject to appropriate safeguards for the fundamental rights and freedoms of natural persons.* According to Art.9(2)(g) GDPR, indeed, the prohibition shall not apply *if the processing is necessary for reasons of substantial interest, on*

the basis of the Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.

Additional conditions are additional to those of the GDPR, see Article 10(5):

'In addition to the provisions set out in Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680, all the following conditions must be met in order for such processing to occur:

*(a) the bias detection and correction **cannot be effectively fulfilled by processing other data, including synthetic or anonymised data;***

*(b) the special categories of personal data are subject to **technical limitations on the re-use** of the personal data, and **state-of-the-art security and privacy-preserving measures**, including pseudonymisation;*

*(c) the special categories of personal data are subject to measures to ensure that the personal data processed are secured, protected, subject to suitable safeguards, including **strict controls and documentation of the access**, to avoid misuse and ensure that only authorised persons have access to those personal data with appropriate confidentiality obligations;*

*(d) the special categories of personal data are **not to be transmitted, transferred or otherwise accessed by other parties;***

*(e) the special categories of personal data are **deleted** once the bias has been corrected or the personal data has reached the end of its retention period, whichever comes first;*

(f) the records of processing activities pursuant to Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680 include the reasons why the processing of special categories of personal data was strictly necessary to detect and correct biases, and why that objective could not be achieved by processing other data.'

3.4.2 Council of Europe Convention 108+

The Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No. 108) opened for signature on 28 January 1981 and was the first legally binding international instrument in the data protection field. Under this Convention, the parties are required to take the necessary steps in their domestic legislation to apply the principles it lays down in order to ensure respect in their territory for the fundamental human rights of all individuals with regard to processing of personal data.

Importantly, Article 6 states the following:

1. The processing of:

— genetic data;

— personal data relating to offences, criminal proceedings and convictions, and related security measures;

— biometric data uniquely identifying a person;

— personal data for the information they reveal relating to racial or ethnic origin, political opinions, trade-union membership, religious or other beliefs, health or sexual life,

shall only be allowed where appropriate safeguards are enshrined in law, complementing those of this Convention.

2. Such safeguards shall guard against the risks that the processing of sensitive data may present for the interests, rights and fundamental freedoms of the data subject, notably a risk of discrimination.

4.0 Mitigation framework

Artificial Intelligence and Machine Learning can cause a variety of **legal, ethical or fairness-related harms**²⁹:

- ▶ They can unfairly allocate opportunities, resources, or information;
- ▶ They can also fail to provide the same quality of service to some people as they do to others;
- ▶ They can reinforce existing societal stereotypes;
- ▶ They can over- or underrepresent groups of people, or even treat them as if they don't exist;
- ▶ They can denigrate people by being actively derogatory or offensive.



Artificial Intelligence and Machine Learning biases pose severe risks to the use of AI and require ongoing attention and evaluation by competent experts.

A constant balance is required. Since the AI and ML biases pose severe risks to the use of AI, they require attention and evaluation by competent experts. People with knowledge in the field can play a key role in mitigating potential biases. Successful mitigation methods do not necessarily require knowledge of data science; it is often sufficient to have knowledge of the domain concerned.

Table 5 shows 24 methods that can mitigate the above-mentioned eight biases within the CRISP-DM process. Notably, a particular bias can be mitigated by several methods, and a particular method can mitigate multiple biases. In addition, a mitigation method that is applied in one phase can address biases that occur in the respective phase or in the later stages of the ML project.³⁰

²⁹ See: Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. Available at: <https://doi.org/10.1145/3313831.3376445>

³⁰ For more background and detail, see: van Giffen et al (2022).

Table 5: Overview of 24 mitigation methods for addressing biases within the CRISP-DM process phase

Overview of 24 bias mitigation methods allocated to CRISP-DM process phases

Phase Bias	Business Understanding	Data Understanding	Data Preparation	Modelling	Evaluation	Deployment
Social Bias			Rapid Prototyping Reweighting Data Massaging Disparate Impact Remover Learning Fair Representation Optimised Pre-Processing	Prejudice Remover Adversarial Debiasing Equalised Odds Multiple Models Latent Variable Model Model Interpretability		
Measurement Bias			Rapid Prototyping			
Representation Bias	Diversity in Teams Exchange with Domain Experts	Proxy Estimation	Reweighting Data Augmentation	Model Interpretability		
Label Bias	Diversity in Teams	Data Plotting Exchange with Domain Experts	Data Massaging			
Algorithmic Bias		Exchange with Domain Experts	Rapid Prototyping	Exchange with Domain Experts Resampling Model Interpretability Multitask learning		
Evaluation Bias				Resampling	Representative Benchmark Subgroup Validity Data Augmentation	
Deployment Bias	Diversity in Teams Consequences in Context		Rapid Prototyping			Monitoring Plan Human Supervision
Feedback Bias						Human Supervision Randomness

Key methods w/o technical knowledge


1. Diversity in Teams
2. Exchange with domain experts
3. Consequences in context
4. Data plotting
5. Rapid prototyping
6. Monitoring plan
7. Human supervision


Apart from the 24 mitigation methods, Table 5 highlights the seven key methods that can most often be implemented without the need or help of a data scientist:

1. Diversity in teams helps to mitigate measurement, and can prevent representation and deployment bias;
2. Exchange with domain experts on project objectives addresses emerging measurement bias and prevents representation bias;
3. Discuss social and technical consequences of the ML model (especially to prevent deployment bias);
4. Data plotting can reveal spikes (i.e., one-time phenomena/outliers) that affect empirical conclusions and need to be removed to prevent representation bias;
5. Rapid prototyping is an effective approach for identifying different types of unintended bias;
6. Monitoring plan;
7. Human supervision in the deployment helps to enhance objectivity and mitigates possible occurrence of deployment and feedback bias.


Table 6 explains the advantages of each of these seven key methods as well as a short methodology.

Table 6: Why and how to use the seven key mitigation methods (extended from van Giffen et al. 2022)

Key method	Why doing it?	How to do it?	
1. Diversity in teams	<ul style="list-style-type: none"> ▶ define the ML problem better, select more appropriate features, specify representative populations, and anticipate different use contexts ▶ identify potential harms by introducing different perspectives on the ML task ▶ better reflection of the target population (e.g., demographics, preferences) 	<ul style="list-style-type: none"> ▶ integrate target users in the core team ▶ seek for diverse backgrounds, e.g., along key dimensions that will likely be relevant in the ML task (e.g., gender, background, race, income, preferences and other demographics) 	This is one of the most important measures you can take!
2. Exchange with domain experts	<ul style="list-style-type: none"> ▶ Understand key relations between data categories (=making sense) ▶ consider possibly affected populations ▶ domain experts help designing the ML model with appropriate and measurable target variables and features 	<ul style="list-style-type: none"> ▶ Seek for experts with deep domain knowledge (years of experience and scars on their back) ▶ establish a shared (qualitative and quantitative) understanding of sensitivities, e.g., what is considered good/bad, fair/unfair, and performance means 	Introduce the context and prediction task with domain experts to stimulate a discussion about (salient) assumptions and potential risks in the particular context.
3. Consequences in context	<ul style="list-style-type: none"> ▶ Consider, envision and understand social context and prevailing moral (and legal) situation early on ▶ Articulate constraints regarding the 	<ul style="list-style-type: none"> ▶ Establish a comprehensive understanding of the social and technical deployment context 	Establish this user research activity in your ML project plan!

Key method	Why doing it?	How to do it?	
	applications on other use contexts clearly	<ul style="list-style-type: none"> ▶ Use a systematic method such as AEIOU-Analysis³¹ ▶ Reflect and discuss observations and their implications for the ML application 	
4. Data plotting	<ul style="list-style-type: none"> ▶ reveal spikes, one-time, phenomena, or outliers through visualization 	<ul style="list-style-type: none"> ▶ Plot key dimensions (or combinations) of selected features in adequate diagrams ▶ Identify and review potential outliers and their causes (if possible) ▶ Decide/calculate if the identified values (or series of value) is representative of your target population and context. 	Great for non-data scientists to better understand available data!
5. Rapid prototyping	<ul style="list-style-type: none"> ▶ creating and testing a prototype of the ML model can reveal discriminative effects, e.g., resulting from social bias, test and proxy variables regarding their suitability to predict the outcome of interest ▶ address measurement bias and uncover overlooked sections of the population to 	<ul style="list-style-type: none"> ▶ Prototypes can have different purposes ▶ Purpose can range from testing the quality of predictions (stochastically) to analyse interactions, usability or user experience ▶ (Potentially) involve data science and/or user (experience) researchers at various stages of your project 	This is a largely underestimated method in AI/ML development!

³¹ This method is commonly used in Design Thinking projects and focuses on understanding: Activities, Environments, Interactions, Objects, Users.

Key method	Why doing it?	How to do it?	
	prevent representation bias		
6. Monitoring plan	<ul style="list-style-type: none"> ▶ A monitoring plan for an algorithm helps detect drifts in the data. ▶ For example, if demand for a certain product slowly decreases over time and the algorithm is not regularly re-trained the demand forecasts gradually gets worse over time 	<ul style="list-style-type: none"> ▶ Account for changes in the algorithm when the context evolves. ▶ Build your monitoring plan on quantitative and qualitative metrics ▶ Assign responsibilities for monitoring the deployed AI models 	This is an ongoing activity once the model is in deployment!
7. Human supervision	Algorithmic recommendations should not be accepted “blindly” because they cannot be expected to be bias-free	<ul style="list-style-type: none"> ▶ Assess criticality of false predictions (false positives/negatives) ▶ For medium to high risks systems, include humans in the application loop to analyse and question algorithmic recommendations regularly 	Responsible humans should be trained and sensitised about how to handle irregularities!

Some analysis techniques or mitigation methods (modifying learning algorithms to mitigate biases) that are aware of social fairness or discrimination, but that need the help of a data scientist are³²:

1. Prejudice remover regulariser: this approach involves incorporating regularisation terms or constraints to address social bias. The method takes into account variations in the learning algorithm's classification of attributes like race, gender, or ethnicity (protected and non-protected), and subsequently applies penalties

³² Based on: van Giffen et al. 2022.

to the overall loss based on the extent of these differences.³³ It is a technique that reduces indirect prejudice.

2. **Adversarial debiasing:** this in-processing method maximizes accuracy while simultaneously removing the ability to identify protected attribute(s).³⁴ The method involves constructing two models.³⁵ The first model predicts the target variable using the training data, incorporating any feature engineering and pre-processing steps already performed. The second model serves as the adversary, attempting to predict the sensitive attribute based on the predictions made by the first model. In an unbiased scenario, the adversarial model should struggle to accurately predict the sensitive attribute. Consequently, the adversarial model drives adjustments to the original model, modifying its parameters and weighting, in a way that reduces the predictive capacity of the adversarial model until it can no longer accurately predict the protected attributes based on the outcomes.
3. **Equalised odds:** this post-processing approach mitigates social bias by accessing only aggregated data and ensures that true positive and false positive rates are equal across protected groups.³⁶ This method is based on a fairness metric that checks if, for any particular label and attribute, a classifier predicts that label equally well for all values of that attribute. Its goal is to ensure that a ML model performs equally well for different groups.



Managing and mitigating AI bias to ensure fairness is a complex concept and deeply contextual³⁷

- *There is **no single definition of fairness** independent of the context that will apply equally to different AI applications.*
- *Given the many complex sources of unfairness, it is **not possible to fully “debias” a system** or to guarantee fairness; the goal is to **detect and to mitigate** fairness-related harms as much as possible, but there will be a point that we have to accept that a certain distribution is represented in the model in a certain way.*
- *Prioritising fairness in AI systems often means making **trade-offs based on competing priorities**. It is therefore important to be **explicit and transparent** about priorities and assumptions.*
- *There are **seldom clear-cut answers**. It is therefore important to **document your processes** and considerations (including priorities and trade-offs), and to **seek help** to experts and users when needed.*
- *Therefore, it is important to open the model to critics, as detecting and mitigating fairness-related harms requires **continual attention and refinement**.*

Fairness in AI systems is a sociotechnical challenge as these systems can behave unfairly for a variety of reasons - some social, some technical, and some a combination of both. In other words, AI systems can behave unfairly

³³ Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. *Machine Learning and Knowledge Discovery in Databases*, 35–50. Available at: https://link.springer.com/chapter/10.1007/978-3-642-33486-3_3; and Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). Fairness Constraints: Mechanisms for Fair Classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

³⁴ Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

³⁵ For a description aimed for practitioners, see Mahmoudian, H. (2020). Using Adversarial Debiasing to Reduce Model Bias: One Example of Bias Mitigation in In-Processing Stage. Available at: <https://towardsdatascience.com/reducing-bias-from-models-built-on-the-adult-dataset-using-adversarial-debiasing-330f2ef3a3b4>

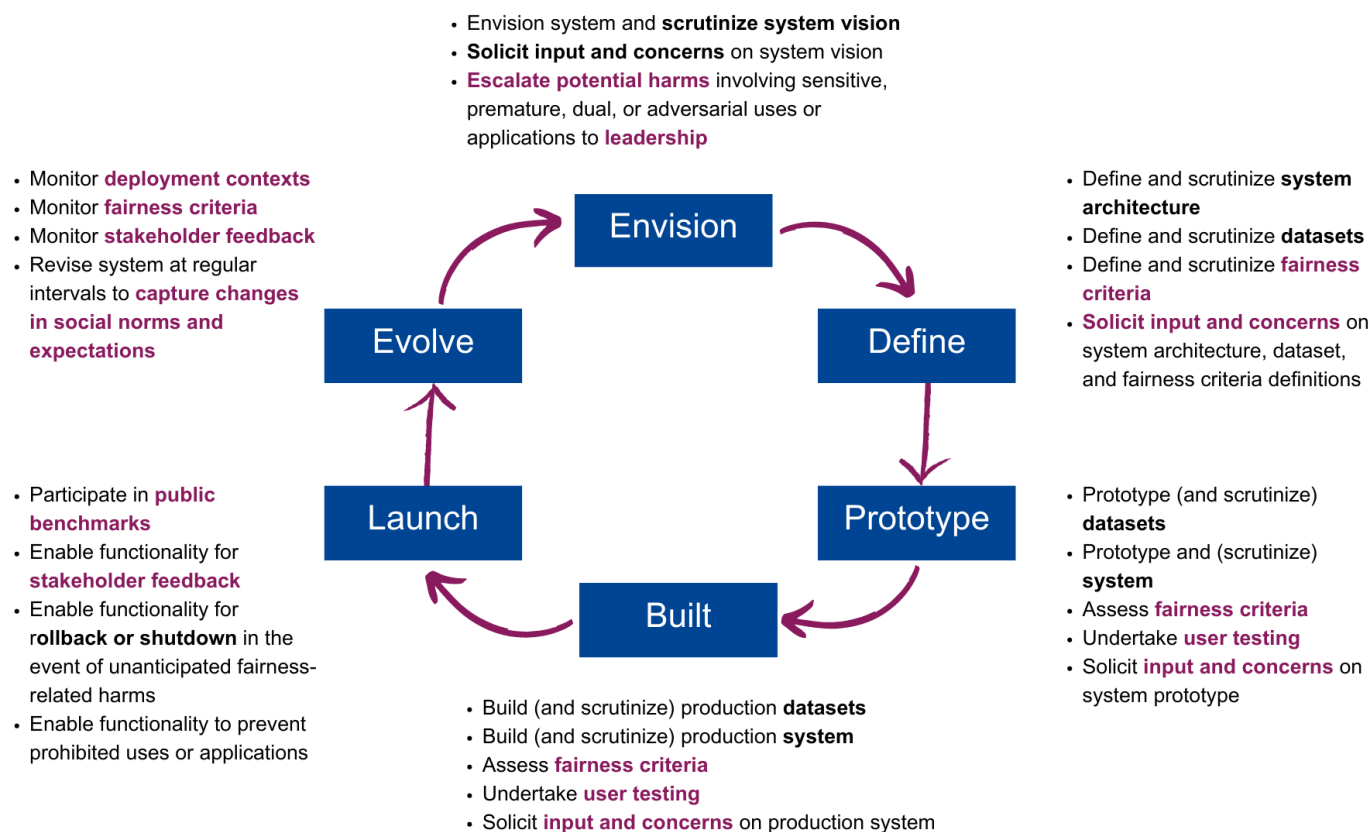
³⁶ Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. In arXiv [stat.ML]. Available at: <http://arxiv.org/abs/1808.00023>. See also: Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 3315–3323. Available at: <https://arxiv.org/abs/1610.02413>

³⁷ Excerpt from M. Madaio, et al. (2020).

because of societal biases reflected in the datasets used to train them, which are explicitly or implicitly reflected in the decisions made during AI development and the deployment lifecycle.

The following **AI fairness checklist** represents one of the first process models to generate AI systems that make “fair” predictions and/or decisions.

Figure 6: AI Fairness checklist



*** require human judgement & reasoning based on human values**

Source: Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. Available at: <https://doi.org/10.1145/3313831.3376445>.

The AI fairness checklist as represented in **Error! Reference source not found.6** is similar to the CRISP-DM process phases. Many humans, customers and users fear the use of Artificial Intelligence, for different reasons. Bias is highly relevant for the deployment of AI systems and AI bias can emerge in every project phase. It is important to have in mind that fairness in AI is a complex concept and deeply contextual. Its conceptualisation is important because a change in the real world has impacts on the ML model. Data bias can be mitigated but requires significant human judgement.

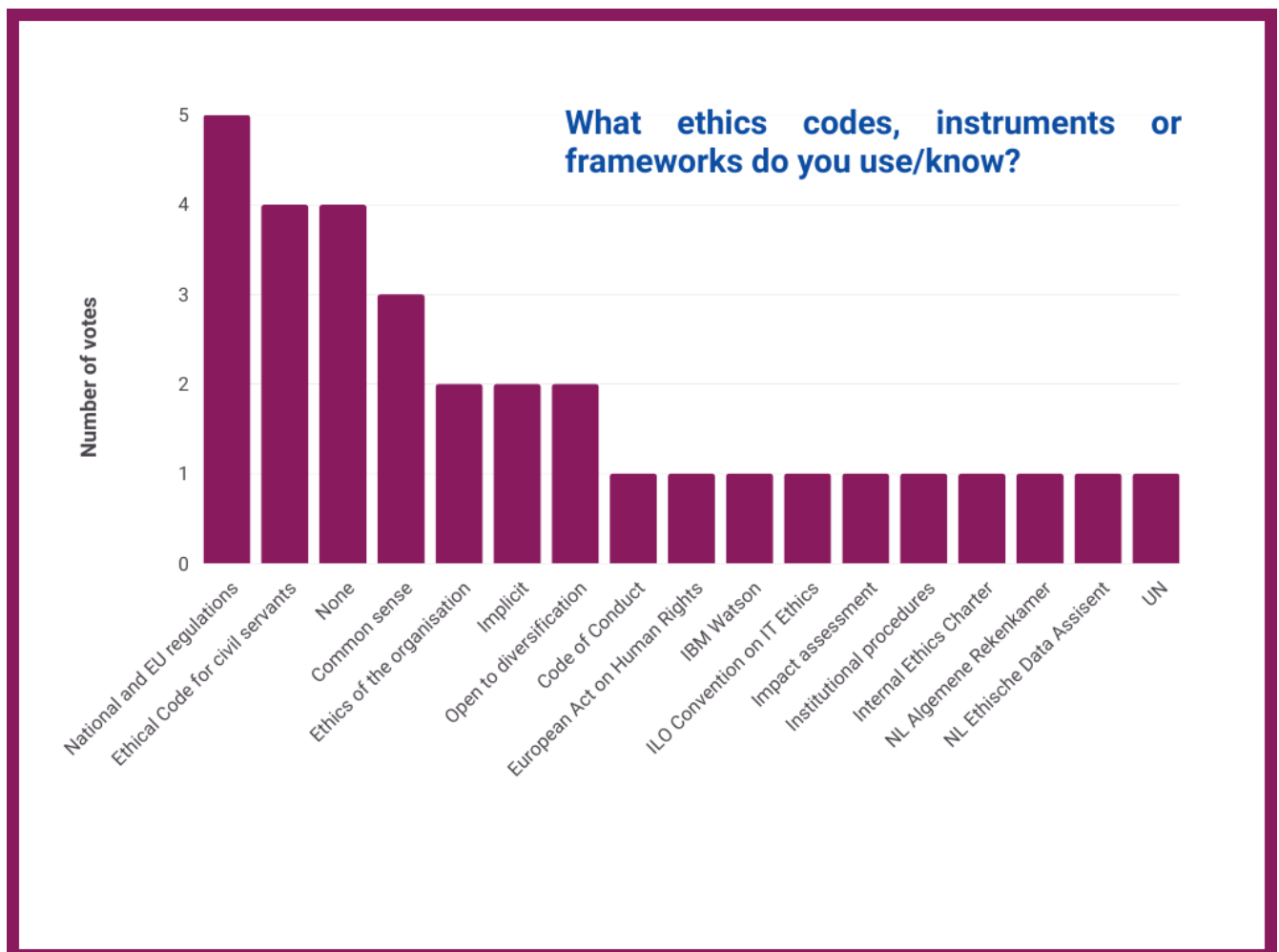


Consider the projects you are involved in. Your expertise matters. Seek for diversity and be willing to try, fail and learn when using AI/ML for societal benefit.

5.0 Key ethical requirements. Beyond the law, what ethical requirements can support non-discriminatory AI?

While the European Union has a strict legal framework in place to ensure, *inter alia*, the protection of personal data and privacy and non-discrimination, to promote gender equality, environmental protection and consumers' rights, Section 3 has shown that the existing rules on non-discrimination can continue to apply in relation to Artificial Intelligence and related technologies, although certain adjustments of specific legal instruments may be necessary to reflect the digital transformation and to address new challenges posed by the use of AI. Further, and in addition to (adjustments to) existing legislation, private companies³⁸, NGOs³⁹, research/academic institutions⁴⁰ and public sector organisations⁴¹ have issued principles and guidelines for ethical AI.

Participant input 5: What ethics codes, instruments or frameworks do you use or know?



³⁸ e.g. IBM AI Fairness 360, OpenAI Charter or Google Principles.

³⁹ e.g. Amnesty and Access Now's Toronto Declaration 2018.

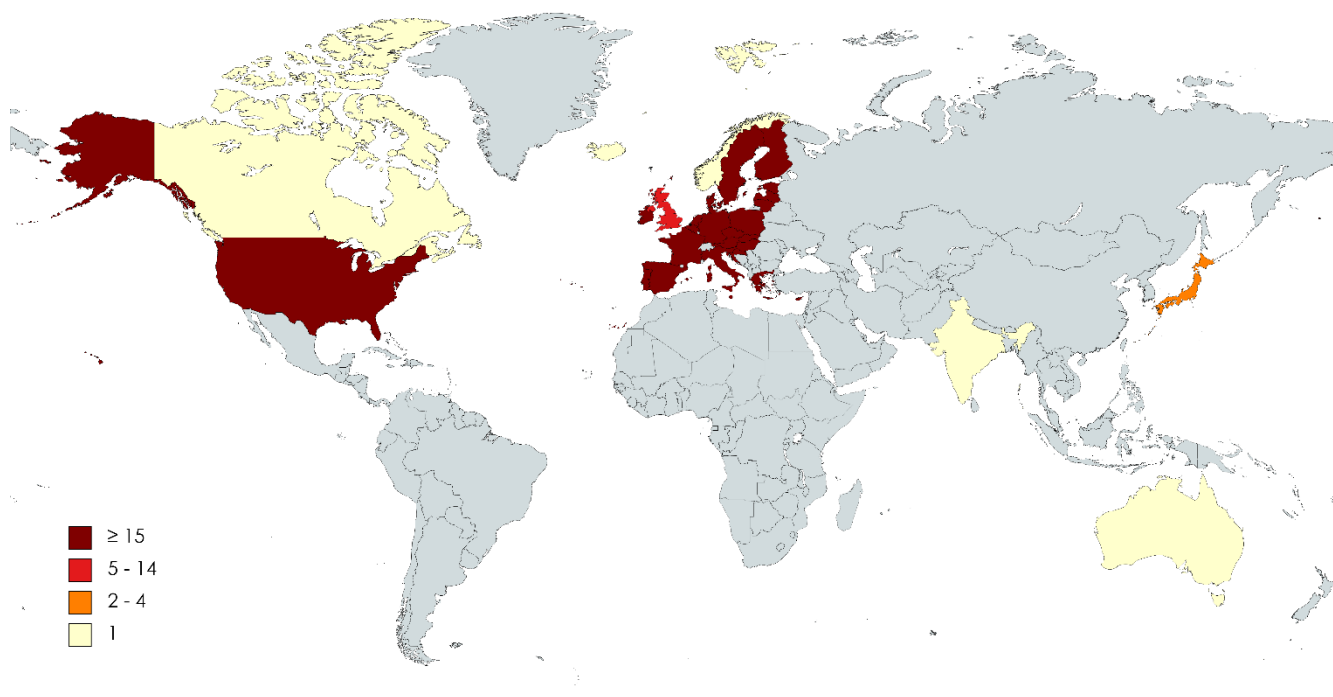
⁴⁰ e.g. Alan Turing Institute's guide for 'Understanding artificial intelligence ethics and safety' 2019. See also: AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, 2018.

⁴¹ e.g. UNESCO Recommendations on the Ethics of Artificial Intelligence, 2021 or OECD AI principles 2019. See also: Council of Europe, European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment, 2019.

Even though AI has been around for 70 years, ethical AI guidelines started proliferating in the second half of the 2010s and there is currently a myriad of ethics guidelines, codes, principles, frameworks, and tools on AI.⁴² A great part of these soft law tools have been issued in the period 2017-2019, with a geographical distribution mainly concentrating in the US and Europe (see Figure 7)**Error! Reference source not found..**⁴³ It is noted that ethics guidelines are produced in a certain context and by certain actors; this could potentially be a source of bias.

This handbook reports only upon a few different tools that are used to nudge into the direction of an “ethical AI” and does not aim at describing the entire landscape.

Figure 7: Geographical distribution of ethical AI guidelines



Source: Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, Vol. 1, pages 389–399, available at: <https://doi.org/10.1038/s42256-019-0088-2>

Ethical questions relating to AI technologies should be addressed through an effective, comprehensive and future-proof framework that closes existing legal loopholes and that increases legal certainty for businesses and citizens alike.⁴⁴ However, numerous existing ethical recommendations – aimed at both providers and users of AI and algorithmic systems – rely on self-regulation, and are therefore being heavily criticised (conflict of interest, ethics washing, bias, etc.), especially those developed by private organisations.

5.1 Ethics Guidelines for Trustworthy AI

The independent High-Level Expert Group on Artificial Intelligence (AI HLEG), set up by the European Commission in June 2018, prepared a document entitled “Ethics Guidelines for Trustworthy Artificial Intelligence (AI)”⁴⁵ in 2019.

⁴² Antonio A. Casili, “An ‘End-to-End’ Approach to Ethical AI”, Institut Polytechnique de Paris. See Casili’s AI talk @ ETUI available at <https://www.etui.org/events/ai-talks-etui-what-really-ethical-ai>

⁴³ Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI). (2020). AI Ethics Guidelines: European and Global Perspectives. Provisional report by Marcello Ienca and Effy Vayena, available at: <https://rm.coe.int/cahai-2020-07-fin-en-report-ienca-vayena/16809eccac>. See also Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, Vol. 1, pages 389–399, available at: <https://doi.org/10.1038/s42256-019-0088-2>.

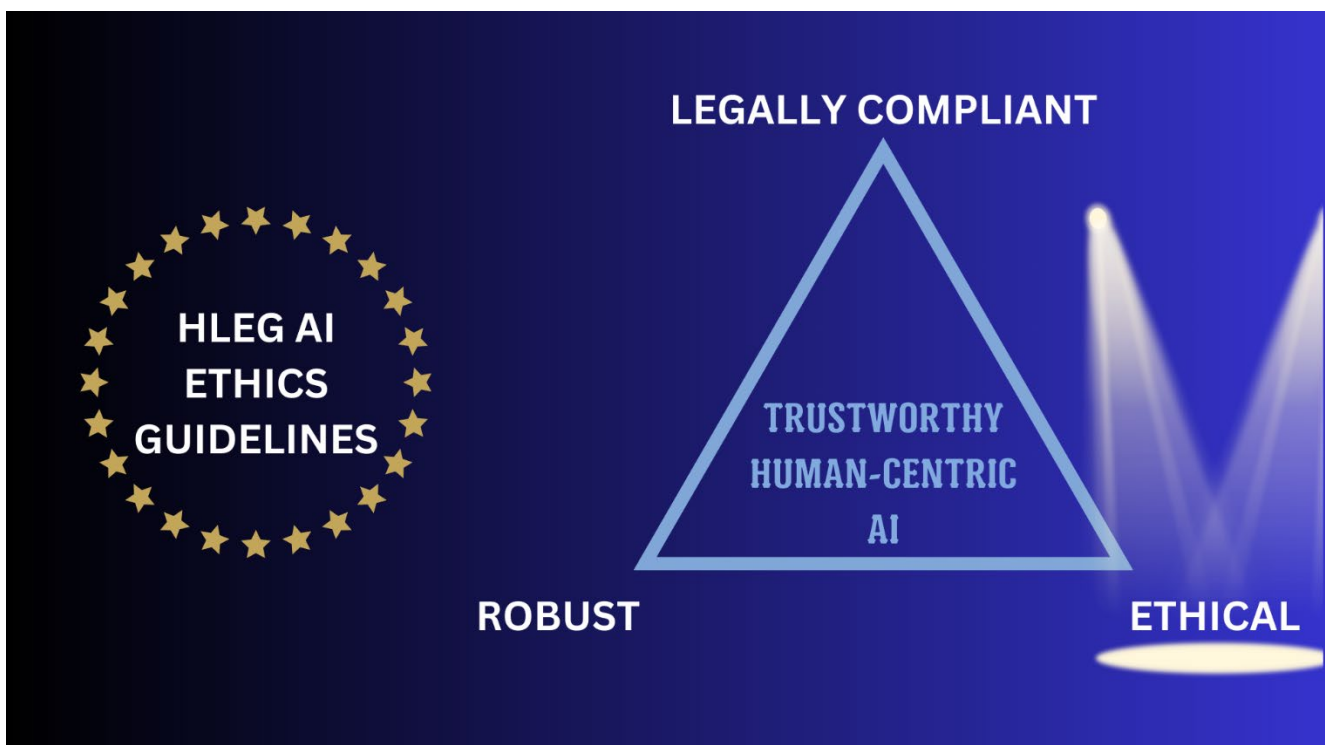
⁴⁴ European Parliament, Framework of ethical aspects of artificial intelligence, robotics and related technologies. EP Resolution of 20 October 2020.

⁴⁵ Available at: <https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>

The Guidelines resulted from a consultation of more than 500 contributors. Stakeholders welcomed the practical nature of the Guidelines as they offer concrete guidance for developers, deployers and users of AI on how to ensure the technology's trustworthiness.

The aim of the Guidelines is to promote trustworthy and human-centric AI. Trustworthy AI has three components, which should be met throughout the system's entire life cycle: (1) it should be **lawful**, complying with all applicable laws and regulations (2) it should be **ethical**, ensuring adherence to ethical principles and values and (3) it should be **robust**, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm.

Figure 8: The three components of trustworthy AI.



The above-mentioned milestone document for AI ethics in the EU calls for the respect for four principles for a trustworthy AI, in all AI models.

- **Respect for human autonomy:** *'AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans'.*
- **Prevention of harm:** *'AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. (...) Vulnerable persons should receive greater attention and be included in the development, deployment, and use of AI systems. Particular attention must also be paid to situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information'.* It is very important to prevent harm, and thus to avoid dangerous feedback loops. AI biases can make vulnerable groups even more vulnerable. Asymmetries of power or information can also lead to discrimination.
- **Fairness:** *'The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. (...) The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them.'* Fairness is a very broad and multifaceted concept, that insists on two dimensions: 1) the substantive dimension of fairness implies a

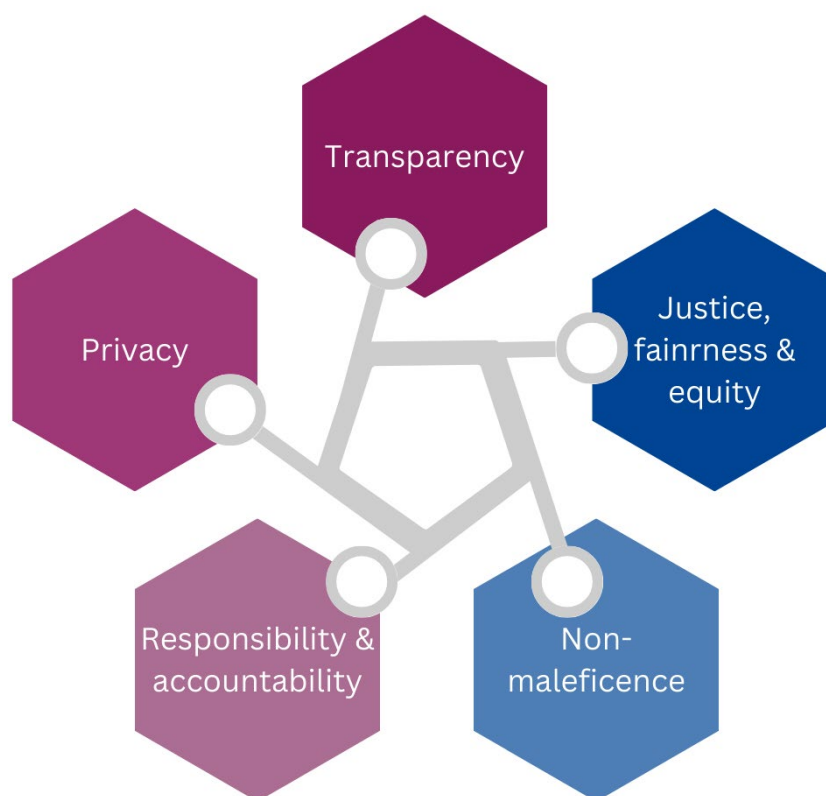
commitment to ensuring equal and just distribution of both benefits and costs; 2) the procedural dimension of fairness that, for instance, is embedded in the transparency obligations, implies that individuals have access to information and context and seek effective redress against decisions made by AI systems and by the humans operating them.

- **Explicability:** *'This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected.'*

5.2 The five converging ethics principles

Mapping the current corpus of principles and guidelines on ethical AI reveals global convergence around five ethical principles.⁴⁶ Nevertheless, there is “substantive divergence in relation to how these principles are interpreted; why they are deemed important; what issue, domain or actors they pertain to; and how they should be implemented”.⁴⁷ Figure 9 displays these five converging ethics principles, while Table 7 sets out their meaning/application and their usability, as well as some alternative keywords.

Figure 9: The five converging ethics principles



Source: Ienca, M., & Vayena, E. (2020). On the responsible use of digital data to tackle the COVID-19 pandemic. *Nature Medicine*, Volume 26(4), pages 463-464.

Transparency means that the algorithms and data processing methods as well as human practices related to the design, development and deployment of AI systems should be transparent. There are two main ways to apply the transparency principles, through technical measures (explainable AI, interpretability methods) and through non-

⁴⁶ Jobin A., Ienca M. & Vayena E. (2019), and Ienca, M. & Vayena, E. (2020).

⁴⁷ Jobin, A., Ienca M. & Vayena, E. (2019).

technical measures (audits, information disclosure, non-secrecy (e.g. open source code, accessibility / auditability of training data)).

Justice, fairness and equity of algorithmic systems means that human decision-makers, developers, data scientists should serve to maintain and foster democratic processes, with equal respect for the moral worth and dignity of all human beings.⁴⁸

The third principle of **non-maleficence** is basically a no harm principle, i.e. a duty to prevent bias to the best of one's knowledge.

The fourth principle is **responsibility and accountability**. It requires that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.

Finally, **privacy** is a fundamental right that must be guaranteed when AI systems are deployed, just as data protection, throughout a system's entire lifecycle. It covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.⁴⁹

Table 7: Explanation of the five ethics principles⁵⁰

Ethics principle	What?	How?	Alternative keywords
Transparency	<ul style="list-style-type: none"> ▶ Algorithms and data processing methods ▶ Human practices related to the design, development and deployment of AI systems 	<ul style="list-style-type: none"> ▶ Technical: explainable AI, interpretability methods ▶ Non-technical: audits, information disclosure, non-secrecy (e.g. open source code, accessibility / auditability of training data) 	access to meaningful information, explainability, intelligibility, explicability
Justice, fairness & equity	<ul style="list-style-type: none"> ▶ Algorithmic systems ▶ Human in/on the loop, human decision-makers, developers, data scientists... 	<ul style="list-style-type: none"> ▶ Diversity and inclusion in: <ul style="list-style-type: none"> ▷ data collection and use (e.g. representativeness) ▷ the design of AI systems (e.g. equality by design) ▷ the deployments in society (e.g. non-discriminatory impact) ▶ Equal representation and participation in developers teams (diverse backgrounds, etc.) 	non-discrimination, bias prevention/mitigation

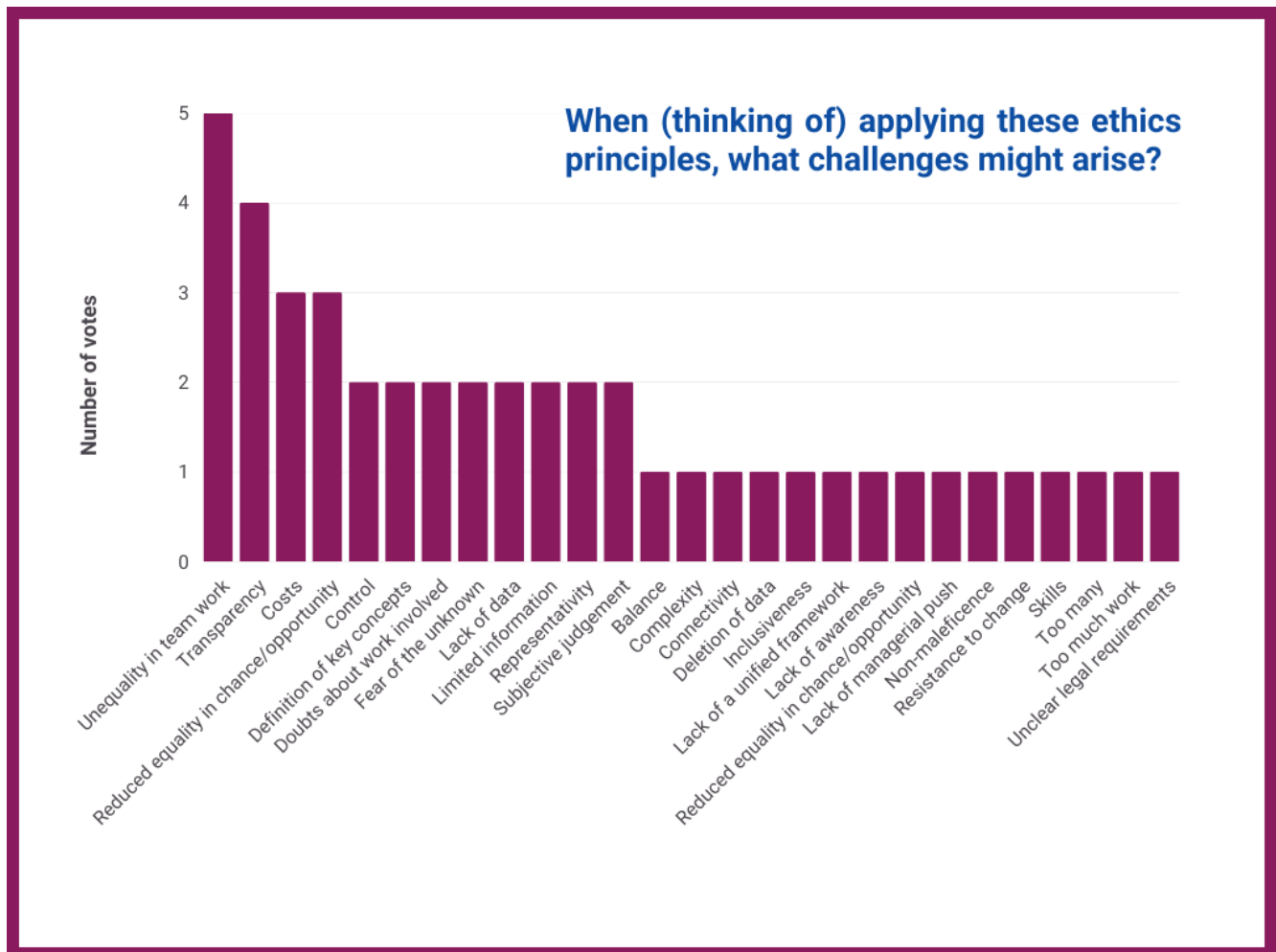
⁴⁸ See: European Commission (2019), Ethics Guidelines for Trustworthy AI. High-level Expert Group on Artificial Intelligence, available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

⁴⁹ For detailed definition, see Ienca, M. & Vayena, E. (2020).

⁵⁰ See ibid.

Ethics principle	What?	How?	Alternative keywords
		<ul style="list-style-type: none"> ▶ Right of 'appeal' and remedy of algorithmic decisions/recommendations ▶ 'De-biasing' (but numerous limitations) ▶ Bias training 	
Non-maleficence	<ul style="list-style-type: none"> ▶ No harm principle (foreseeable or not) ▶ Safety and security ▶ No misuse/ abuse ▶ Privacy concerns 	<ul style="list-style-type: none"> ▶ Compliance tests, audits, monitoring, assessments ▶ Governance structures 	beneficence, 'AI for good', ethical AI, human-centric AI
Responsibility & accountability	<ul style="list-style-type: none"> ▶ Algorithmic systems ▶ Industry sector 	<ul style="list-style-type: none"> ▶ Legal compliance ▶ Providing access to meaningful information ▶ Providing right to challenge system and decisions ▶ Informing about the use of AI or algorithmic systems ▶ Goes hand in hand with transparency 	'Responsible AI', Trustworthiness, Trust
Privacy	<ul style="list-style-type: none"> ▶ Right: end users, data subjects ▶ Value: linked to trust in the AI industry 	<ul style="list-style-type: none"> ▶ Legal compliance with data protection regulation ▶ Technical measures for data safety ▶ Public awareness and information (e.g. data breaches...) 	data governance, freedom, autonomy, self-determination

Participant input 6: When applying, or thinking of applying, these ethics principles, what challenges might arise?



5.3 Applying ethics principles

Human rights impact assessments are tools that can help to implement the above-mentioned ethics principles and to apply these principles in fostering compliance with non-discrimination law and mitigating other unfair biases in algorithmic systems. Several such tools have been developed in the past years. This handbook only refers to two examples that are particularly relevant and that can provide inspiration for those wanting to develop a system involving AI or machine learning techniques.

The Dutch 2022 '[Fundamental Rights and Algorithms Impact Assessment](#)' (FRAIA) was developed by the Dutch Ministry of the Interior and Kingdom Relations, in cooperation with Utrecht University.⁵¹ FRAIA is a deliberation and decision-making tool that helps to map the risks to human rights in the use of algorithms and to take measures to address these risks. It is mainly addressed to governmental players and commissioned professionals working on the development or deployment of an algorithmic system. It contains a number of useful references to other tools such as ethics guidelines, specific impact assessment or auditing frameworks developed at national or European level. FRAIA comprises a roadmap of the various stages of preparation, development and deployment of algorithmic systems with sets of questions for each step. These questions can be categorised into

⁵¹ Gerards, J., Schäfer, M.T., Vankan, A. & Muis, I. Impact Assessment: Fundamental rights and algorithms (Netherlands Ministry of the Interior and Kingdom Relations, 2022), available at: <https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms>.

four core 'chapters'.⁵² A first set of questions ('**why?**') revolves around the rationales behind commissioning or developing an algorithmic system, the intended use and purpose of such a system and the problems it is meant to mitigate or solve. A second set of questions concerns the type of system to be developed and the data to be used in this context ('**what?**'). This section can be used to think about choices in terms of modelling, type of system, performance and accuracy, data collection, data quality and data representativeness, testing strategies, transparency and explainability, in particular in relation to bias issues. A third set of questions focuses on the deployment of the algorithmic system ('**how?**'). It prompts questions about the use of algorithmic outputs in decision-making and the involvement of human decision-makers, possible discriminatory effects of algorithmically supported decisions, procedural safeguards, considerations related to the context of use, accountability and auditing measures. Finally, the last set of questions pertains to possible breaches of '**fundamental rights**' by the system and available mitigation strategies. In particular, it can be used to think about whether a system breaches the fundamental right to equal treatment through direct or indirect discrimination against protected groups in light of applicable legislation and considerations of proportionality.⁵³

A second example of ethical implementation tool that can deal as a helpful reference for the users of this handbook is the '[Handbook on non-discriminating algorithms](#)' developed by a team of researchers from Tilburg University, Eindhoven University of Technology, Vrije Universiteit Brussel, and the Netherlands Institute for Human Rights, and commissioned by the Netherlands Ministry of the Interior.⁵⁴ The 'Handbook' can be used for thinking about algorithmic deployment in both the public and the private sector. It consists of ten 'rules' for ensuring 'non-discrimination by design' in the various stages of development and implementation of algorithmic decision-making tools. It invites thinking about the 'legal', 'technical' and 'organisational' aspects of the context of development and deployment of algorithmic tools and offers concrete examples. Throughout six 'phases' -- problem definition, data collection, data preparation, modelling, implementation, and evaluation -- it sets the following ten 'rules': stakeholder involvement, appropriate reflection and questioning, contextual assessment, bias awareness and testing, establishing clear objectives, monitoring throughout, expert involvement, assessment of indirect discrimination, legitimacy and documentation.⁵⁵

The aim of this section was to offer two examples of helpful tools that can be consulted and used as inspiration to implement the ethical principles covered in section 5.2. in relation to non-discrimination and the non-discrimination legal framework covered in section 3 of this handbook. Importantly, these and similar impact assessment and evaluation instruments should be used as deliberative support and thinking guidance, as opposed to sets of checkboxes. The aim is to think about open-ended questions that call for qualitative and reasoned answers.

⁵² As explained in: Ministry of the Interior and Kingdom Relations (March 2022), Impact Assessment Fundamental rights and algorithms, available at: <https://www.government.nl/binaries/government/documenten/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms/fundamental-rights-and-algorithms-impact-assessment-fraia.pdf>

⁵³ FRAIA is available open access at <https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms>.

⁵⁴ Van der Sloot, B., Keymolen, E., Noorman, M., Het College voor de Rechten van de Mens, Weerts, H., Wagenveld, Y. & Visser, B., Handreiking Non-discriminatie by design (Netherlands Ministry of the Interior, 2023), available at: <https://www.tilburguniversity.edu/nl/over/schools/law/departementen/tilt/onderzoek/handreiking>.

⁵⁵ The full Handbook and a short version are available at: <https://www.tilburguniversity.edu/about/schools/law/departments/tilt/research/handbook>.

List of References

- Adams-Prassl, J., Binns, R. and Kelly-Lyth, A. (2022). Directly Discriminatory Algorithms. *Modern Law Review*, Vol. 86, Issue 1, pages 144-175.
- Allhutter, D., Cech, F., Fischer, F., Grill, G., & Mager, A. (2020). Algorithmic profiling of job seekers in Austria: how austerity politics are made effective. *Frontiers in Big Data*, 5
- Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Sorelle A. Friedler, Christo Wilson, editors, *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA. Volume 81 of Proceedings of Machine Learning Research*, pages 77-91, PMLR, 2018
- Casili, A. "An 'End-to-End' Approach to Ethical AI", Institut Polytechnique de Paris. See Casili's AI talk @ ETUI, available at: <https://www.etui.org/events/ai-talks-etui-what-really-ethical-ai>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R., "CRISP-DM 1.0 step-by-step data mining guide," 2000
- Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. In arXiv [stat.ML]. Available at: <http://arxiv.org/abs/1808.00023>
- Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI). (2020). AI Ethics Guidelines: European and Global Perspectives. Provisional report by Marcello Ienca and Effy Vayena, available at: <https://rm.coe.int/cahai-2020-07-fin-en-report-ienca-vayena/16809eccac>
- Dastin, J. (2018), 'Amazon scraps secret AI recruiting tool that showed bias against women' (10 October), available at: www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
- European Commission, Directorate-General for Justice and Consumers, Gerards, J., Xenidis, R., (2021). Algorithmic discrimination in Europe: challenges and opportunities for gender equality and non-discrimination law, Publications Office, available at: <https://data.europa.eu/doi/10.2838/544956>
- European Parliament and Council, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) OJ L OJ L 2024/1689 [2024] <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Gerards, J., Schäfer, M.T., Vankan, A. & Muis, I. Impact Assessment: Fundamental rights and algorithms (Netherlands Ministry of the Interior and Kingdom Relations, 2022), available at: <https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms>
- Gupta, AH. (2019), 'Are Algorithms Sexist?', *The New York Times* (15 November), available at: www.nytimes.com/2019/11/15/us/apple-card-goldman-sachs.html
- Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, Vol. 55, Issue 4, pages 1143 – 1185
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 3315–3323. Available at : <https://arxiv.org/abs/1610.02413>
- Ienca, M., & Vayena, E. (2020). On the responsible use of digital data to tackle the COVID-19 pandemic. *Nature Medicine*, Volume 26(4), pages 463-464

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, Vol. 1, pages 389–399, available at: <https://doi.org/10.1038/s42256-019-0088-2>

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. *Machine Learning and Knowledge Discovery in Databases*, 35–50. Available at: https://link.springer.com/chapter/10.1007/978-3-642-33486-3_3

Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. Available at: <https://doi.org/10.1145/3313831.3376445>

Ministry of the Interior and Kingdom Relations (March 2022), Impact Assessment Fundamental rights and algorithms, available at: <https://www.government.nl/binaries/government/documenten/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms/fundamental-rights-and-algorithms-impact-assessment-fraia.pdf>

Timmer, A. (2016). Gender Stereotyping in the Case Law of the EU Court of Justice. *European Equality Law Review*, Issue 1, p. 38-9

Van der Sloot, B., Keymolen, E., Noorman, M., Het College voor de Rechten van de Mens, Weerts, H., Wagensveld, Y. & Visser, B., Handreiking Non-discriminatie by design (Netherlands Ministry of the Interior, 2023), available at: <https://www.tilburguniversity.edu/nl/over/schools/law/departementen/tilt/onderzoek/handreiking>

van Giffen, B., Herhausen, D., & Fahse, T. (May 2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, Vol. 144, pages 93-106, available at: <https://www.sciencedirect.com/science/article/pii/S0148296322000881>

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). Fairness Constraints: Mechanisms for Fair Classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340

Case law

C-414/16 - Egenberger (Judgment of the Court (Grand Chamber) of 17 April 2018, Vera Egenberger v Evangelisches Werk für Diakonie und Entwicklung e.V., ECLI:EU:C:2018:257)

C-177/88 - Dekker v Stichting Vormingscentrum voor Jong Volwassenen (Judgment of the Court of 8 November 1990, Elisabeth Johanna Pacifica Dekker v Stichting Vormingscentrum voor Jong Volwassenen (VJV-Centrum) Plus, ECLI:EU:C:1990:383)

C-668/15 - Jyske Finans (Judgment of the Court (First Chamber) of 6 April 2017, Jyske Finans A/S v Ligebehandlingsnævnet, acting on behalf of Ismar Huskic, ECLI:EU:C:2017:278)

C-443/15 - Parris (Judgment of the Court (First Chamber) of 24 November 2016, David L. Parris v Trinity College Dublin and Others, ECLI:EU:C:2016:897)

C-354/13 - FOA (Judgment of the Court (Fourth Chamber), 18 December 2014, Fag og Arbejde (FOA) v Kommunernes Landsforening (KL), ECLI:EU:C:2014:2463)