July 2026

Algorithmic fairness in the financial sector

Discussion paper

AUTHORS
Cyril Chhun, Olivier Fliche, Julien Uri
Directorate for Innovation, Data and Technological Risks

ACPR
BANQUE DE FRANCE

Summary

Algorithmic fairness refers to the set of principles and techniques aimed at designing, assessing and regulating algorithmic systems in order to prevent them from generating unjustified inequalities between individuals or groups, particularly where those inequalities are directly or indirectly linked to personal attributes considered "sensitive". In the financial sector, it touches upon a key question: how to **differentiate** between individuals on the basis of their level of risk – a prerequisite for business model sustainability – without, however, resorting to **unfair or even discriminatory treatment**, despite sensitive attributes being frequently **correlated with observed risks?**

To address this challenge, the long-dominant approach known as "**fairness-through-unawareness**" has consisted in excluding sensitive variables from statistical processing. While this method has always been subject to debate, it has now been **rendered largely obsolete by the development of artificial intelligence (AI) models**, which are capable of reconstructing the information contained in these variables. More broadly, the rapid growth of AI is reshaping how issues of algorithmic fairness are addressed, making it necessary to develop approaches that are more explicit, more robust and better suited to the complexity of current systems.

Against this backdrop, this discussion paper first sets out the **legal framework** for algorithmic fairness in the financial sector. It shows that although non-discrimination is a firmly established principle, its implementation is more complex when decisions are based on statistical models, and even more so on AI systems. The **European Regulation on Artificial Intelligence** (the "**AI Act**") **adds to this framework by setting out fairness requirements for so-called "high-risk" AI systems** and by reaffirming a principle of non-discrimination for all AI systems. In the financial sector, this regulation dovetails with the **rules on customer protection** in the banking and insurance industries, which also set out requirements for fairness, often based on a principle of "protection through abstention" (from granting or selling), whereas the AI Act places greater emphasis on the risks of exclusion. A comparison between the legal frameworks in other countries then highlights the diversity of approaches.

This paper then explains the **main concepts involved**, particularly distinguishing between three levels of analysis – disparity, bias and discrimination – in order to clarify notions that are often confused in debates on algorithmic fairness. **It emphasises that disparity does not necessarily constitute discrimination, as discrimination implies a normative and contextual judgement.** The paper then outlines the main **sources of bias** and sets out the **amplification potential** of certain models. More broadly, the paper stresses the **structural conflict between predictive performance and non-discrimination**, given that the relevant variables for measuring risk are often correlated with sensitive attributes. It also considers the distinction between **individual fairness** – treating comparable individuals in a similar manner – and **group fairness**, which aims to reduce disparities in treatment between groups. While individual fairness is theoretically appealing, especially in terms of performance, in practice the conditions for its implementation appear particularly demanding, which justifies **prioritising approaches based on group fairness**.

With regard to group fairness, the scientific literature distinguishes between **three main families of metrics: independence, separation and sufficiency**, which each correspond to a distinct **normative conception** (parity of outcomes, parity of error rates, or parity of decision reliability) and each lead to **different practical implications**. The literature also shows that when levels of risk differ between groups, it is impossible to satisfy all these different requirements

simultaneously. **Choosing a fairness metric** therefore necessarily involves striking a balance between competing objectives – fairness, performance and inclusion – which should be made explicit.

The paper then sets out the **methods for assessing and correcting algorithmic bias**. It first examines how to measure them, focusing on estimating the uncertainty that surrounds fairness metrics, particularly through the use of confidence intervals. It then lists the main **correction methods**, distinguishing between those applied upstream of the models (pre-processing of the data), during their training (modifying the optimisation function) or downstream (adjusting decisions). There are advantages and limitations in each of these approaches, which in practice leads to trade-offs between fairness, performance and implementation simplicity.

As for the concrete conditions for the **implementation** of algorithmic fairness in the financial sector, the paper stresses that it cannot be reduced to a technical issue solely at the discretion of the modelling teams. On the contrary, it is a **cross-cutting issue**, involving **strategic choices and trade-offs** that fall within a financial institution's overall responsibility. Considerations of fairness should be integrated at **every level of decision-making** – strategic, business and technical – and throughout the systems' lifecycle. This implies the **defining explicit objectives, documenting the choices made and implementing tailored control mechanisms**, consistent with existing model risk management frameworks.

The paper goes on to examine the main **operational choices** facing financial players: the identification and use of sensitive variables; the estimation of biases and the definition of analysis groups; the choice of metrics; the determination of thresholds; and the selection of methods for correcting biases. It **provides operational reference points to help inform these decisions**, while emphasising that they cannot be entirely standardised. On the contrary, they must be **adapted to the context** of use, the data available and the objectives pursued, based on a **proportionate and risk-based approach**.

Lastly, the paper briefly and prospectively addresses the case of **generative AI**, which is developing rapidly in the financial sector, even though it is not currently deployed on a large scale for use cases that are likely to pose significant algorithmic fairness challenges. It appears that the bias assessment methods designed for traditional predictive models **cannot be directly applied** to these new systems, whose operating methods and output formats differ significantly. Nevertheless, **methods for evaluating the fairness of a generative AI system have already been developed**, combining three layers (representational, behavioural and allocative) to be used concurrently.

Contents

Introduction

Algorithmic fairness refers to the set of principles and techniques aimed at designing, assessing and regulating algorithmic systems in order to prevent them from generating **unjustified inequalities** between individuals or groups, particularly where those inequalities are directly or indirectly linked to personal attributes considered "sensitive".

In the financial sector, the issue of algorithmic fairness is far from new. It has been a concern since statistical models were first used to inform decision-making (granting loans, insurance pricing, etc.). Fundamentally, the question is how to **differentiate** between individuals on the basis of their level of risk – a prerequisite for business model sustainability – **without, however, resorting to unfair or even discriminatory treatment**, despite sensitive attributes being frequently **correlated with observed risks**.

To address this challenge, the long-dominant approach known as "**fairness-through-unawareness**" has consisted in excluding sensitive variables from statistical processing. Although this approach has always been somewhat contentious, it has now been rendered **largely obsolete** by the development of models based on **artificial intelligence (AI)**: the **high dimensionality** of the data and **strong collinearity** between variables mean that AI algorithms can identify **proxy variables**, enabling them, in practice, to reconstitute the information contained in the sensitive variables.

The rapid growth of AI is thus profoundly transforming the ways in which issues of fairness are addressed. On the one hand, AI-based models are more powerful, which could theoretically help to reduce the risk of unfair treatment, but on the other hand, the mechanisms underlying discrimination become more diffuse and harder to identify and interpret, due to the complexity and opacity of the systems.

Furthermore, Regulation (EU) 2024/1689 on Artificial Intelligence (the "**AI Act**") sets out **fairness requirements** for so-called "high-risk" AI systems, while more broadly reaffirming the **principle of non-discrimination** enshrined in the Charter of Fundamental Rights of the European Union (EU) for all AI systems deployed within the EU.

In this context, the analytical frameworks and operational tools used to address the issue of fairness in the financial sector need to be updated.

In the spring and autumn of 2025, the *Autorité de contrôle prudentiel et de résolution* (ACPR – the French Prudential Supervision and Resolution Authority) held a **series of technical workshops** with volunteer financial sector participants to gain insights into these issues. Their aim was to understand how the banks and insurers surveyed dealt with issues of fairness in their processes and governance on a concrete basis. The discussions notably revealed that institutions have **high expectations** of public authorities – and financial supervisors in particular – to provide clarification and precision on the implementation of the applicable rules on fairness.

This paper aims to provide a **structured analytical framework of the issues of algorithmic fairness in the financial sector** by bringing together the contributions from scientific literature, the legal requirements in force and observed on-the-ground practices. It also aims to provide **operational reference points** to help players in the sector understand and effectively implement fairness requirements in their processes.

This paper is divided into **six sections**. The **first section** presents the **legal framework** applicable to fairness in the financial sector and draws together non-discrimination rules, sector-specific regulations and requirements introduced by the AI Act. The **second section** offers **clarification on the concepts of bias and fairness**, notably making a distinction between group fairness and individual fairness, and also their conceptual bases. The **third section** details the **main families of group fairness metrics** found in the scientific literature – independence, separation and sufficiency – by analysing their properties, advantages and limitations. The **fourth section** identifies the **main methods for assessing and correcting bias**. The **fifth section** addresses the **challenges associated with practical implementation**, offering methodological guidance for fairness governance within institutions, for selecting relevant metrics and defining appropriate thresholds.

This discussion paper focuses primarily on "traditional" predictive systems. Indeed, these currently constitute the bulk of systems deployed at scale in the financial sector that may give rise to fairness concerns. The case of **generative AI**, which is developing rapidly, is addressed at the end of the paper, in a **sixth section**: it appears that the bias assessment methods designed for traditional predictive models cannot be directly applied to these new systems, whose operating methods and output formats differ significantly.

This discussion paper was drafted by the ACPR's technology risk department, based on a review of the scientific literature and the discussions mentioned above. **It is not intended to provide an exhaustive view of all the issues relating to algorithmic fairness, nor to express an official position taken by the ACPR.** Its aim is to present an initial analysis of the ways in which algorithmic fairness requirements can be implemented, with a view to discussing them with stakeholders, particularly from within the profession, during public consultation.

The authors would like to thank the reviewers of this paper for their invaluable contributions, and in particular Jean-Michel Loubès (Inria), Félicien Vallet and Maxence Gérard (CNIL), Samy Chali and Shaden Shabayek (PEReN).

1 Algorithmic fairness requirements for financial sector firms

1.1 The principle of non-discrimination and sensitive data under law

1.1.1 Discrimination and differentiation

The principle of non-discrimination is widely enshrined in the different legal systems. It derives primarily from the **principle of equality**. Under French law, Article 1 of the Constitution of 4 October 1958 provides that the Republic "*shall ensure the equality of all citizens before the law, without distinction of origin, race or religion. It shall respect all beliefs*".[1] This principle is further clarified in Article 225-1 of the French Penal Code, which states that, on the basis of a number of **protected characteristics**,[2] "*any distinction made between natural [and legal] persons constitutes discrimination*".

Under European law, Article 20 of the **Charter of Fundamental Rights of the European Union**, which became legally binding following the entry into force of the Treaty of Lisbon on 1 December 2009, states, "*Everyone is equal before the law.*" Article 21 of the Charter provides that "*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*". Non-discrimination is also upheld by the **European Convention on Human Rights**, in particular Article 14.[3]

While the list of protected characteristics may vary slightly from one text to another, these various sources all agree on one same requirement: prohibiting discrimination based on protected personal characteristics. In the financial sector, this prohibition specifically targets unfair customer segmentation (see Section 2), the automatic refusal to enter into a relationship, or the application of less favourable terms based on personal criteria with no direct link to the actual risk.

However, the **principle of non-discrimination does not prohibit all differences in treatment between individuals**, provided that those differences are based on **objective and relevant criteria**, particularly economic criteria, linked to the nature of the decision being taken. The principle of equality may indeed lead to **differing treatment of objectively distinct situations,** in order to ensure effective equality. This approach, consistently upheld by European case law, requires a case-by-case assessment of situations.[4] Thus, a price differentiation or a denial (e.g.

---

[1] Furthermore, the Preamble to the 1946 Constitution, recognised by the French Constitutional Council as a text of "constitutional value", states in its first paragraph that "*each human being, without distinction of race, religion or creed, possesses sacred and inalienable rights*". Furthermore, "*the law guarantees women equal rights to those of men in all spheres*" (paragraph 3).

[2] Article 225-1 of the *Code Pénal* (French Penal Code) lists 26 protected characteristics in France.

[3] Article 14 of the Convention: "*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.*"

[4] European Court of Human Rights, 13 November 2007, D.H. and Others v. the Czech Republic, No. 57325/00, § 175: "*The Court has established in its case-law that discrimination means treating differently, without an objective and reasonable justification, persons in relevantly similar situations (see Willis v. the*

loan, claim, etc.) may be lawful where it is based on demonstrable economic factors (income, financial stability, indebtedness, behaviour presenting risks, etc.), but becomes unlawful if it is based directly or indirectly on protected characteristics.

Box 1: Fairness and competition in the insurance sector

Insurance operates on a simple principle: **risk pooling** – everyone's contribution helps to compensate the losses suffered by the least fortunate. In a given market, the insurer must therefore set a total level of premiums that is sufficient to cover the risk. Under an insurance **monopoly**, it would therefore suffice to apply a single rate to the entire population (equal to the level of premiums required divided by the number of policyholders).

However, in a **competitive** insurance market, **customer segmentation** – grouping policyholders into homogeneous risk classes – can enable an insurer to offer more favourable rates to a below-average-risk group, thereby capturing a larger market share. Therefore, it is above all the competition between insurers that leads them to offer different rates to each group of policyholders.

Consequently, each insurer must be able to **optimally segment** the various groups according to levels of risk: to this end, a key role of an actuary is to **select the most relevant variables**.

1.1.2 The problem of algorithmic discrimination

While this legal framework appears to be clearly defined in principle, **its implementation raises difficulties in practice**, due to the complexity of the mechanisms through which discrimination may occur.

These difficulties stem particularly from the fact that discriminatory mechanisms can arise unintentionally. European legislation thus distinguishes between *direct discrimination*, based explicitly on a protected characteristic, and *indirect discrimination*, which results from a provision or practice that is apparently neutral but that is likely to place certain people at a particular disadvantage. This distinction shows that discrimination can be a by-product of certain rules or practices, regardless of discriminatory intent.

This point is particularly important in the case of **algorithmic systems use**. **Algorithmic discrimination** refers to situations in which an automated system produces differences in treatment between individuals or groups due to the data used, the model chosen or the variables selected, without any explicit discriminatory intent.[5] Compared to discrimination in the traditional legal sense, it is characterised by more **indirect and diffuse mechanisms**, which may arise at all stages of the design and use of systems, and which may lead to the reproduction – or even amplification – of pre-existing discrimination. Moreover, such discrimination may take on an

United Kingdom, no. 36042/97, § 48, ECHR 2002-IV, and Okpisz v. Germany, no. 59140/00, § 33, 25 October 2005)."

[5] Défenseur des Droits, *Lutter contre les discriminations produites par les algorithmes et l'IA*, February 2024 (in French only): "*Discriminatory mechanisms are frequently based on the bias of the data selected and used by an algorithm. This bias may be linked to a lack of representativity in the data in relation to the context in which the algorithm is to be deployed. It may also be linked to the fact that the data is the mathematical result of past often discriminatory practices and behaviour and of systemic discrimination present in society*."

intersectional dimension, combining multiple protected characteristics, which further complicates its identification.[6]

The **core issue underlying model-driven** discrimination lies in the fact that **protected variables** (such as ethnic origin, gender, age or place of residence) **are often correlated with financial risk, not because of a direct causal link, but due to indirect mechanisms linked to pre-existing social and economic inequalities**. For example, structural inequalities in education, employment or access to credit can lead certain groups to exhibit statistically different risk profiles, which the model then learns and exploits. Even when protected variables are explicitly excluded, proxy variables[7] such as income, contract type, banking history, geographical area, etc. may be sufficient to reconstitute sensitive information and reproduce discriminatory effects. The model thus faces a **fundamental conflict**: ignoring these correlations may undermine predictive performance, but incorporating them amounts to institutionalising historical disadvantages, by treating as "neutral" signals that reflect social inequalities rather than intrinsic risk.

These difficulties are particularly pronounced in the use of AI systems, due to their **complexity** and **opacity**. This notably results in increased challenges with regard to the **burden of proof**, as demonstrating discrimination may prove difficult if those affected lack access to the data or models, or even due to a lack of transparency regarding the very use of algorithmic processing. However, **algorithmic discrimination is not subject to a specific legal framework**, as the relevant general rules are intended to apply regardless of the technology used.

1.1.3 Protected characteristics and sensitive variables

Nevertheless, the **use of personal data** by automated systems is subject to a specific legal framework. In this respect, the **General Data Protection Regulation** (GDPR)[8] is the authoritative legal reference text; it regulates the collection, use and protection of personal data, in order to strengthen individuals' rights and make the organisations that process such data more accountable. It is based on several fundamental principles, such as the lawfulness and transparency of treatment, data minimisation, purpose limitation, security and accountability of data controllers.

The GDPR notably identifies "special categories of personal data" – commonly referred to as "**sensitive data**" – which require **enhanced protection**. These include data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, sexual orientation, or related to health (see the table 1 below for the full list). **Processing these data is, in principle, prohibited**, except in a limited number of cases strictly governed by the GDPR, such as the data subject giving explicit consent or a requirement to comply with specific legal obligations.

The protected characteristics in terms of discrimination and sensitive data within the meaning of the GDPR only partially overlap. Some data fall into both categories, such as ethnic origin or religious beliefs. Other data, however, are protected under non-discrimination law but are not classified as sensitive by the GDPR: this is particularly the case for age or gender, which

[6] Intersectional discrimination eludes traditional detection methods, as it results from the combination of several protected characteristics and is not necessarily visible when analysed separately.
[7] In modelling, a proxy variable acts as a substitute for another variable, which is generally more difficult to collect or measure.
[8] 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.

are considered “ordinary personal data”. Conversely, the GDPR may restrict the processing of certain data that are not directly linked to a risk of discrimination, because of their particularly intrusive nature to a person’s private life. By way of illustration, Table 1 below provides an overview of the situation under European law.

Table 1: Protected characteristics and sensitive data under European law

Category	Charter of Fundamental Rights of the European Union (protected characteristics)	GDPR (Sensitive data)	Comment
Ethnic origin	✓	✓	Direct coverage
Religion/beliefs	✓	✓	Direct coverage
Political opinions	✓	✓	Direct coverage
Sexual orientation	✓	✓	Direct coverage
Health/disability	✓	✓	Partial coverage[9]
Genetic data	✓	✓	Direct coverage
Biometric data	✗	✓	GDPR protected only
Trade union membership	✗	✓	GDPR protected only
Sex	✓	✗	Charter protected only
Age	✓	✗	Charter protected only
Nationality	✓	✗	Charter protected only
Social origin/property	✓	✗	Charter protected only
Language	✓	✗	Charter protected only

In fact, the legal framework for the protection of fundamental rights – in terms of non-discrimination – and the legal framework for the protection of personal data pursue **complementary but distinct objectives**, which in practice results in different approaches. Non-discrimination law focuses primarily on the **effects** of decisions (equal treatment between groups), while the GDPR regulates the **means** (data collection and use). For AI systems in the financial sector, the interplay between the two requires a balance between controlling the data used for training and exercising greater vigilance over the results produced.

Lastly, it should be noted that this table **does not consider proxy variables**, which, without formally being protected characteristics, may be **strongly correlated** with them. For example, postcodes or place of residence are not protected characteristics or sensitive data, but they may be strongly correlated with income, ethnic or social origin, or political opinions. Using proxy variables may therefore raise legal issues relating to non-discrimination when they lead, even

[9] Disability in the Charter, health in the GDPR. However, a person's state of health is a protected characteristic under French law.

indirectly, to differences in treatment based on protected characteristics. Thus, in practice, legal risk assessment **cannot be limited** solely to the apparent nature of the variables used, but must also consider their potential as proxies.

1.2 The AI Act: requirements regarding fairness

The European AI Act,[10] which came into force in August 2024, introduced a uniform regulatory framework for AI,[11] with the twofold objective of protecting citizens' health, safety and fundamental rights, and promoting the development of a single European market for "trustworthy AI". The AI Act categorises AI systems according to their risk: the core of its provisions concerns so-called "high-risk" systems, defined primarily by their purpose, and which apply to the financial sector in two use cases.[12]

With regard to fairness,[13] the AI Act first reaffirms a simple principle: the **right to non-discrimination** is one of the fundamental rights protected by European law; **it must therefore also be respected by all AI systems deployed in the EU**.[14] In fact, the right to non-discrimination is one of the arguments used to prohibit manipulative systems (Recital 28) or social scoring (Recital 31). The risk of discrimination is also cited to explain the classification of AI systems used to evaluate the creditworthiness of natural persons as high-risk,[15] as well as in the case of systems intended for risk assessment and pricing for health and life insurance.[16]

However, the provisions relating to algorithmic fairness are only really specified for high-risk systems.[17] The central provision of the AI Act on the issue of bias and discrimination is found in

---

[10] Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence came into force on 1 August 2024.

[11] In particular, the AI Act distinguishes between providers and deployers of AI systems: the provider of an AI system is the entity that developed it and placed it on the market or put it into service, and the deployer of that AI system is an entity that uses it for a business activity. The majority of the requirements of the AI Act apply to providers.

[12] Annex III, (5b): "*AI systems 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*"; and (5c): "*AI systems intended to be used for risk assessment and pricing in relation to natural persons in the case of life and health insurance*".

[13] The term "fairness" is only marginally referred to in the AI Act, which makes more frequent use of the terms "bias" and "discrimination".

[14] See, for example, Recital 27 of the Act: "*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*".

[15] "*AI systems used for those purposes may lead to discrimination between persons or groups and may perpetuate historical patterns of discrimination, such as that based on racial or ethnic origins, gender, disabilities, age or sexual orientation, or may create new forms of discriminatory impacts*" (AI Act, Recital 58).

[16] "*Moreover, AI systems intended to be used for risk assessment and pricing in relation to natural persons for health and life insurance can also have a significant impact on persons' livelihood and if not duly designed, developed and used, can infringe their fundamental rights and can lead to serious consequences for people's life and health, including financial exclusion and discrimination*" (AI Act, Recital 58).

[17] Lastly, it should be noted that the issue of bias is also addressed in the case of general-purpose AI models: the technical documentation drawn up by their providers must include "[...*] all other measures to detect the unsuitability of data sources and methods to detect identifiable biases*" (Annex XI of the AI Act). At a minimum, it aims to provide stakeholders further down the value chain – such as potential providers of high-risk systems built on general-purpose models – with the means to comply with their regulatory obligations.

Article 10, which deals with requirements relating to data and data governance.[18] Thus, **Article 10(2)(f)** provides that "*the training, validation and testing datasets*" of high-risk systems are subject to "*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*". Furthermore, **Article 10(2)(g)** adds the requirement for "*appropriate measures to detect, prevent and mitigate possible biases identified according to point (f)*".

Articles 10(2)(f) and 10(2)(g) therefore formalise a requirement for fairness in high-risk systems under the AI Act, covering the entire chain of processing biases that could lead to discrimination (prevention, detection and mitigation).[19] Furthermore, although the wording of Article 10(2) appears to refer only to the input data of the models ("*training, validation and testing datasets*"), its requirements can in reality only relate to the outputs of AI systems (or the decisions resulting from them), as they concern the potential effects of their use.[20]

Lastly, the principle of **human oversight** set out in the AI Act theoretically constitutes another means of combating algorithmic discrimination, by requiring that users be able to detect abnormal or unjustified results and, where necessary, to **intervene to correct, suspend or invalidate an automated decision**. Human oversight should thus act as a safeguard where a purely automated decision-making process could undermine the principle of equal treatment.

In practice, however, there is a risk that human oversight is insufficient to fully prevent algorithmic discrimination: on the one hand, **automation bias** often leads human operators to place excessive trust in the recommendations issued by AI systems, even when they are erroneous or questionable, especially if the system is perceived as being technically complex or objectively superior.[21] Human oversight can then become a mere formality, without properly calling into question the decisions produced by the machine. On the other hand, **human decision-makers are not necessarily neutral**: they themselves exhibit cognitive, social or cultural biases that may be equal to, or even greater than, those of the algorithms, and which are likely to sway their judgements.

[18] It should be noted that CEN and CENELEC are currently working on standardisation at the European level. At the request of the European Commission, they are developing harmonised standards to further clarify the requirements of the AI Act, including in relation to algorithmic fairness. However, their work has fallen behind the originally planned schedule. Furthermore, as the standards are intended to be cross-sectoral, they are unlikely to address issues that are specific to the financial sector.

[19] The identification of any discriminatory effects must also be included in the technical documentation for high-risk systems, as detailed in Annex IV of the Act.

[20] Furthermore, model outputs are included in the required examination, as they may form the basis for subsequent retraining of the model (Article 10(2)(f)). This point is confirmed in Article 15(4): "*High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures.*"

[21] To combat this phenomenon, Article 14(4)(b) of the AI Act stipulates that AI systems must be provided in such a way that human overseers remain aware of the existence of this type of bias.

1.3 Other regulatory sources

1.3.1 The banking sector

In the banking sector, **prudential regulations** – in particular the CRR Regulation[22] and the European Banking Authority (EBA) Guidelines on loan origination and monitoring[23] – address the issue of statistical bias but not that of potential discrimination bias. In fact, the purpose of prudential regulation is to safeguard the stability of the financial system; it therefore focuses on **risks to institutions** rather than the risks that might affect customers.

However, European regulations on **consumer protection** in the banking sector, in particular the mortgage credit directive (MCD)[24] and consumer credit directive (CCD),[25] contain provisions relating to fairness. They apply to **all consumer lending models**, which encompass a broader range of use cases than the high-risk systems referred to in the AI Act.

The guarantees take **two forms**. First, professionals (creditors and credit intermediaries) must act **honestly, fairly, transparently and professionally** and take account of the rights and interests of consumers, over the entire lifetime of the credit product (from product design to contract execution).[26] Second, a borrower's **assessment of creditworthiness** is subject to specific rules: it must be based on information that is **relevant, accurate, necessary and proportionate** to the characteristics of the credit, focusing particularly on the consumer's income, expenditure and financial situation.[27] Obtaining information from **social networks** is strictly prohibited.

It is important to note that, under customer protection regulations, the creditworthiness assessments carried out by banks are above all in the **consumer's interest**.[28] Seen from this angle, it is the granting of a loan, rather than its refusal, that constitutes the main risk for the consumer, who may be unable to repay it or become overly indebted. This logic of "**protecting through abstention**", which aims to limit access to credit to prevent excessive indebtedness, differs significantly from the **access-based approach** of the AI Act. The latter emphasises the opposite risk of being **deprived of opportunity** and seeks to prevent borrowers that are in fact creditworthy from being unduly excluded from access to credit.

---

[22] Article 174 of Regulation (EU) 575/2013 on prudential requirements for credit institutions and investment firms (CRR) thus provides that *"[i]f an institution uses statistical models and other mechanical methods to assign exposures to obligors or facilities grades or pools, the following requirements shall be met: (a) the model shall have good predictive power and capital requirements shall not be distorted as a result of its use. The input variables shall form a reasonable and effective basis for the resulting predictions. The model shall not have material biases; [...] (c) the data used to build the model shall be representative of the population of the institution's actual obligors or exposures"*.

[23] See in particular paragraphs 53(e), 54(a) and 55(a).

[24] Directive 2014/17/EU of 4 February 2014 on credit agreements for consumers relating to residential immovable property.

[25] Directive (EU) 2023/2225 of 18 October 2023 on credit agreements for consumers.

[26] Article 32 of the CCD and Article 7(1) of the MCD.

[27] Articles 18(1) and 18(3) of the CCD and Article 20(1) of the MCD.

[28] Prudential regulations also require an assessment of creditworthiness to safeguard an institution's own solvency.

1.3.2 The insurance sector

European regulations for the insurance sector share similar features. In this instance too, the rules on customer protection – in particular the **Insurance Distribution Directive** (IDD)[29] – establish requirements for fairness and those requirements apply to a scope of AI systems broader than those classified as high-risk under the AI Act.

Thus, the IDD requires all insurance distributors to act **honestly, fairly and professionally, in accordance with the best interests of their customers**.[30] The IDD also introduces **product governance** requirements to ensure that the products sold meet the needs and characteristics of the customers for whom they are intended, thereby reducing the risk of mis-selling. In particular, insurers must identify a **target market** for each of their products and segment their customer base to minimise the risk of mis-selling or abusive sales practices.

Again, as in the banking sector, insurance regulations on customer protection are therefore primarily designed to protect customers through abstention (from selling), in contrast to the access-based approach of the AI Act. However, this contrast must be **qualified**, as **certain types of insurance**, such as home or car insurance, **are compulsory**.[31] Consequently, **specific measures** have been put in place under French law to guarantee an effective right of access to insurance, such as regulated pricing mechanisms or the intervention of central pricing offices, which help to limit insurance refusals.

It is also worth noting that the European Insurance and Occupational Pensions Authority (EIOPA) published its ***Opinion on AI governance and risk management*** in 2025.[32] Although the Opinion is not legally binding, it notably recommends that insurance organisations: (i) identify and, where possible, remove or at least mitigate potential bias, including proxy discriminatory variables; (ii) regularly monitor AI systems, in particular by **using fairness and non-discrimination metrics**;[33] and (iii) develop internal guidelines and training on fairness for their staff.

[29] Directive (EU) 2016/97 of 20 January 2016 on insurance distribution.
[30] Article 17(1) of the IDD.
[31] The right to a bank account follows the same logic in the banking sector. In France, anyone without a bank account has the right to obtain one from an institution designated by the Banque de France. This scheme is intended to ensure financial inclusion by guaranteeing everyone access to essential banking services.
[32] EIOPA, Opinion on AI Governance and Risk Management, 6 August 2025.
[33] Annex I of the Opinion provides several examples of fairness metrics.

Box 2: The case of gender-based pricing in insurance

The Test-Achats judgment of the Court of Justice of the European Union (CJEU) of 2011[34] resulted in the **prohibition**, effective from December 2012, of **all pricing based on gender in insurance, whether in relation to calculating premiums or setting the level of benefits**. The CJEU held that differences in pricing between men and women infringed the principle of equality protected under EU law, even where they were based on statistical data (for example, in the motor insurance sector, women statistically have fewer accidents than men), insofar as they relied on a protected personal characteristic. From December 2012, insurers were therefore required to revise their actuarial models to eliminate any segmentation based on gender, which led to a redistribution of costs among policyholders.

This prohibition theoretically extends to variables used as proxies for gender, which must be removed from models, unless their use is justified by a legitimate objective and is appropriate and necessary. The European Commission illustrates this situation with the following examples: price differentiation based on the size of a car engine in the field of motor insurance should remain possible, even if statistically men drive cars with more powerful engines, as engine size is **directly correlated with risk**. By contrast, it is prohibited to apply price differentiation based on the size of a person in motor insurance, as men are generally taller than women, and this difference has **no objective correlation with risk**.[35]

1.4 Financial companies' risk management policy and internal policy

Financial sector companies take fairness into account at two complementary levels: first, through **compliance policies**, which set out the applicable legal obligations internally, and second, through **additional voluntary frameworks**, such as ethical charters or AI governance principles, which reflect a company's trade-offs between performance, profitability and fairness.

These trade-offs are of **particular consequence in the insurance sector**, where the logic of risk segmentation forms the very foundations of the business model: creating homogeneous risk classes so that premiums can be adjusted to an expected level of risk necessarily leads to differentiation between policyholders and may result in deviations in treatment between individuals or groups (see Section 2). As such, internal frameworks of insurance undertakings – whether they relate to compliance or ethical considerations – play a central role in framing risk assessment and pricing by defining permissible variables, organising bias controls, and setting acceptable limits on potential deviations in treatment between individuals or social groups.

In the banking sector, similar issues may arise, particularly in creditworthiness assessment and lending processes, but the **conflict between fairness and differentiation is more diffuse** and is more a matter of risk management rather than a direct consequence of the business model.

[34] Court of Justice of the European Union (CJEU), Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des Ministres, C-236/09, 1 March 2011.
[35] European Commission guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats).

1.5 Regulatory frameworks outside the European Union

1.5.1 The United States

In the United States, the legal approach to fairness primarily falls within the broader non-discrimination framework, which stems in particular from the Civil Rights Act of 1964. The Civil Rights Act shaped two main concepts of discrimination: first, **disparate treatment**, which refers to intentional discrimination by which an individual is treated differently because they belong to a protected group; and second, **disparate impact**, which refers to situations where a seemingly neutral policy or action causes, in practice, disproportionate effects on certain groups. This second concept has historically played an important role in assessing the fairness of models, as it allows for the identification of discrimination regardless of explicit intent.

In the financial sector, these general principles have been enshrined in **specific legislation**, in particular in the Equal Credit Opportunity Act of 1974. This law prohibits any discrimination in the granting of credit, regardless of the technology used,[36] and requires lenders to justify any adverse decision. This law is complemented by the Fair Credit Reporting Act of 1970 as it regulates the use of data in credit decisions (considerations of quality and accuracy, privacy, right of access, etc.).

In practice, US authorities have historically relied on empirical tools to **assess** disparate impact, particularly in the field of employment. The Uniform Guidelines on Employee Selection Procedures thus stipulate, in Section 4(D), that "*a selection rate for any race,[37] sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact*". This empirical rule has gradually been applied in a **wide range of fields**, and notably the financial sector, and has inspired, more or less explicitly, certain contemporary practices for measuring algorithmic fairness (see Section 3).

It should also be noted that recent developments have marked a **significant shift in the approach** to fairness in the United States. An executive order signed by President Trump in August 2025[38] prompted federal agencies to **abandon** the use of "disparate impact" in their supervisory activities. Agencies such as the Office of the Comptroller of the Currency (OCC)[39] and the Consumer Financial Protection Bureau (CFPB)[40] subsequently announced that they would no longer refer to the concept. This shift could therefore reduce the use of fairness metrics in favour of an approach that centres on evidence of intentional discrimination.

1.5.2 The United Kingdom

In the United Kingdom, assessments of algorithmic discrimination are primarily based on the general legal framework under the Equality Act 2010,[41] which sets out the characteristics used to

[36] Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms | Consumer Financial Protection Bureau.
[37] Term to be considered in a US context.
[38] Executive order 14281: Federal Register: Restoring Equality of Opportunity and Meritocracy.
[39] Fair Lending: Removing References to Disparate Impact | OCC.
[40] Fair Lending Report of the Consumer Financial Protection Bureau for 2024.
[41] It extends and builds upon a number of earlier domestic law, in particular the Sex Discrimination Act 1975 and the Race Relations Act 1976, which laid the initial foundations for combating discrimination in the United Kingdom.

identify discriminatory practices. In particular, the concept of **indirect discrimination** (Section 19 of the Equality Act) plays a central role: it covers situations in which apparently neutral provisions or models produce, in practice, disproportionate effects on certain groups without objective justification – particularly relevant to AI systems, whose biases are often systemic and unintentional. This framework does not include **standardised thresholds or metrics** to characterise the effects: analysis relies largely on a case-by-case assessment, combining empirical evidence and legal reasoning, and thus leaves judges and regulators significant scope for interpretation.

In the financial sector, the Financial Conduct Authority (FCA) applies these principles by adopting an approach focused on the **practical impacts on consumers**, particularly within the framework known as the "Consumer Duty".[42] Fairness assessment is not limited to formal model compliance, but also considers the potential effects arising from their use (e.g. the exclusion of certain profiles from access to credit, less favourable terms for vulnerable groups, or excessive market segmentation). Recent FCA guidance emphasises that AI systems must neither violate individual rights nor generate unjustified discrimination, and that they must be designed taking into account the fairness criteria suitable to their context of use. This approach is complemented by **oversight of model governance processes**, in particular an institution's ability to identify, measure and correct bias throughout the systems' lifecycle.

1.5.3 Singapore

In Singapore, fairness regulation in the financial sector is based on a **principles and operational** approach and largely structured around the role of the Monetary Authority of Singapore (MAS). In 2018, the MAS laid the foundations of this approach with the **FEAT** (Fairness, Ethics, Accountability, Transparency) principles, which provide a reference framework for the use of AI and data in the financial services. Though non-binding, the FEAT principles are highly influential and are intended to ensure that systems do not produce **unjustified differentiation at a systemic level** for certain individuals or groups, while also imposing requirements on governance, responsibility for decision-making and explainability.[43] In particular, the MAS asserts that "fairness-through-unawareness" is no longer suitable for AI models (see Section 2.2).

To put these principles into operation, the MAS – in partnership with the financial sector – has developed the **Veritas Initiative**; one of the most advanced schemes internationally for assessing algorithmic fairness. At Veritas' core is a **structured assessment methodology** that covers the entire lifecycle of AI systems (design, development, deployment, monitoring) and is designed to enable financial institutions to translate the FEAT principles into operational processes.

This methodology includes tools to identify sensitive variables, detect and measure bias, select the most relevant fairness metric (including through decision trees) and document any trade-offs made. It has gradually been incorporated into an **open-source toolkit**,[44] developed by a consortium of public and private sector actors, which enables certain analyses (such as the calculation of fairness metrics) to be automated and assessment practices to be standardised.

---

[42] Consumer Duty | FCA.

[43] This approach goes hand-in-hand with broader guidelines on fair dealing, which require institutions to design and distribute financial products tailored to clients' needs, to explain their decisions, and to be able to justify any differential treatment between categories of clients.

[44] Veritas Toolkit 2.0, open-source resource.

A striking feature of the Veritas approach is its **holistic and contextual** nature: fairness is not reduced to a single statistical indicator, but is integrated into a far broader set of considerations: ethics, governance and transparency. The proposed methods thus combine both quantitative (bias tests, inter-group comparisons) and qualitative (justification of design choices, risk documentation, internal control mechanisms) analyses. Furthermore, the MAS insists on a **risk-based** approach, in which the level of requirements depends on the potential impact of the system.

2 The concept of fairness in the financial sector

2.1 The concept of discrimination bias

When discussing fairness, the terms “bias” and “discrimination” are sometimes used imprecisely or interchangeably, which risks blurring the analysis. To avoid confusion, this paper distinguishes between **three levels of analysis**, corresponding to different conceptual and operational realities.[45]

A **disparity** is a systematic difference in the behaviour, outputs or performance of a model across groups. Disparities are frequent and are not necessarily unjustified. For example, a model for granting loans that uses income as a variable will, by design, produce different score distributions for groups with different income distributions. A disparity is therefore a simple statistical observation: it does not imply the existence of bias, let alone discrimination.

A **bias** is a disparity that reflects a systematic deviation from a given assessment standard. This standard is often of accuracy (e.g. a model making more errors for a particular group), but it may also relate to service quality (e.g. conversational AI understanding certain customer groups less than others; see the final section of this paper on generative AI), or to robustness (e.g. a model proving to be more fragile when applied to certain populations). Characterising bias therefore requires specifying the standard against which the deviation is assessed: it is not disparity in itself that constitutes bias, but rather the deviation it exhibits from a given standard.

Discrimination is a bias that is considered unacceptable given context of use, the nature and extent of the harm (actual or potential), the rights affected, and the applicable legal framework. For a supervisory authority, the underlying standard is a legal standard, based on characteristics protected by European or national law (see Section 1.1). For a financial institution, this standard may also stem from its internal policy: a company may decide to go beyond what is legally required by incorporating broader ethical considerations[46] into its assessment of fairness, for example.

From an operational perspective, these three concepts correspond to distinct levels of analysis. A disparity can be observed and measured statistically, subject to issues of uncertainty (see Section 4.1). Identifying a bias also involves specifying the standard used for the assessment and justifying its relevance to the use case. Lastly, qualifying a case as discrimination involves a contextual judgement on the acceptability of the deviation, taking into account the issues at stake, the rights involved and the legal framework. Confusing these different levels can mean that all disparities are classified as discrimination or, conversely, that discrimination is only recognised in extreme or intentional cases.

These distinctions clarify the legal concepts of differentiation and discrimination introduced in Section 1.1. Differentiation based on objective and relevant characteristics constitutes a legitimate disparity with regard to the use case. Direct discrimination refers to bias based explicitly on a protected characteristic. Indirect discrimination, on the other hand, results from

[45] Loubes et al., 2026.

[46] Two ethical standards are traditionally contrasted in scientific research: the world as it is, and the world as it should be.

correlation (or proxy) effects between variables; its legal classification depends on the existence of an objective and proportionate justification.[47]

2.2 Algorithmic bias and Big Data in the financial sector

Fundamentally, the issue of fairness in the financial sector is caught in a **structural conflict** between (i) the **economic necessity of differentiating** between individuals on the basis of risk, in order to ensure the sustainability of business models and (ii) the **prohibition of discrimination** based on non-objective or protected characteristics. This conflict is made all the more acute by the fact that certain protected variables **statistically correlate with observed risks**, due to indirect causal mechanisms, which can mean that their exclusion can be detrimental in terms of predictive accuracy.

To counter this difficulty, an approach that has long prevailed has been to "protect" sensitive variables by excluding them from statistical treatment.[48] **Today, this "fairness-through-unawareness" approach, has been rendered obsolete by the high dimensionality of datasets, which results in strong collinearity between protected and unprotected variables.** Algorithms can thus detect **proxy variables** that enable them to "reconstitute" the information contained in protected variables. For example, postcode, type of telephone, or spending habits can be strongly correlated with a protected variable, such as ethnic origin.

An AI model can therefore accurately discriminate between individuals without ever "seeing" the sensitive variable. Moreover, this indirect discrimination can sometimes be less transparent and less controllable than explicit discrimination. **Discrimination can thus arise as a collateral consequence of big data processing.** In this context, some research proposes that collection and use of sensitive variables should not be prohibited, but instead should be used to combat discrimination[49] (see Box 7 in Section 5.2).

2.3 Different sources of bias

Discrimination bias may primarily stem from the data. First, there are **historical biases**: the data may reflect past discriminatory practices (such as credit or insurance refusals, or reduced access to certain products), which models tend to replicate and normalise. Second, there are **representation biases**, linked to incomplete or unbalanced data: certain groups are under-represented, poorly observed or described by less relevant variables, which tends to undermine the quality of the predictions made in their regard. A third major source is **bias arising from the measurement and quality** of the data (errors, approximations, inappropriate aggregations), which generally affect certain populations disproportionately, particularly those whose incomes or life trajectories are not "standard". Lastly, bias may arise from the use of **proxy variables**, where seemingly neutral indicators (place of residence, marital status, etc.) are strongly correlated with protected characteristics and result in indirect discrimination.

[47] Research shows that what is considered an acceptable use of certain variables may change over time. Charpentier and Barry (2022) point out, for example, that at the end of the 19th century, insurers in the United States deemed the link between skin colour and life expectancy scientifically established, and applied it in practice in their models. More generally, modelling choices may contain an element of subjectivity, when they are rooted in a specific historical and social context (Glenn, 2000).
[48] Simon, 1988.
[49] Williams, Brooks and Shmargad, 2018.

Discrimination can also stem from the modelling itself – from the technical and conceptual choices that underpin processing. One source lies in the **choice of objectives and optimisation functions**: by seeking exclusively to maximise overall performance (profitability, accuracy, risk reduction), models may produce results that are unfair to certain groups. A second source concerns the choice of **parameters and decision rules**, such as thresholds, or trade-off or automatic rejection rules, which can have very different outcomes depending on the population, without being explicitly taken into account.

Third, there are **amplification and feedback effects** specific to algorithmic systems: the pursuit of statistical performance can itself contribute to amplifying biases present in the data through a range of mechanisms. Firstly, optimising the mean error across the entire population automatically leads the model to **favour majority groups or the most frequent cases**, at the risk of tolerating larger errors for minority groups. Secondly, machine learning algorithms tend to **prioritise the strongest correlations** in the data, which often reflect social structures or historical bias: these correlations can thus be reproduced in the model's decision-making and even reinforced. Non-linear models can also **amplify** initially modest differences by combining them with other variables to generate, in certain situations, **disproportionate effects**. Moreover, when the model's decisions influence future data (by conditioning access to credit or certain services, for example), **feedback loops** may exacerbate these biases over time. All these mechanisms conspire to create a model that could potentially **accentuate existing deviations** between groups.

Lastly, recent research suggests that bias amplification results not only from the properties of the final model, but also from the model's **training path**.[50] In particular, **the model tends to learn regularities (patterns) associated with majority groups during the early stages of optimisation, while the regularities specific to minority groups are learned at a later stage.** In these circumstances, a model whose training is terminated prematurely (early stopping) or whose capacity is limited may amplify bias because it lacked the time or resources required to integrate the structures specific to minority groups. Similarly, the choice of optimisation algorithm (gradient descent, Adam, etc.) can influence the speed of learning of minority regularities. These factors demonstrate that **from the perspective of fairness, training duration, model capacity, and the choice of model optimiser are not neutral parameters.** As such, they should be explicitly taken into account in model validation processes.

[50] Bachoc, Bolte, Boustany and Loubes, 2026.

Box 3: Bias decomposition methods

The approach often put forward in the scientific literature is to break down an observed disparity into **two components**: one attributable to data, and another attributable to the algorithmic processing.

The principle involves **comparing the deviation between groups in the model's outputs with the corresponding deviation for a reference variable**, such as the observed target (for example, the actual defect), an independent external data point, or a corrected target. If the deviation in the outputs is greater than the deviation in the reference variable, the model has **amplified** a pre-existing disparity; if it is comparable, the model has **transmitted** it without any significant change; if it is smaller, the model has **mitigated** it.

This framework is consistent with the reasoning behind the AI Act, which distinguishes between requirements relating to data quality (Article 10) and those concerning the design and operation of the models (Article 15).

The practical operational utility of this decomposition is that it steers corrective action. When a model significantly amplifies a disparity, this suggests that action should be taken on the model's training phases (choice of loss function, fairness constraints, optimisation) or on the model's outputs (post-processing). Conversely, when the model merely reproduces a pre-existing disparity, the levers for action are primarily upstream, at the level of data collection, selection, transformation or rebalancing.

In practice, the degree of amplification can be quantified using various **metrics** of distance or divergence between distributions (total variation distance, Wasserstein distance, Kullback–Leibler divergence, etc.). The choice of metric depends on the use case and the nature of the outputs (binary, ordinal, continuous). As with any statistical measure, associating these estimates with confidence intervals is recommended (see Section 4.1).

2.4 Group fairness and individual fairness

Research traditionally distinguishes between **group fairness**, which is based on a collective view of fairness, founded on a division of the population into distinct groups, and **individual fairness**, which is based on the legal rights of individuals.

Individual fairness is based on the idea that **similar profiles should receive similar treatment**. For example, when granting a loan, two applicants with an equivalent objective risk of default should be assigned similar credit scores, regardless of their personal characteristics. **Individual fairness is therefore above all a matter of procedural equality.** This conception of fairness theoretically allows for a **better performance** than group fairness, as the model should tend to better reflect each individual's risk profile. This paradigm, however, relies on **two major assumptions that are difficult to verify in practice**. First, it assumes that it is possible to define an objective measure of similarity between two individuals, which is not always guaranteed, and poses a particular problem when the measure of similarity involves sensitive attributes. Second, it is based on the assumption that the training data fairly represents the reality, which amounts to assuming that there is no discrimination bias.

Group fairness, on the other hand, is based on equality of results: it requires the model to perform similarly, from a quantitative perspective, across different subgroups of the population. For example, men and women *as a whole* should collectively benefit from the same credit

acceptance rate; in this case, the focus is not directly on individuals, but on social groups. **Group fairness is generally easier to verify using statistical tools** and helps to reduce **systemic inequalities** by rebalancing differences in treatment deemed unjustified. However, group fairness tends to reduce the model's overall performance, as imposing approval rate parity or error rate parity will generally lead to an increase in classification errors (see below).

Thus, **two legitimate but distinct conceptions of fairness** coexist. Individual fairness reflects a **more juristic approach**: every individual placed in a comparable situation must be treated equally throughout a procedure, and any difference in treatment must be objectively justified at individual case level. Conversely, group fairness reflects a **more statistical approach**, historically prevalent in finance, which reasons in terms of **risk groups** (particularly in insurance): decisions are based on regularities observed within populations, regardless of the specific situation of each individual. This **collective logic** lies at the heart of risk assessment and insurance pricing,[51] but sits in tension with the **legal principle that no one should be penalised due to their belonging to a certain group**.

The dichotomy between individual fairness and group fairness, however, deserves greater **nuance**. In principle, the requirements of **consistency** ("similar treatment for similar cases") and the **equality of odds** ("removing disadvantages not attributable to individuals' behaviour") are, at least to some extent, **compatible, and could in practice form the basis for both individual and group metrics.**[52] **Furthermore**, in financial practice, these two approaches are neither entirely incompatible nor interchangeable: group fairness helps to structure decisions that are consistent, predictable and justifiable at the portfolio level, while individual fairness reminds us of the need to recognise specific circumstances.

Regardless, due to the practical difficulties in implementing individual fairness metrics, **this paper focuses on measures of group fairness**.

2.5 Groups to be considered when analysing fairness

Comparing differences in treatment between two or more groups naturally raises the question of how those different groups should be composed. Here again, two broad conceptions coexist: groups can be defined according to **a single sensitive variable** (**univariate analysis**), or by **combining several sensitive variables** (**multivariate or intersectional analysis**).

Univariate analysis involves composing groups based on one variable (such as gender or age) at a time. It thus involves comparing the way in which a model treats two groups (men and women for the sensitive "gender" variable, for example) or more (such as age, broken down into groups by 10-year age brackets).

Univariate analysis has **undeniable advantages**. First, it is very **simple to implement**, as the number of groups to be studied is limited. It is also **more legible**, in that all parties can easily understand the composition of the groups being compared. However, research has revealed a **major limitation** of this approach: **it can hide major disparities within subgroups**. For example,

[51] Ewald, 2011.

[52] Some studies go even further and demonstrate a form of equivalence, for example, between demographic parity and counterfactual fairness (a measure of individual fairness). See Rosenblatt and Witter, 2023.

a model may seemingly treat groups of men and women fairly (overall), while significantly disadvantaging women under the age of 25.[53]

More broadly, research shows that **there is no satisfactory theoretical justification for restricting the analysis of differences in treatment to a single sensitive variable**. Discrimination is not necessarily produced by **addition**: it can arise from **combination**. In other words, biases can be amplified at the intersection of several sensitive attributes. **This is why there is robust consensus in the scientific literature recommending that differences in treatment between defined groups should be examined (multivariate or intersectional analysis).**

Dimensional cross-analysis nevertheless assumes that certain **practical difficulties** are overcome. First, cross-analysing several sensitive variables mechanically fragments the population into a **larger number of groups**, some of which may have **very small populations**, rendering the results unstable or statistically non-robust. Multivariate analysis generates a more fundamental conceptual tension: when subgroups are too finely defined, analysis shifts towards the individual level and thus **conflicts with the very logic of modelling**, which involves exploiting statistical regularities in order to differentiate treatment.

Box 4: Fairness in each institution and global fairness

It is important to note that **the requirement for fairness applied solely to the customers of a bank or insurance company does not necessarily guarantee fairness for the population as a whole**. This is primarily because **each institution's marketing and sales policies have a direct impact on the composition of its client base**: even without practising any form of discrimination, products designed for a specific type of client, more favourable rates for a particular category, or advertising targeted at certain segments of the population lead to the overselection of particular social groups (and the underselection of others). To take an extreme example, an insurer that decided that only men would be acceptable as customers could present a perfectly equitable male-female pricing structure, but without it having any practical application. Thus, the customers of a particular financial institution generally form a **statistically biased sample** of the total population.[54]

Furthermore, each financial institution only has **access to its own data**, by design. Even if it wanted its customer base to faithfully reflect the general population, it would not necessarily have the means to ascertain the distribution of all relevant characteristics. There are, however, instances where **certain customer data is pooled**, which could be used to promote global fairness. In the banking sector, for example, this type of data pooling is practised in most OECD countries through credit bureaux, which aggregate customer information (credit history, account transactions, etc.) from all banks in the market.[55]

[53] This phenomenon is sometimes qualified as fairness gerrymandering. See Kearns et al., 2018.
[54] Côté, Côté and Charpentier, 2024.
[55] Data pooling can also have positive effects for the entry of new players, who can access data to train their models.

3 Group fairness

3.1 The three main families of group fairness metrics

Numerous group fairness metrics can be found in the scientific literature, but they can be grouped into **three main families – independence, separation, and sufficiency** – each associated with a criterion defining a form of fairness.

In this section, each of these broad metrics is examined using a concrete example: **granting a loan**, considered here, for the sake of simplicity, as a **binary decision** between acceptance and refusal.[56] Here, the **target variable** is therefore the decision itself, and not the risk score, contrary to the most widespread practice.[57] To this end, we take a population of 200 loan applicants: 100 from **Group A** (favoured) and 100 from **Group B** (disadvantaged), with the following characteristics:

- Group A: 80 people repay their loan in full, and 20 default, giving a **baseline rate of** creditworthy **borrowers**[58] **of 80%**.
- Group B: 50 people repay their loan in full, and 50 default, **giving a baseline rate of** creditworthy **borrowers of 50%**.

3.1.1 Independence

- Main associated metric: demographic parity.
- Mechanism: this criterion requires that the model's decision be independent of the protected attribute (e.g. gender). In statistical terms, in a classification framework, **the probability of a loan application being approved must be the same for all groups**.[59] The algorithm does not consider whether applicants are actually capable of repaying their loan or not; it simply ensures that acceptance rates are equal.
- Functional logic: Acceptance rate for Group A = Acceptance rate for Group B.

[56] This is therefore a classification problem. Two main types of problems are usually distinguished: classification problems, which involve predicting a discrete value (where applicable, a binary value, such as 0 or 1 for the acceptance or refusal of a loan), and regression problems, which involve predicting a continuous value (for example, the interest rate on the loan granted).

[57] Granting a loan is most often based on the estimation of a risk score (regression problem). The decision on whether or not to grant the loan is then made by applying a score threshold: below the threshold, the loan is refused; above it, the loan is approved.

[58] Here, "creditworthy" is understood to be an intrinsic attribute of the individual, observed retrospectively. A creditworthy borrower is therefore, in this context, a borrower who actually repays their loan in the future, and not a borrower predicted to be creditworthy by the model.

[59] In a regression problem, demographic parity requires that the scores of Groups A and B have an equivalent distribution. This condition may be relaxed to require that the *mean* scores of each group be equal.

Implication: to satisfy this criterion, a bank could be forced to accept riskier profiles within a disadvantaged group simply to achieve statistical balance.

Scenario 1: Imposing **independence** in the loan approval model.

Objective: The bank wants the approval rates to be identical: borrowers in Group A must have the same approval rate as those in Group B.

Model action: The bank sets an approval rate of **60%** for each group.

- For Group A: The model selects the top 60. As there are 80 creditworthy borrowers, the model easily finds 60.
  → Result: 60 creditworthy borrowers approved.
- For Group B: The model must find 60 borrowers, but there are only 50 creditworthy borrowers in total. It manages to identify the 50 creditworthy borrowers but includes 10 non-viable borrowers to meet the quota of 60.
  → Result: 50 creditworthy borrowers and 10 non-viable borrowers approved.

3.1.2 Separation

- Main associated metric: error rate parity.[60]
- Mechanism: this criterion requires mathematical independence between the decision and the protected variable, conditional on reality. In other words, the model must have **equal acceptance rates across groups of individuals who exhibit the same actual behaviour** (the same value of the target variable). Statistically speaking, this amounts to constraining the model to check for **equal error rate parity**[61] between the two groups under consideration, i.e. equal false positive rates (proportion of non-viable borrowers accepted by the model), as well as equal false negative rates (equivalent, in a binary classification, to equal true positive rates;[62] as the latter is more intuitive, it is used hereafter).
- Functional logic:
  - Acceptance rate for creditworthy borrowers in Group A = Acceptance rate for creditworthy borrowers in Group B; **AND**
  - Acceptance rate for non-viable borrowers in Group A = Acceptance rate for non-viable borrowers in Group B.
- Note: The calculation of these rates is based on observing the actual repayment behaviour of all applicants. However, this can only be known for borrowers who were effectively granted a loan. Their estimation therefore requires the use of heuristics.[63]

[60] Also referred to as "equalised odds".
[61] This condition is sometimes relaxed, in order to examine the equality of true positive rates only; it is then generally referred to as an "equal opportunity" metric in the literature.
[62] In a binary decision framework, these two quantities are complementary, i.e. their sum is necessarily equal to 1.
[63] Several "reject inference" methods exist for estimating the rate of false negatives, with mixed rates of success. See, for example, Ehrhardt, Biernacki, Vandewalle, Heinrich and Beben, 2021.

- Implication: good profiles are not disadvantaged by their belonging to a given group; at the same time, risk-taking errors are not concentrated in any particular group.

Scenario 2: Imposing **separation** in the model.

Objective: The bank wants the error rates to be identical: creditworthy borrowers in Group A must have the same approval rate as those in Group B, with the same applying to non-viable borrowers.

Model action: The bank adjusts the model to ensure a true positive rate of 90% and a false positive rate of 10%.

- For Group A: The model approves 74 borrowers: 72 creditworthy and 2 non-viable.
  - True positive rate = 72/80 = 90%.
  - False positive rate = 2/20 = 10%.
- For Group B: The model approves 50 borrowers: 45 creditworthy and 5 non-viable.
  - True positive rate = 45/50 = 90%.
  - False positive rate = 5/50 = 10%.

3.1.3 Sufficiency

- Main associated metric: predictive value parity.[64]
- Mechanism: this criterion requires that **the reality** (in this case, a default) **be independent of group membership, conditional on the prediction made by the model**. Thus, if the actual default rate observed ex-post is 5% for approved borrowers in Group A, the default rate must also be 5% for approved borrowers in Group B. Statistically speaking, this amounts to requiring equality in positive predictive values (the proportion of positive predictions that are actually correct, i.e. the actual repayment rate of approved borrowers)[65] between Groups A and B, as well as equality in negative predictive values (the proportion of negative predictions that are actually correct, i.e. the actual default rate of rejected borrowers).[66]
- Functional logic:
  - Default rate for approved borrowers in Group A = Default rate for approved borrowers in Group B; **AND**
  - Default rate for rejected borrowers in Group A = Default rate for rejected borrowers in Group B.
- Note: The second condition usually necessitates the use of estimates, as the institution does not normally know the default rate of the borrowers it has rejected.

[64] Also called calibration.
[65] Or, equivalently, default rates.
[66] As with error rate parity, it is possible to relax this constraint to require only parity in positive or negative predictive values.

- Implication: In order to maintain a consistent level of reliability across the predictions of different groups, sufficiency may lead to the reproduction or even accentuation of existing disparities.

Scenario 3: Imposing **sufficiency** in the model.

Objective: The bank wants the predictive values to be identical: approved borrowers in Group A must have the same default rate as those in Group B, and the same applies to rejected borrowers.

Model action: The bank adjusts the model to ensure a positive predictive value of 95% and a negative predictive value of 80%.

- Group A: The model approves 80 borrowers (76 creditworthy and 4 non-viable) and rejects 20 (4 creditworthy and 16 non-viable).
    - Positive predictive value = 76/(76 + 4) = 95%.
    - Negative predictive value = 16/(4+16) = 80%.
- Group B: The model approves 40 borrowers (38 creditworthy and 2 non-viable) and rejects 60 (12 creditworthy and 48 non-viable).
    - Positive predictive value = 38/(38 + 2) = 95%.
    - Negative predictive value = 48/(12+48) = 80%.

3.2 Impossibility theorem

The scientific literature shows that **it is mathematically impossible to construct a model that simultaneously satisfies two of the three group fairness criteria** outlined above (independence, separation, sufficiency) whenever the target variable is correlated with a sensitive variable (in other words, whenever baseline rates differ between groups).[67] A fortiori, it is therefore also impossible to satisfy all three criteria in this type of situations.

[67] Barocas, Hardt and Narayanan, 2019.

The following tables illustrate this point using the examples discussed above:

Taking **scenario 1**:

1. Independence (respected):
    - Group A: Approval rate = 60/100 = **60%**
    - Group B: Approval rate = 60/100 = **60%**

2. Separation (**violated**):
    - Group A: True positive rate = 60/80 = **75%**, False positive rate = 0/20 = **0%**
    - Group B: True positive rate = 40/50 = **80%**, False positive rate = 10/50 = **20%**
    - Conclusion: A creditworthy borrower in Group B will be approved more often than a creditworthy borrower in Group A, but a non-viable borrower in Group B will be approved by mistake more often than a non-viable borrower in Group A.

3. Sufficiency (**violated**):
    - Group A: Positive predictive value = 60/60 = **100%**
      Negative predictive value = 20/40 = **50%**
    - Group B: Positive predictive value = 50/60 ≈ **83%**
      Negative predictive value = 40/40 = **100%**
    - Conclusion: An approved borrower in Group B will default more often than an approved borrower in Group A, and a rejected borrower in Group A would have been more likely to have repaid the loan than a rejected borrower in Group B.

Taking **scenario 2**:

1. Independence (**violated**)
    - Group A: Approval rate = 74/100 = **74%**
    - Group B: Approval rate = 50/100 = **50%**
    - Conclusion: Group A obtains far more loans.

2. Independence (respected):
    - Group A: True positive rate = 72/80 = **90%**, False positive rate = 2/20 = **10%**
    - Group B: True positive rate = 45/50 = **90%**, False positive rate = 5/50 = **10%**

3. Sufficiency (**violated**):
    - Group A: Positive predictive value = 72/74 ≈ **97%**
      Negative predictive value = 18/26 ≈ **69%**
    - Group B: Positive predictive value = 45/50 = **90%**
      Negative predictive value = 45/50 = **90%**
    - Conclusion: An approved borrower in Group B will default more often than an approved borrower in Group A, and a rejected borrower in Group A would have been more likely to have repaid the loan than a rejected borrower in Group B.

Taking **scenario 3**:

1. Independence (**violated**):
    - Group A: Approval rate = 80/100 = **80%**
    - Group B: Approval rate = 40/100 = **40%**
    - Conclusion: Group A obtains far more loans.

2. Separation (**violated**):
    - Group A: True positive rate = 76/80 = **95%**, False positive rate = 4/20 = **20%**
    - Group B: True positive rate = 38/50 = **76%**, False positive rate = 2/50 = **4%**
    - Conclusion: A creditworthy borrower in Group A will be approved more often than a creditworthy borrower in Group B, but a non-viable borrower in Group A will be approved by mistake more often than a non-viable borrower in Group B.

3. Sufficiency (respected):
    - Group A: Positive predictive value = 76/80 = **95%**
      Negative predictive value = 16/20 = **80%**
    - Group B: Positive predictive value = 38/40 = **95%**
      Negative predictive value = 48/60 = **80%**

3.3 A comparison of the three families of metrics: underlying assumptions and practical implications

The previous sections have illustrated the differences between independence, separation and sufficiency through a credit granting example. This section offers **a more in-depth comparison** of these three criteria, highlighting their advantages, their limitations and the normative implications they entail.

3.3.1 Independence

Independence (demographic parity metric) is the simplest of the three criteria because it **only takes into account the model's predictions**, which is both an advantage and a drawback. It is the **easiest to implement** from an algorithmic perspective, as it requires neither observation of the target variable (repayment or default, in the lending example in Section 3.1) nor modelling of model errors, but simply a direct comparison of decisions between groups.

However, this is not always enough to guarantee "fairness". Returning to the example of granting a loan, demographic parity between Groups A and B can easily be achieved by simply approving half of the borrowers from each group on a **purely random** basis, i.e. without taking their "financial viability" into account at all. It can therefore lead to a situation where non-viable borrowers are approved **at the expense of** creditworthy borrowers, which is not a very "fair" situation (see also Box 5 below, on the distributional effects of each fairness family).

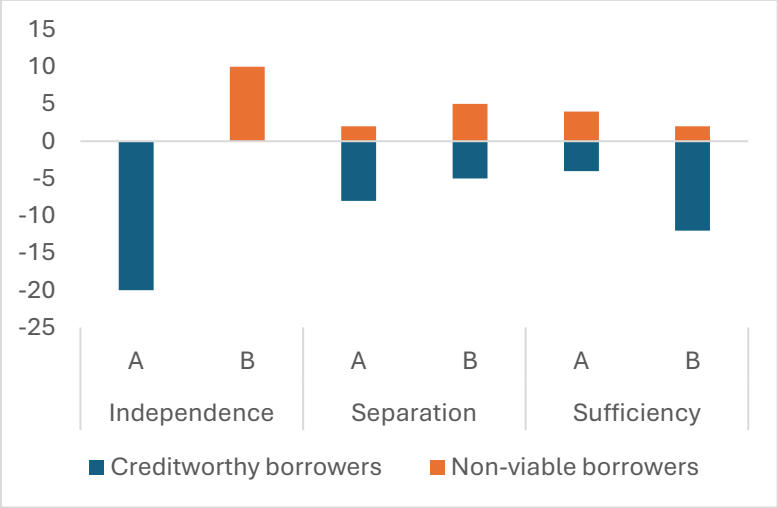
Box 5: Summary of the distributional effects of each fairness family in the described example

The chart below summarises the figures given in the example presented in the previous sections, illustrating the distributional effects associated with each fairness scenario. It compares the result obtained when the model is successively constrained by the criteria of independence, separation and sufficiency with the credit distribution considered optimal under the initial assumptions: the applications of the 80 creditworthy borrowers in Group A and the 50 creditworthy borrowers in Group B would be accepted, while all other applications would be rejected.[68]

The chart shows the main distributional trends for each fairness family (analysed in greater detail in this section). It is important to stress, however, that this is for illustrative purposes only: this exercise is based on deliberately simplified and stylised assumptions, and its implications should be interpreted with caution, particularly as financial sector models can have numerous use cases that cannot be reduced to a binary classification between a favoured group and a disadvantaged group.

Chart: Comparison between the result of each fairness scenario (independence, separation and sufficiency) and an optimal credit distribution

(In number of applications accepted or rejected)



The chart shows that each fairness family results in different distributional outcomes between the two groups, A (favoured) and B (disadvantaged). With **independence**, the constraint of global parity in decisions leads to a significant correction in distribution in favour of Group B: the number of non-viable borrowers accepted increases significantly (+10), while a number of creditworthy borrowers from Group A are rejected (-20), reflecting a significant loss of effectiveness. **Separation** seemingly strikes a balance: it reduces the overall differences between Groups A and B, at the cost of a certain under-selection of creditworthy borrowers in Group A (-8) and classification errors in Group B. Lastly, **sufficiency** performs quite well in the selection of creditworthy borrowers from Group A, but severely penalises creditworthy borrowers from Group B (-12).

[68] The number of applications accepted is broadly comparable across the three scenarios: 120 applications accepted for independence and sufficiency; 124 for separation due to mathematical constraints.

More generally, the demographic parity metric **does not necessarily take into account the risks** associated with individuals or groups. It therefore tends to **overlook differences in behaviour** that could explain the differing results between groups, interpreting them as the result of a historical inequality of odds – in other words, as artefacts of history rather than as intrinsic differences between the groups under consideration.[69]

3.3.2 Separation

The criterion of separation (error rate parity metric) aims to correct the main limitation of the independence criterion (demographic parity metric) by **explicitly taking into account real-world outcomes**, that is, the **actual behaviour** of individuals. In other words, separation is a definition of fairness that introduces the **concept of risk** (default on a loan, behaviour presenting risks in the case of non-life insurance, etc.). It therefore overcomes the main criticism levelled at demographic parity. Separation can be viewed as a **form of compromise** between independence and sufficiency, in that it seeks to take into account the influence on the actual situation of historical and social factors on the one hand, and to tailor the model's predictions to this situation, on the other.

Furthermore, fairness as envisaged under the principle of separation takes into account the fact that **different social groups may suffer unequal harm as a result of the use of automated decision-making**. In particular, models often produce higher error rates for historically marginalised and disadvantaged groups, thereby inflicting further harm upon them.[70]

However, fairness as envisaged by the separation criterion has **several drawbacks**. First, this form of fairness tends to mechanically **reduce the model's performance**. When risk distributions differ between groups, an unconstrained optimised model will naturally adjust its decision thresholds to minimise overall errors. Imposing similar error rates across groups amounts to deviating from this optimal trade-off: it is then often necessary to deliberately increase certain types of errors in one group to align them with those observed in another (see, for example, Box 5). The objective of fairness introduces an additional constraint that restricts the space of possible solutions and generally prevents the simultaneous achievement of maximum predictive performance.

Second, the separation criterion leads to reasoning based exclusively on the observed target variable (for example, borrower default), which is assumed to be a reliable benchmark. However, in many financial use cases, **this target variable is itself prone to imperfections**: selection bias

[69] In this approach, particular importance is therefore generally accorded to groups that may have experienced historical discrimination and/or exhibit systemic differences in key indicators of living standards (e.g. poverty rates) compared with the rest of the population.

[70] The example of the COMPAS algorithm, used by some courts in the United States to assess an offender's likelihood of recidivism (by classifying them as high risk or low risk), illustrates this issue well: the organisation ProPublica found that the algorithm was biased against blacks, showing that black defendants had a false positive rate nearly twice that of white defendants (45% compared with 23%), and that, conversely, the false negative rate was significantly higher among white defendants than black defendants (48% compared with 28%). In other words, COMPAS was twice as likely to wrongly classify a black defendant as high risk than a white defendant, and twice as likely to wrongly classify a white defendant as low risk than a black defendant. Conversely, those supporting the algorithm highlighted comparable levels of reliability in positive predictions between white and black defendants (in other words, among those classified as high-risk, the proportion of recidivism is similar across the two groups, which corresponds to a metric of sufficiency: predictive value parity).

(only certain profiles have been subject to a decision and therefore observed), dependence on past decisions (credit history constructed under constraints), or measurement noise. By explicitly conditioning fairness constraints on this target, the metric can contribute to **perpetuating these imperfections**.

Third, in configurations where risk distributions differ between groups, the practical implementation of the separation metric generally leads to the use of decision thresholds[71] that vary depending on the group. Consequently, two individuals with comparable levels of risk may be treated differently simply because of their belonging to a different group.

Fourth, **the ethical and regulatory implications of the different errors may be asymmetrical**. In the example of credit granting, a false negative corresponds to the refusal of a loan to a creditworthy borrower, with a risk of being deprived of opportunity, while a false positive corresponds to the granting of a loan to a non-viable borrower, with a risk of default or over-indebtedness. These two types of error are linked by the decision threshold applied to the risk score: therefore, making the model stricter reduces false positives – and therefore the risk of default or over-indebtedness – but mechanically increases false negatives, i.e. the unjustified refusal of a loan to creditworthy borrowers. Consequently, the separation constraint amounts to striking a balance between these conflicting objectives, which gives rise to an **ethical reflection**: in the case of a loan, is it more acceptable to refuse loans more often to creditworthy borrowers in Group B than to those in Group A, or to grant loans more often to non-viable borrowers in Group B than to those in Group A? It all depends on the underlying assessment of the benefits and risks involved.

3.3.3 Sufficiency

Sufficiency metrics are based on the idea that the decisions taken by a model must have the **same signification regardless of the group to which an individual belongs**. In the example above, this means that when the model decides to grant a loan, the probability that the borrower will repay it must be the same for all groups: a positive decision must not be more reliable for one group than for another.[72] From a probabilistic perspective, this translates into **equality of predictive values** (both positive and negative) between groups: among the individuals accepted, the proportion of creditworthy borrowers must be comparable, and among those rejected, the proportion of non-viable borrowers must also be comparable. This requirement is closely linked to the concept of **calibration**: a given score (for example, an 80% probability of repayment) must correspond to the same empirical reality, regardless of the group. Sufficiency thus guarantees a form of **fairness in the interpretation of decisions**.

However, when groups, on average, differ – particularly due to economic or social factors – this requirement has significant consequences. If a group has a lower baseline rate (in our example, a lower proportion of creditworthy borrowers), then maintaining the same precision mechanically imposes the need to be **more selective** for that group. In other words, the **decision threshold**

[71] In fact, the separation criterion is generally defined in terms of final decisions (in this case: acceptance or refusal), rather than the risk scores themselves. It is possible to define variants of the criterion directly on the scores (i.e. on a regression problem rather than a classification problem), for example by requiring that their distribution be independent of the group and conditional on the actual risk. These conditions are, however, more demanding and rarely verified in practice.

[72] Although sufficiency can be assessed on binary decisions, it is primarily a property of the risk scores themselves, in that it requires that, for a given score level, the probability of the event of interest be identical between groups.

has to be raised to accept only the least risky cases, because **in a group where sound profiles are rarer, applying the same level of selectivity as in the other group would mechanically lead to the risk of accepting more non-viable borrowers**.

This results in a **potentially significant reduction** in the number of positive decisions for the group (as shown, for example, in Box 5). Sufficiency therefore guarantees **ex-post fairness**, in the sense that the decisions taken all have the same "predictive quality", but it does not guarantee **ex-ante fairness of access**: individuals belonging to a disadvantaged group may in reality have a much lower chance of obtaining a favourable decision, even with comparable characteristics. In fact, sufficiency metrics **do not directly limit the number of errors made regarding genuinely** creditworthy **individuals**: it is therefore possible, while adhering to these metrics, to reject a higher proportion of creditworthy applicants in one group than in another, provided that the decisions made remain broadly reliable.

Furthermore, these structural effects can be amplified by the characteristics of the data and the model. In practice, disadvantaged groups are often **less well represented or more heterogeneous**, which can impair the quality of the scores: as uncertainty is higher, the model is less able to distinguish between creditworthy and non-viable profiles. To maintain a consistent level of precision, the model then adopts a **more cautious** approach, **further tightening selectivity** for more disadvantaged groups.

More fundamentally, sufficiency is based on the baseline rates observed in the data, which themselves reflect economic, social or historical realities. By aligning decisions with these baseline rates, it therefore tends to **reproduce existing differences** between groups, and even **reinforce** them where they are already pronounced. In other words, the inequalities present in the data are not corrected, but incorporated into the decision-making logic. Thus, while sufficiency offers a strong guarantee in terms of probabilistic consistency and the reliability of decisions, it can also lead to the increased and lasting **exclusion** of certain groups, particularly when these groups present a higher average level of risk.

3.3.4 Summary table

The points discussed in this section are summarised in the table below.

Fairness criterion	Independence (demographic parity)	Separation (error rate parity)	Sufficiency (predictive value parity)
Measure	Same approval rate between groups.	Same false positive and false negative rates between groups.	Same positive and negative predictive values between groups.
Principle	“The proportion of approved borrowers is the same for all groups.”	“Among creditworthy borrowers, the proportion of approved borrowers is the same for all groups.”	“Among approved borrowers, the proportion of creditworthy borrowers is the same for all groups.”
Normative implications	- Assumes that current inequalities are due to historical artefacts rather than intrinsic differences.	- Assumes that classification errors affect groups unequally partly because of historical artefacts.	- Assumes that current inequalities are due to intrinsic differences rather than historical artefacts.
Advantages	- Easy to implement, particularly as it requires only the model’s predictions - Promotes access to credit for borrowers from disadvantaged groups	- Takes into account the creditworthiness of borrowers - Promotes access to credit for creditworthy borrowers from disadvantaged groups	- Takes into account the creditworthiness of borrowers - Consistency of decisions: a prediction has the same signification for all groups (same score = same level of risk)
Drawbacks	- Fails to take into account the creditworthiness of borrowers - Leads to accepting a different treatment for similar individuals (in terms of risk) belonging to different groups	- Tends to mechanically reduce model performance - Leads to accepting a different treatment for similar individuals belonging to different groups - A same decision may correspond to different risk levels depending on the group	- Increased risk of exclusion: stricter criteria for disadvantaged groups - No uniform protection for low-risk individuals: risk of rejecting more creditworthy applicants in certain groups - Reflects and may exacerbate existing inequalities between groups

Table 2: Comparison of group fairness criteria

4 Assessing and correcting bias

4.1 Assessing bias in practice: taking uncertainty into account

As with any statistical estimate, **measuring bias is prone to uncertainty**. Relying on a point estimate without assessing its **precision** can therefore lead to mistakenly identifying a random deviation or, conversely, to failing to detect a real bias. Improperly taking this uncertainty into account can undermine the credibility of assessments and lead to the unnecessary commitment of resources to rectify the situation.

It is therefore important to take into account **four main sources** of uncertainty.[73]

- **Sampling variability**: fairness metrics are calculated on finite datasets, and **their precision depends directly on sample size**. A deviation may be highly visible in large datasets, yet indistinguishable from pure chance in small datasets. **Fairness analyses should therefore be systematically associated with confidence intervals**, obtained using analytical methods (normal approximation, Fisher's exact test, etc.) or resampling approaches (bootstrapping). The method chosen should be **specified and documented**.
- **Variability related to training**: results may depend on **random factors specific to the learning process** (parameter initialisation, order of data presentation, regularisation mechanisms, etc.). Thus, two training runs of the same model, using the same data, may result in a measurement of two different levels of bias. A bias measurement derived from a single training run is merely one of many other possible outcomes. For the most critical systems, it is therefore advisable to **repeat the training runs using different random seeds and to take into account the distribution of the measurements obtained**.
- **Multiplicity of tests**: fairness analyses often rely on numerous comparisons (multiple groups, metrics, etc.). However, this multiplicity **mechanically increases the probability of detecting at least one significant deviation by pure chance**.[74] Standard adjustment procedures (Bonferroni, Holm, Benjamini-Hochberg) help to manage this risk and maintain the reliability of the conclusions.
- **Statistical power in small groups**: analyses involving small groups are characterised by **greater uncertainty**. In these situations, the absence of a statistically significant deviation is not proof of fairness, but may simply reflect a lack of test power. It is therefore advisable to enhance the results with an **estimate of the minimum detectable amplitude**, i.e. the smallest difference that the analysis can identify given the available data.

[73] Besse, del Barrio, Gordaliza, Loubes and Risser, 2022.

[74] To illustrate, evaluating 20 subgroups without correction results in a 64% chance of finding at least one significant deviation at the standard 5% threshold, even though the model is perfectly fair (Cook, Gebski, & Keech, 2004).

4.2 Bias correction methods

Several methods exist for correcting discrimination biases observed in machine learning models. A common way of classifying them is to divide them into three categories:[75]

1. **Pre-processing** methods are designed to correct underlying discrimination upstream by analysing or transforming the training data;
2. **In-processing** methods are designed to reduce discrimination during the model training process, by modifying the objective function or imposing constraints;
3. **Post-processing** methods are designed to correct discrimination after training, by adjusting the model's output scores.

4.2.1 Pre-processing methods

Pre-processing methods consist of a set of approaches designed to act **prior** to the training phase, by **modifying the data or their representation** in order to reduce, or even remove, dependencies between sensitive variables (such as gender or age) and the other characteristics used by the model. The central idea is to construct a **"cleansed" data space**, in which potential biases have been corrected even before the model is trained. This approach has the advantage of being largely **independent of the algorithms used**: once the data has been adequately transformed, any model trained on these data is expected to inherit, at least in part, the sought-after fairness properties. In this sense, pre-processing is a **cross-cutting and often modular approach**, which can be integrated into various processing chains. However, its effectiveness depends directly on the quality of the transformations carried out, as well as on the ability to preserve information useful for prediction.

In practice, there are several broad categories of methods. "**Blinding**" approaches aim to neutralise the influence of sensitive variables by structuring the data according to predefined subgroups and fairness criteria. **Causal methods**, on the other hand, seek to identify the mechanisms responsible for the discrimination by explicitly modelling the relationships between variables, in order to correct biases at their source, despite the difficulties involved in implementation. Other techniques rely on **dividing** the dataset into subgroups and on **sampling** strategies, in order to better represent disadvantaged populations and assess disparities. **Transformation methods** are designed to construct new representations of the data that are less correlated with sensitive attributes but still maintain their predictive power. Lastly, more operational approaches involve directly modifying the data (**re-labelling, perturbation**) or their relative weight in the learning process (**rebalancing**), in order to correct observed imbalances.

4.2.2 In-processing methods

In-processing methods involve **acting directly during model training** by explicitly incorporating fairness objectives into the **optimisation process**. In contrast to pre-processing approaches, which modify the data upstream, these methods **adjust the behaviour of the model itself** so that it complies with certain fairness constraints while maximising its predictive performance. In principle, they therefore strike a **better balance between fairness and accuracy**, as both objectives are optimised jointly. However, they present **significant operational constraints**: they require **full access** to the data and learning algorithms, and are often **specific to certain**

[75] Caton and Haas, 2024.

types of models or problems, which limits their portability and generalisability in heterogeneous environments.

There are several main approaches. **Regularisation and constrained optimisation** methods involve modifying the model's **objective function** by incorporating **penalties** related to fairness gaps, so as to steer the learning process towards less discriminatory solutions. **Adversarial learning** methods introduce an "adversary" mechanism tasked with detecting information relating to sensitive variables in the model's predictions or representations, and encouraging the model to remove them. Other approaches, such as **algorithmic bandits**, fall within a framework of sequential and adaptive learning: they are designed to make fair decisions as they go along, by balancing exploration and exploitation, and by incorporating fairness as a dynamic performance criterion.

4.2.3 Post-processing methods

Post-processing methods are applied **once the model has been trained**, by **adjusting its predictions** to meet a given fairness criterion. Unlike the approaches applied during or prior to training, they require neither data modification nor model retraining, making them particularly useful in contexts where the model is **already deployed** or operates as a "**black box**".[76] This flexibility is their main advantage: they can be applied to any type of model, without any additional training costs. However, the fact that they generally rely on a more overt use of group belonging (for example, by setting different acceptance thresholds) may raise legal or ethical questions.

Among the main approaches, **calibration methods** aim to adjust the scores produced by the model – by aligning predicted probabilities with observed frequencies – so that they can be interpreted consistently between groups. They are particularly relevant when the model's outputs are used as an aid in decision-making rather than for automated decision-making. **Thresholding methods**, on the other hand, involve setting – potentially differently depending on the groups – decision thresholds that reconcile certain measures of fairness and performance. They particularly target ambiguous situations, close to decision thresholds, where the risks of bias are highest.

4.2.4 Summary table

The points discussed above are summarised in the table below.

[76] Particularly when the model is purchased "off the shelf" from a third-party provider.

Table 3: Comparison of different bias correction methods

Approach	Advantages	Drawbacks
Pre-processing	- Transforms the variables space so that it is **independent of the sensitive attribute** prior to model training, making it largely reusable in various downstream applications.	- This approach **does not directly optimise the estimator** to reconcile both fairness and predictive performance during training.
In-processing	- May offer the **best performance**, as the model is optimised by directly incorporating the fairness constraint into the learning process.	- Requires **access** to raw data as well as the training procedure. - This approach is **less general**, as it can often only be applied to certain classes of models or specific optimisation schemes.
Post-processing	- **Works with any model, even "black-boxes"**. - Does not require model **retraining**, which is useful when the original training is complex or unavailable.	- This approach often relies explicitly on **group membership**, for example by setting different thresholds for different groups, which can present a challenge.

5 Practical implementation

This section builds on the points discussed above, as well as the workshops conducted by the ACPR with a number of volunteer institutions (see Box 6 below), to examine **practical arrangements for implementing the** principles relating to fairness.

Box 6: Fairness workshops carried out by the ACPR with volunteer financial sector participants

In the spring and autumn of 2025, the ACPR held a series of workshops with volunteer financial sector participants to gain insights into issues of fairness. Their aim was to understand how the banks and insurers interviewed dealt with these types of issues in their processes from a **technical** perspective and incorporated them into their **governance** on a concrete basis.

The workshops helped to provide an overview of the participants' views and practices with regard to the various fairness-related issues (see below). They notably showed that in the financial institutions interviewed – which were likely among the most advanced on these issues at the time of the interviews – the work undertaken on fairness issues within the organisations had taken a variety of forms (pilot schemes, raising awareness across different levels of defence, internal guidelines, etc.). In all cases, the work undertaken was **relatively recent and limited**, as fairness issues were deemed complex. From this perspective, the participants interviewed stated that they had high expectations of financial supervisors – and even of national and European legislators – to clarify the applicable rules.

These workshops also highlighted the fact that most financial players collect **relatively little protected or sensitive data**, limiting themselves to gender, age and place of residence in most financial use cases, as well as health-related data for certain insurance policies.

5.1 General considerations

5.1.1 Fairness and governance in the financial sector

It is clear from the points discussed above that **algorithmic fairness cannot be regarded as a purely technical issue**. Consequently, and contrary to what we often see in practice, it cannot be solely entrusted to the discretion of data scientists. **On the contrary, it involves decisions that fall within the remit of strategy, risk management and, more broadly, financial institutions' accountability**. As such, it must be treated as a **governance issue** and involve all levels of decision-making within the organisation.

The **highest level of governance** is therefore responsible for defining the **broad guidelines** on fairness, just as it does for an organisation's risk appetite, for example. The board of directors could therefore demonstrate its commitment by approving a **written policy** on AI fairness and by requesting regular reports on fairness-related key performance indicators (KPIs). The broad principles adopted by senior management can then be **rolled out** across the **various business lines**, and adapted to the different use cases – such as consumer credit, mortgage lending or fraud prevention – with their significantly different challenges, available data or risks. Responsibility for **translating these guidelines into concrete practices** would then lie at the **technical level**, employing suitable methods (selection of fairness metrics, bias correction techniques, validation procedures, etc.) and, whatever the situation, **state-of-the-art** approaches. Furthermore, fairness considerations should be incorporated into the **three lines of**

defence traditional to the financial sector: (i) model developers, (ii) independent model validation and compliance, and (iii) internal audit.

It is important to note that governance with regard to fairness – like the governance of AI systems more generally – does not necessarily require the creation of new structures, but **can be perfectly well integrated into existing model risk management frameworks**. The challenge is rather to integrate fairness issues into these frameworks explicitly, alongside the traditional concerns of performance or robustness, in order to ensure a **consistent and systematic approach**.

In this context, financial sector institutions could find it useful to implement a **structured review process** for each AI use case centred on **some key questions**:

- What biases do we wish to prevent or correct? Answering this question first requires clarification of the benchmark used to determine potential discrimination. This benchmark is primarily legal: financial institutions must comply with the non-discrimination requirements set out in European and national law. Beyond these requirements, they may choose to adopt more stringent standards that reflect their ethical commitments or strategic priorities. These reviews can thereby act as an opportunity for corporate governance to set out its ambitions with regard to fairness and to clarify its choices by answering a number of questions. Does it already have a clear understanding of the disparities that its marketing and sales policies or its history may have created in the composition of its customer base, access to its services, or pricing? Does it consider these disparities to be legitimate, for example in light of its competitive positioning or its statutory duties, or does it wish to reduce them? Does it limit itself to avoiding the introduction of further biases relative to the observed data, or does it wish to actively correct existing inequalities?
- What technical or organisational measures are in place to prevent or correct bias?
- Do these measures introduce new risks, for example, by impairing the model's performance, increasing its opacity or generating other forms of bias?
- Lastly, how is the trade-off between these different, often competing, effects managed?

5.1.2 Taking fairness into account throughout the system's lifecycle

In order to take the issues associated with algorithmic fairness into account, fairness considerations should be incorporated – from the **development stage** onwards – among the model's explicit objectives, in the same way as performance or robustness. This requires the **prior identification of any sensitive data** used (see Section 5.2) as well as the **groups** likely to be affected differently by the model (see Section 5.3 on group selection). Generally speaking, it is important to closely examine the **datasets**, conducting analyses of representativity and quality by group, and documenting imbalances, historical biases (see Section 2.3) and known limitations. Furthermore, fostering dialogue between the data science teams and business lines, as well as with legal and compliance experts, can help ensure that fairness issues are better taken into account.

At the model validation stage, results should be systematically assessed on a group-by-group basis, using fairness metrics. To do so, the **metric best suited** to the use case must be defined (see Section 5.4), and the trade-off between reducing disparities and other system objectives (such as performance, see Box 8) explicitly stated. It is good practice to **document and justify** the results of these analyses (see in particular Section 5.5 on the selection of relevant

thresholds), which may, where appropriate, lead to **technical adjustments** (see Section 5.6) or **governance decisions** on the acceptability of residual risk.

At the system deployment stage, taking fairness issues into account may require the implementation of organisational safeguards. In some cases, the institution will have to ensure that system users understand its limitations, particularly with regard to differentiated performances depending on the groups. Mechanisms for human oversight or to contest decisions or escalate concerns can then be put in place, particularly for cases with a significant individual impact. The **system's operating parameters** (decision thresholds, full or partial automation, integration with existing processes, human intervention) can have **as great an impact on fairness as the algorithm itself,** which justifies designing and documenting them to the same exacting standard.

Lastly, **monitoring the system over time** is important for identifying potential deviations. Data, the populations concerned and uses evolve, which can create new disparities or exacerbate existing biases. In this case, it is worthwhile for institutions to **implement continuous monitoring of fairness indicators by group**, alongside **periodic reviews and warning mechanisms**. Furthermore, as with other aspects such as performance or explainability, user feedback and internal or external audits can contribute to a process of continuous improvement.[77]

Irrespective of the case, institutions should take fairness issues into account on the basis of a **proportionate, risk-based approach**: thus, the standards required, the depth of the analyses conducted, and the governance arrangements should be **tailored** to the model's potential impacts on individuals.

5.2 Use of protected characteristics and sensitive variables

The personal characteristics of existing or potential customers collected by financial players have **varying statuses**, both from a legal perspective and in terms of their potential effects (see Section 1.1.3). The following discussion provides an illustration, without claiming to exhaustively cover all the situations that are likely to arise.

A piece of information commonly requested by financial players is the **age** of their customers. Under non-discrimination law, this is a protected characteristic but the assessment of its use is relatively **flexible**. Indeed, in many practical use cases, age is directly and objectively correlated with **relevant risk assessment factors**, such as the repayment horizon, the stability and trajectory of income, and also the investment horizon and the ability to absorb financial risks over the long term. These factors generally **justify differences in treatment** based on age. Furthermore, under the GDPR, age falls within the category of "ordinary" personal data, which can be processed provided that the general principles of data protection are respected.

The situation is significantly more restrictive for **gender**. Although it is not classified as sensitive data under the GDPR, it constitutes a protected characteristic in terms of non-discrimination. Direct differential treatment based on gender is **prohibited** in principle, particularly in insurance following the introduction of gender-neutral pricing requirements in Europe (see Box 2 in Section

[77] Toolkits are available to automate some of these aspects. In addition to the Veritas tool (MAS, Singapore) mentioned above, other examples include IBM's AI Fairness 360 and Aequitas, which offer metrics and methods for identifying and correcting bias.

1.3.2). Furthermore, gender is one of the characteristics particularly exposed to risks of **indirect discrimination**, insofar as it can be reconstituted from numerous proxy variables in the models.

Financial players also generally collect information on customers' **place of residence**.[78] This is not a protected characteristic in itself, nor is it classified as sensitive data under the GDPR. However, it can act as a **proxy for protected characteristics** (ethnic origin, socio-economic status, etc.), thereby posing a risk of indirect discrimination. In this case, its use must be **specifically justified** to demonstrate that it is based on legitimate and proportionate grounds.

Lastly, the insurance sector sometimes collects **data on health** (particularly for loan insurance and death and disability insurance). These **data are sensitive** under the GDPR, and, in principle, cannot be processed unless a specific exemption applies (notably, the explicit consent of the person concerned). Furthermore, a person's state of health is a protected characteristic under French law (see Section 1.1.3).[79] However, in the insurance sector, these data have historically been central to risk assessment and pricing (particularly in loan insurance and death and disability insurance). Its use is permissible, but it must therefore **reconcile** legitimate **actuarial requirements** with the **heightened protection requirements** for individuals' rights.

Box 7: Must we collect more sensitive data to detect bias?

Practically speaking, detecting bias means that we have to be able to measure it: by construction, **we cannot detect what we cannot observe**. From this point of view, collecting certain data relating to protected characteristics would appear to be useful in assessing the group fairness of models. This approach has traditionally met with strong resistance, particularly in France, where the use of such data has historically been subject to strict legal regulation.

Recent developments have, however, introduced **some targeted leeway**. In particular, **the AI Act permits, under strict conditions, the processing of sensitive data where this is necessary for bias detection and correction in AI systems**, particularly high-risk AI systems (Article 10(5)).[80]

Collecting more sensitive data therefore seems plausible, provided that the objective is indeed to combat discrimination (and that it cannot be achieved by other means) and that certain **guarantees** are respected, particularly with regard to security, access restrictions, non-reuse and the prohibition of sharing with third parties. Lastly, it must be duly justified, documented and time-limited, with the data deleted as soon as it is no longer needed.

[78] Sometimes only the postcode.

[79] It is important to note that these data are also protected by specific mechanisms, such as rules relating to the right to be forgotten or sector-specific agreements (see Section 1.1.3).

[80] This point is currently under discussion as part of the European Commission's "Digital Omnibus" package, which notably aims to make adjustments to the AI Act.

5.3 Bias identification: statistical uncertainty and univariate or multivariate analysis

Firstly, when **taking statistical uncertainty into account**, **any measure of bias should go hand-in-hand with a measure of its precision**[81] (see Section 4.1). As such, **three elements** should be included as standard in the fairness analyses carried out by financial sector institutions:

- a **confidence interval** for each reported metric, with a specific mention of the estimation method;
- where the model is trained in-house, an **indication of the variability of the measures** obtained across multiple training runs (distinct random seeds);
- for each group in which no significant deviation is detected, an estimate of the **minimum detectable deviation amplitude** given the available sample size.

As for the **groups to be compared**, the ACPR's workshops showed that, among financial players, analyses focused **almost exclusively on groups defined by a single sensitive variable** (such as gender). However, there is some agreement in the scientific literature on favouring the idea of comparing groups defined by the intersection of several variables (multivariate analysis), where the data allow (see Section 2.5).

In the financial sector, the **practical difficulties of multivariate analysis for most use cases do not appear insurmountable**. Indeed, the number of sensitive variables collected is generally limited (see Box 6). Therefore, in most cases, it is a matter of **intersecting two or three dimensions**, which limits the risk of combinatorial explosion and reduces the complexity of the analysis.

Given that, implementing multivariate analyses requires, above all, a **definition of the minimum size** of the groups included in the analysis, in order to ensure the statistical robustness of the comparisons. In small groups, the variance of the estimator for fairness metrics increases mechanically, and the absence of a statistically significant deviation may simply reflect a **lack of test power** in the comparison (see Section 4.1).

The scientific literature therefore recommends **determining the minimum group size on a case-by-case basis**, depending on the characteristics of the population being studied, rather than applying uniform rules.[82] Where certain groups are too small, it may be necessary to **merge them or to adjust the way in which continuous variables are "divided" into categories** (for example, in the case of age, by modifying the initially defined age brackets).[83]

---

[81] Moreover, this principle is consistent with the general requirements for statistical robustness applicable to internal models in the financial sector.

[82] A minimum sample size of 30 individuals is sometimes cited as an empirical rule (linked to the central limit theorem, which states that the distribution of a sample mean – after normalisation and scaling – converges to a standard normal distribution as the sample size increases), but this does not guarantee validity. In fact, the conditions for applying this theorem (notably the independence of observations and the existence of a finite variance) are not always met in practice. More generally, the power of a test comparing groups depends on several factors: the magnitude of the deviation sought, the variability of the data, the sample size of the groups, and the chosen significance level.

[83] An alternative approach, developed in recent scientific literature, involves using an algorithmic method to identify the groups (characterised by combinations of variables) for which the model shows the most marked deviations in results. This type of approach means that biases can be identified in sub-populations that would not have been examined spontaneously.

Ultimately, while univariate analyses provide a minimum baseline, financial institutions **are encouraged to expand them with multivariate approaches** in order to identify potentially more complex and severe biases.

5.4 Regarding the choice of metrics

The use of a group fairness metric makes it possible to assess the impact of decisions made by an AI model on different population groups. It has already been shown that three main families of metrics co-exist in the scientific literature, each with its own advantages and drawbacks (see Section 3), but that it is impossible to satisfy them all simultaneously – **choices must be made** (although it may be useful to examine the results obtained using several fairness metrics in turn, in order to analyse the model's behaviour).

This section is therefore intended to help guide a company's choices, recognising that, in principle, **no family of metrics can be favoured or ruled out in all cases**, but that choices must depend on the socio-technical context of the use case in question, as well as on the company's general policy (see Section 5.1).

These choices can be guided by **considering three questions in turn**.[84]

1/ Is there a regulatory requirement for demographic parity?

Such a requirement was, for example, traditionally present in the **United States**' regulatory framework (see Section 1.5). In France and the European Union, the question may notably arise in relation to **insurance pricing**. Since the CJEU's Test-Achats judgment, insurers may no longer differentiate premiums and benefits on the basis of gender (see Section 1.3). A precise analysis of the judgement shows, however, that it **does not require insurers to ensure demographic parity between men and women,**[85] **but rather prohibits the use of gender in pricing models**. In other words, it places a constraint on the variables input into the models, not on the results produced.

Insurers must therefore refrain from directly using gender; but the use of **variables correlated with gender** is permitted provided that it is based on risk factors that are objectively justified, proportionate and relevant to the actuarial assessment of risk. As such, variables such as vehicle characteristics, place of residence or no-claims bonus coefficients may, under current law, be taken into account. **As a result, differences in premiums between men and women may persist in practice.**

More generally, our analysis leads us to conclude that, to date, the French and European legal frameworks **do not appear to impose any requirement for demographic parity** in the financial sector's use cases.

2/ Is the model applied to groups with significant differences?

Independence metrics (demographic parity) **are only relevant when risk levels and the distributions of explanatory variables are relatively similar between groups**. They require

[84] The choice of a fairness metric can be formalised using decision trees. Although this is a fairly general approach, one example is the model developed by the University of Chicago (Saleiro et al., 2018).

[85] It is important to remember that in a regression problem such as pricing, demographic parity means that the distribution of prices must be similar between the groups being compared. A less strict version of demographic parity requires only that average rates be the same.

decisions to be equal regardless of underlying differences in risk. They are therefore **ill-suited to situations in which different social groups have substantially divergent baseline rates** (see Section 3). Similarly, when the frequency distributions of the explanatory variables – and the resulting scores – **differ significantly** between groups, the pursuit of demographic parity may necessitate substantial adjustments (particularly to decision thresholds), or even changes to the model. Adjustments such as these can result in a loss of information and, ultimately, reduced predictive effectiveness.

In these situations, it is generally preferable to apply **separation or sufficiency criteria**, which are more consistent with the **heterogeneity** of the risk observed.

When choosing between these two families of metrics, an analysis of the data characteristics is a key factor in the trade-off decision. Thus, when the **differences in baseline rates** between groups are significant, using sufficiency criteria may lead to a markedly stricter selection of disadvantaged groups. In this context, **favouring separation metrics can be justifiable**, as they ensure that individuals with the same actual level of risk are treated comparably, thereby helping to limit differences in access to positive decisions.

Taking into account the **distribution of explanatory variables** is **more delicate**. When these distributions differ significantly between groups, the scores produced by the model tend themselves to reflect these differences. In this case, **sufficiency often seems more consistent** with a risk-based approach, as it ensures that decisions retain the same probabilistic meaning. Conversely, separation may require more significant adjustments to compensate for these differences.

However, **the interpretation of these differences is decisive**. The core question, then, is: do the differences observed reflect a heterogeneity of risk that we actually want to take into account in the decision? Where they correspond to **factors deemed relevant and legitimate**, sufficiency allows this information to be preserved and ensures a consistent interpretation. However, where these differences primarily reflect **historical bias or structural inequalities** that should not be reproduced, it may be preferable to use separation metrics in order to limit their impact on decisions.

3/ Depending on use case and company policy, what form of statistical relevance should be prioritised?

The selection of the most relevant family of metrics – between separation and sufficiency – also depends on the **type of statistical performance** chosen to be prioritised. More specifically, the choice involves a fundamental trade-off between **two forms of reliability**: **precision** (or positive predictive value), i.e. the probability that a positive prediction[86] is correct; and **recall** (or sensitivity), i.e. the ability to correctly classify as positive all individuals who are actually positive. Prioritising precision means ensuring that the decisions made are **reliable**, while prioritising recall means **not missing any relevant cases**.

Within this framework, **sufficiency** metrics (predictive value parity) are naturally associated with **precision**: they ensure that, among the individuals that receive a positive decision, the proportion of cases that are actually positive is comparable between groups. Conversely, **separation** metrics (error rate parity) are associated with **recall**: they aim to ensure that individuals who are

[86] Here, the analysis is conducted within a framework of binary classification between a positive class and a negative class.

actually positive have equivalent rates of positive classification regardless of the group to which they belong.

These two approaches reflect **different priorities**. Sufficiency focuses on the **consistency of decisions**: the same score or decision must correspond to the same level of risk for all groups. It is therefore particularly well-suited to situations where **false-positive**-type errors (granting a benefit to an ineligible individual) are costly or socially undesirable. Conversely, separation focuses on **access to opportunities**: it aims to prevent certain groups from being under-represented among correctly identified individuals, and is particularly sensitive to **false-negative**-type errors (failing to recognise an eligible individual).

In our example of granting a loan, using sufficiency metrics fulfils an objective of **prudence**: limiting the granting of loans to borrowers likely to default, in order to protect both the institution and individuals from becoming overly indebted. Conversely, using sufficiency metrics fulfils an objective of **inclusion**: ensuring that creditworthy borrowers are not unfairly excluded from access to credit due to model errors.[87] The choice between these two approaches therefore amounts to striking a balance between a rationale of **positive lending decision reliability** and a rationale of **non-exclusion of eligible individuals**.

[87] This echoes the objectives of the AI Act.

Box 8: Do we need to strike a balance between fairness and performance?

The question of the trade-off between fairness and performance – understood in this instance as the accuracy of predictions – is a **central, but contentious, issue** in the scientific literature.

In some studies, this trade-off is **inherent**, in that it is **impossible to design** models that fully satisfy both objectives. Differences in distribution between groups (for example, distinct default rates) can thus make certain fairness metrics difficult to reconcile with high overall precision. Furthermore, the introduction of fairness constraints may limit the information the model can utilise, or exacerbate the limitations of imperfect (biased, noisy or incomplete) data, which can lead to a reduction in performance without fully correcting the underlying imbalances.

Other studies question the notion that a trade-off between fairness and performance is inevitable. They highlight the decisive role played by **design choices**: the selection of variables, the quality and representativity of the data, and the machine learning methods employed. From this perspective, taking greater account of the diversity of situations and using tailored techniques can help to improve both fairness and performance simultaneously, or, at the very least, ease the conflict between these objectives.

In practice, the nature of this trade-off is decisive. If it is **structural**, any improvement in fairness necessitates an explicit trade-off, which then comes down to **normative choices** between effectiveness and inclusion. If it is partly **contingent**, **technical scope** exists for simultaneously improving fairness and performance, which should be explored. In both cases, these **trade-offs** cannot be solely entrusted to the discretion of technical teams: they must be **explicitly stated, documented and incorporated** into suitable governance frameworks.

In the absence of a consensus, adopting a **pragmatic approach** ultimately seems essential. Institutions could thus demonstrate that they have actively sought solutions to improve both fairness and performance, notably through data quality and traceability, tailored testing, and the use of multiple assessment metrics, within the framework of strengthened model validation processes.

5.5 Regarding the thresholds to take into account

Once the relevant metric has been determined, the next step is to establish a threshold that defines a **problematic difference in treatment**.

The 80% threshold – often referred to as the "four-fifths rule" – applied to the demographic parity metric **figures prominently in the debate**, particularly in the United States, where it is commonly used, including in the financial sector (see Section 1.5). This threshold means that differences in treatment between two groups[88] must not exceed 20%: thus, the number of men who have been granted a given loan must not exceed that of women by more than 20%. However, this threshold is based on a **largely pragmatic legal rationale**, intended to provide a simple warning indicator, **rather than on any robust scientific foundations**. Indeed, **there is no basis for considering that a 20% difference represents** – in general and regardless of context – **a valid dividing line between a fair situation and an unfair situation**. Moreover, a mechanical application of this

[88] Either between two groups defined in binary terms (e.g. men and women), or between a reference group (e.g. adults aged 20 to 45) and other groups against which it is compared.

threshold could prove **unsuitable in certain situations**, for example when baseline selection rates are very low or very high,[89] or when sample sizes differ significantly between groups.

More generally, **the academic literature is largely in agreement that there is no universal, scientifically grounded threshold** that would allow a group fairness metric to be qualified as "acceptable" or "unacceptable" **irrespective of the context**. Research shows that, regardless of the metric considered, the relevance of a threshold depends on numerous factors: the size and structure of the populations compared, baseline rate levels, the objectives pursued by the model, etc.

Consequently, when determining the thresholds to be taken into account, **adopting a contextual approach** – based on an analysis of the magnitude of the differences between groups, their statistical robustness and their practical implications, potentially rounded out with sensitivity analyses[90] – is advisable, **rather than mechanically applying immutable thresholds**.

5.6 Regarding the choice of bias correction methods

Choosing an appropriate method for correcting bias depends on several aspects of the task at hand: the identified causes and types of bias, the degree of control over the AI system, the level of regulatory constraint, and so on. In this respect, the scientific literature suggests that it is **preferable to combine several of these methods**. A number of factors that can help in the decision are outlined below.

- If **several different models are likely to be used on the same dataset**, pre-processing methods seem more suitable, as they generally transform the data themselves. For example, if a company wishes to use separate models on a shared dataset for the assessment of creditworthiness and of pricing for loans, it may be worthwhile to use pre-processing to pool bias corrections.

- If the **model training process is firmly controlled**, integrated learning methods may be more appropriate: integrating a fairness constraint directly into the training can help strike an optimal balance between performance and fairness, particularly when focusing on a single protected variable.

- If **only a black-box model is available**, post-processing methods can still be used, provided that the model's numerical outputs (scores, probabilities, etc.) can be accessed. These methods primarily allow the final decisions to be adjusted using thresholds, although it is possible to recalibrate the outputs at the cost of a generally more resource-intensive procedure.

Whatever the case, the institution must **retain full and complete control over the bias correction process**. Regardless of the method(s) chosen, it is essential, first and foremost, to **scrupulously document** the changes made to the model and the data. Furthermore, the academic literature shows that overly strong corrections to one particular dimension can, in some cases, undermine fairness in other areas (for example, an improvement in fairness with regard to gender may come at the expense of fairness in terms of age). It is therefore **necessary to analyse the effects of corrections across all relevant dimensions**, and to make an explicit

[89] In the case of consumer credit with an average approval rate of around 95%, the differences between groups would be very unlikely to exceed 20%.
[90] Aimed at measuring the extent to which conclusions vary when certain parameters or methodological choices are altered (for example, the group formations or the chosen threshold).

trade-off between the expected benefits in terms of fairness and the requirements for modelling **sobriety** and **stability**. These trade-offs must be justified and transparently documented. Lastly, the bias correction process **must not increase the model's opacity**: the choices made must remain understandable and explainable, so as not to undermine confidence in the system and its governance.

6 Anticipating the rise of generative AI in the financial sector

This discussion paper focuses primarily on "traditional" predictive systems. These systems now account for the vast majority of models deployed on a large scale in the financial sector that are likely to pose risks in terms of fairness. However, the use of generative AI[91] (language models, multimodal models) is growing rapidly. Yet the bias assessment methods designed for traditional predictive models cannot be directly applied to these new systems, whose operating methods and output formats differ significantly.

Fairness in generative systems is structurally different

The concepts of bias presented in this paper are based on **two assumptions**: the existence of an observable target variable (such as default or fraud), and a decision that is comparable between groups. Generative systems deviate from this framework in several key respects.

- **Absence of a single unambiguous target**: a system that generates text or images does not predict a unique "true" value. It produces one from many possible responses, meaning that the concept of "error" is not defined in the same way as for a more traditional model.
- **Absence of an explicitly declared group**: the user does not generally indicate that they belong to a sensitive group. However, disparities may emerge from implicit clues, such as the language used, the register of expression or the context of a request.
- **Output that is symbolic in nature**: the model's outputs (text, images, etc.) are not directly observable decisions, but objects whose interpretation depends on the context and the recipient. The same response may therefore be perceived as acceptable in one situation but problematic in another.
- **Harm that is dependent on use**: the potentially biased effects do not stem solely from the content produced, but also from the way in which it is used. For example, a seemingly correct response may convey implicit stereotypes, the impact of which will depend on the context of the interaction and the sensitivities of the recipient.

These specific characteristics do not call into question the relevance of a fairness analysis, but they do profoundly alter the way it is carried out. The assessment cannot be limited to comparing indicators between groups; it must also consider the **representation of the world encoded by the model** and on the concrete forms through which this representation is expressed in the generated content.

[91] Formally, the term "General Purpose AI" (GPAI) should be applied, as in the AI Act. This term refers to systems designed to be used in a wide variety of tasks and contexts, without being limited to a specific use case. Strictly speaking, "generative" AI makes up a specific category of these systems, characterised by its ability to produce new content (text, images, code, etc.) based on a prompt.

Four categories of harm specific to generative systems

General-purpose models – and in particular large language models (LLMs) – can be the cause of several types of harm.[92] **Representational harms** may arise when the system associates certain groups with stereotypical roles or traits, for example in marketing content or interactions with chatbot applications. Moreover, **disparities in service quality** may emerge when performance varies depending on the language or register of expression, which is likely to have a concrete impact on customer relations, particularly for those with a limited command of the language in question.

Furthermore, these systems can lead to an **erosion of diversity**, as they converge towards implicit or dominant profiles in the recommendations they generate, to the detriment of the variety of real-world financial situations. **Allocational harms** may also arise downstream when the content generated feeds into automated or semi-automated decisions (KYC, anti-money laundering, credit applications). In this case, the harm, which is conventional in nature, is mediated by a system whose **operation remains largely opaque**, making it more difficult to identify, and therefore all the more critical.

Bias: an inherent property of the model, not just of its outputs

An important insight from recent scientific research on language models concerns the very nature of bias. Bias is not limited to certain visible responses: it is **rooted more deeply**, in the way the model "organises" its understanding of the world. During training, **the model constructs a representational space based on the correlations present in the data**. Within this space, certain dimensions correspond to social characteristics (such as gender), and many words or concepts are distributed across it according to sometimes stereotypical associations. In other words, **biases are not occasional anomalies, but structural properties of this internal representation**, and they do not disappear spontaneously as models become larger or more effective.

"Alignment" techniques implemented after training (such as reinforcement learning from human feedback,[93] direct preference optimisation[94] or security instructions) can reduce the visible expression of certain biases. However, **they mainly act on the surface**: they influence the form of the responses produced, **without fundamentally transforming the model's internal structure**.[95] In practice, this means that a model may provide seemingly satisfactory responses in simple or highly controlled situations, while allowing **biases to reappear in more complex contexts**, for example with long, indirect, multilingual or unusually phrased questions. Consequently, a robust assessment must explore a diverse range of contexts and formulations in order to better reveal the biases that could arise in real-world use conditions.

[92] Gallegos et al., 2024.
[93] Reinforcement learning from human feedback (RLHF) involves using human assessments to guide the behaviour of models. In practice, human annotators compare different responses generated by the model, and these preferences are then used to train a reward system that guides the model towards responses deemed more useful, safer or more appropriate.
[94] Direct preference optimisation (DPO) involves training the model to directly replicate human preferences between different responses, without resorting to an intermediate reward system, unlike RLHF.
[95] Wolf, Wies, Avnery, Levine and Shashua, 2024.

Three layers of assessment that can be used collectively

More generally, an assessment of the fairness of a generative AI system can be organised into three complementary layers, in order to capture the different risk dimensions:[96]

- **The representational layer** aims to measure bias directly within the model's **internal structure**. It relies on methods such as association tests on embeddings,[97] probing classifiers[98] or salient feature analysis.[99] This layer enables the identification of biases present in the model's initial representations (priors),[100] irrespective of alignment mechanisms. However, **this requires access to internal representations**, which is generally limited to in-house or open-source models.[101]
- **The behavioural layer** analyses bias based on the model's **outputs** by testing sets of prompts[102] that are representative of actual usage, including a variety of formulations (long, multilingual or adversarial prompts).[103] It enables an **assessment of what users actually observe in practice**. However, it may **underestimate latent biases that are not triggered by the tested scenarios**.
- **The allocative layer** examines biases when the model's outputs inform a **downstream decision**. It involves applying **standard tools** for fairness assessment (independence,

[96] Neumann, Kirsten, Zafar and Singh, 2025.
[97] Vector embedding is a way to represent a word, phrase or object as a set of numbers, so that the model can manipulate them mathematically. In this representation space, elements deemed similar (for example, words with similar meanings) are located close to one another. These internal representations structure the way in which the model organises information and establishes associations.
[98] A probing classifier is a simple model, trained on the internal representations of an AI model (embeddings), in order to test what information is present. For example, it can check whether these representations contain information on gender, age or other characteristics. This method allows us to analyse what the model has learnt, without altering how it works.
[99] The analysis of salient dimensions involves identifying, within the model's representation space, the directions (or axes) that capture the most variation or structure in the data. Some of these dimensions may correspond to interpretable characteristics, such as semantic or social categories (for example, gender). Analysing them helps understanding how the model organises information and which distinctions it prioritises in its internal representations.
[100] The model's initial representations (priors) refer to the set of knowledge and associations that the model has learnt during training, prior to any interaction with the user. They reflect the regularities present in the training data and structure the way in which the model interprets queries and generates responses.
[101] When a general-purpose model is provided by a third party, the user organisation does not usually have direct access to the model's internal representations. However, it can rely on the technical documentation that the provider is required to make available under the AI Act. In particular, this documentation must describe, to some extent, the training data, the model's characteristics, and the methods used to detect and mitigate bias (Annex XI of the AI Act).
[102] A prompt is the instruction or request addressed to a generative AI model (a question, instruction, text to be completed, etc.), which serves as the starting point for the response produced. The way in which a prompt is formulated – whether it is more or less precise, long, direct or structured – can significantly influence the content and quality of the model's response.
[103] Adversarial prompting: prompts formulated in such a way as to test the model's limits, or even to provoke errors or circumvent its safeguards. These may include, for example, ambiguous, indirect or deliberately misleading questions. These tests help to assess the system's robustness when confronted with non-standard or malicious use.

separation, sufficiency) to the entire process, incorporating the generative model as an intermediate step.

These three layers provide complementary insights. By limiting analyses to the behavioural layer alone – which is often the most readily accessible – structural biases in the model are overlooked: a system may appear satisfactory in tests while still harbouring latent biases. Conversely, focusing solely on the representational layer does not help in understanding whether biases actually translate into concrete impacts.[104]

A comprehensive assessment could therefore **combine these three levels**, with a degree of detail **proportionate** to the level of risk associated with the use case, in line with the **risk-based approach** adopted throughout this paper.

[104] A growing practice involves using another language model as a "judge" to evaluate the outputs of another system ("LLM-as-a-judge"). This approach offers operational advantages (automation, speed), but appears ill-suited to assessing bias. Indeed, the LLM-as-a-judge model may share the same representational biases as the system being assessed, and may itself have its own biases, which can limit its ability to detect inequalities. Above all, the assessment of biases and risks of discrimination relies on contextual judgement that cannot be fully automated.

References

Alvarez, J. M., Bringas Colmenarejo, A., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., . . . Ruggieri, S. (2024). Policy Advice and Best Practices on Bias and Fairness in AI. *Ethics and Information Technology, 26*(31). doi:10.1007/s10676-024-09746-w

Bachoc, F., Bolte, J., Boustany, R., & Loubes, J.-M. (2026). When Majority Rules, Minority Loses: Bias Amplification of Gradient Descent. *Advances in Neural Information Processing Systems, 38*, 30479–30517. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2025/hash/2bbc73b3d3c2de43743ce2d82c8f3d7d-Abstract-Conference.html

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. https://fairmlbook.org/.

Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.-M., & Risser, L. (2022). A Survey of Bias in Machine Learning Through the Prism of Statistical Parity. *The American Statistician, 76*(2), 188–198. doi:10.1080/00031305.2021.1952897

Binns, R. (2022). On the Apparent Conflict Between Individual and Group Fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 514–524. doi:10.1145/3351095.3372864

Caton, S., & Haas, C. (2024). Fairness in Machine Learning: A Survey. *ACM Computing Surveys, 56*(7), 1–38. doi:10.1145/3616865

Charpentier, A., & Barry, L. (2022, December). L'équité de l'apprentissage machine en assurance. *Statistique et Société, vol. 10, n° 3*.

Cook, D. I., Gebski, V. J., & Keech, A. C. (2004). Subgroup Analysis in Clinical Trials. *The Medical Journal of Australia, 180*(6), 289–291. doi:10.5694/j.1326-5377.2004.tb05928.x

Côté, M.-P., Côté, O., & Charpentier, A. (2024). Selection Bias in Insurance: Why Portfolio-Specific Fairness Fails to Extend Market-Wide. *SSRN*. doi:10.2139/ssrn.5018749

Das Jui, T., & Rivas, P. (2024). Fairness Issues, Current Approaches, and Challenges in Machine Learning Models. *International Journal of Machine Learning and Cybernetics, 15*, 3095–3125. doi:10.1007/s13042-023-02083-2

Deck, L., Müller, J.-L., Braun, C., Zipperling, D., & Kühl, N. (2024). Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness. *European Workshop on Algorithmic Fairness*. Retrieved from https://ceur-ws.org/Vol-3908/paper_39.pdf

Desrosières, A. (1993). *La Politique des grands nombres*. Paris: La Découverte.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemer, R. (2012). Fairness Through Awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Ehrhardt, A., Biernacki, C., Vandewalle, V., Heinrich, P., & Beben, S. (2021). Reject Inference Methods in Credit Scoring. *Journal of Applied Statistics, 48*(13–15), 2734–2754. doi:10.1080/02664763.2021.1929090

EIOPA. (2025). *Opinion on AI Governance and Risk Management*. Retrieved from https://www.eiopa.europa.eu/document/download/88342342-a17f-4f88-842f-bf62c93012d6_en

Estellat, C., De Rycke, Y., & Asselain, B. (2005). Intérêt et limites des analyses en sous-groupes dans les essais thérapeutiques. *Oncologie, 7*, 75–79. doi:10.1007/s10269-005-0298-6

Ewald, F. (2011). Omnes et Singulatim. After Risk. *Carceral Notebooks, 7*, 77–107.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M., Kim, S., Dernoncourt, F., . . . Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics, 50*(3), 1097–1179. doi:10.1162/coli_a_00524

Glenn, B. J. (2000). The Shifting Rhetoric of Insurance Denial. *Law & Society Review, 34*(3), 779–808. doi:10.2307/3115143

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM Journal on Responsible Computing*, 1–52. doi:10.1145/3631326

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385. doi:10.1145/3442188.3445901

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *Proceedings of the 35th International Conference on Machine Learning*, 2564–2572.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4069–4079.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, Mai 23). *How We Analyzed the COMPAS Recidivism Algorithm*. Retrieved from ProPublics: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Li, J., & Li, G. (2025). Triangular Trade-off between Robustness, Accuracy, and Fairness in Deep Neural Networks: A Survey. *ACM Computing Surveys, 57*(6), 1–40. doi:10.1145/364508

Loubes, J.-M., Clayes, E., Eynard, J., Lafargue, V., Rottembourg, B., & Prunkl, C. (2026). A Hitchhiker's Guide to Bias Evaluation. *Preprint: hal-05642033v1*.

Meding, K. (2026). It's Complicated. The Relationship of Algorithmic Fairness and Non-Discrimination Provisions for High-Risk Systems in the EU AI Act. *Workshop on Regulatable ML*. doi:10.48550/arXiv.2501.12962

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR), 54*(6), 1–25. doi:10.1145/3457607

Napoletani, D., Panza, M., & Struppa, D. C. (2011). Agnostic Science. Towards a Philosophy of Data Analysis. (Springer, Ed.) *Foundations of Science, 16*(1), 1-20. doi:10.1007/s10699-010-9186-7

Neumann, A., Kirsten, E., Zafar, M. B., & Singh, J. (2025). Position is Power: System Prompts as a Mechanism of Bias in Large Language Models (LLMs). *Proceedings of the 2025 ACM*

Conference on Fairness, Accountability, and Transparency, 573–598. doi:10.1145/3715275.3732038

OECD. (2026). Supervision of Artificial Intelligence in Finance: Challenges, Policies and Practices. *OECD Artificial Intelligence Papers*(54). doi:10.1787/92743dc1-en

Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys (CSUR), 55*(3), 1–44. doi:10.1145/3494672

Rosenblatt, L., & Witter, R. T. (2023). Counterfactual Fairness Is Basically Demographic Parity. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(12), 14461–14469. doi:10.1609/aaai.v37i12.26691

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., . . . Ghani, R. (2026). Aequitas: A Bias and Fairness Audit Toolkit. *arXiv*(18111.05577). doi:10.48550/arXiv.1811.05577

Simon, J. (1988). The Ideological Effects of Actuarial Practices. *Law & Society Review, 22*(4), 771–800. doi:10.2307/3053709

Verma, S., & Rubin, J. (2018). Fairness Definitions Explained. *FairWare '18: Proceedings of the International Workshop on Software Fairness*, 1–7. doi:10.1145/3194770.3194776

Westerstrand, S. (2025). Fairness in AI Systems Development: EU AI Act Compliance and Beyond. *Information and Software Technology, 187*(107864). doi:10.1016/j.infsof.2025.107864

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy, 8*, 78–115. doi:10.5325/jinfopoli.8.2018.0078

Wolf, Y., Wies, N., Avnery, O., Levine, Y., & Shashua, A. (2024). Fundamental Limitations of Alignment in Large Language Models. *Proceedings of the 41st International Conference on Machine Learning, 235*, 53079–53112. Retrieved from https://dl.acm.org/doi/abs/10.5555/3692070.3694246