European Review of Digital Administration & Law - Erdal
2026, Volume 7, Issue 1, pp. 183-204
ISSN 2724-5969 – ISBN 979-12-218-2846-7 – DOI 10.53136/9791221828467 17

*Can Inequality Be Measured? Techniques for Assessing Algorithmic Bias**

Pere Simón Castellano
(Associate Professor of Constitutional Law at International University of La Rioja (UNIR), Spain)

ABSTRACT This article asks whether inequality produced by artificial intelligence systems can be measured in a manner that is technically meaningful and legally useful. It argues that algorithmic bias is measurable only when three conditions are met: the protected or vulnerable groups are identified before testing, the statistical metric is selected by reference to the legal and factual context of use, and the result is interpreted through a governance process rather than treated as a self-executing answer. The analysis is situated in European Union law, especially Regulation (EU) 2024/1689 (Artificial Intelligence Act), and is methodologically grounded in the Spanish AI Supervisory Agency's (AESIA) Guide on data and data governance, Annex C, and UNE-CEN/CLC ISO/IEC/TS 12791:2024. The article develops an operational framework for employment-related and biometric systems, using adverse impact, true positive rate differences, false negative rate differences and statistical robustness as examples of measurable thresholds. Empirical illustrations from automated recruitment and facial recognition show both the promise and the limits of measurement. The conclusion is deliberately cautious: metrics can reveal inequality, structure regulatory evidence and trigger mitigation duties, but they cannot replace the normative judgment required by equality law, data protection law and fundamental rights impact assessment. The article's working hypothesis is that algorithmic inequality can be measured as evidence, but cannot be resolved through metrics alone for the purposes of legal compliance. Its practical contribution lies in proposing a rights-sensitive assessment model that integrates data provenance, disaggregated performance metrics, context-dependent acceptance thresholds, lifecycle governance controls, and residual risk documentation.

KEYWORDS: Algorithmic bias - Non-discrimination - Artificial Intelligence Act (AI Act) - Fairness metrics - Fundamental rights impact assessment - High-risk AI systems - Bias mitigation - Data governance

TABLE OF CONTENTS: 1. Introduction. – 2. Bias as a Lifecycle Risk. – 2.1. Why Bias Cannot Be Reduced to Bad Data. – 2.2. The Legal Object of Measurement. – 3. Techniques for Measuring Algorithmic Inequality. – 3.1. Disaggregated Performance as the Minimum Baseline. – 3.2. Group Fairness Metrics. – 3.3. Counterfactual and Proxy Analysis. – 4. A Methodological Framework for Acceptance Criteria. – 4.1. From Inherent Risk to Residual Risk. – 5. Empirical Illustrations. – 5.1. Automated Recruitment. – 5.2. Facial Recognition. – 5.3. Case-Law Summary: SyRI. – 6. Measurement Under EU Regulatory Governance. – 6.1. Provider Duties: Data, Documentation and Risk Management. – 6.2. Deployer Duties and Fundamental Rights Impact Assessment. – 6.2.1. From Legislative Silence to the Article 27 Duty. – 6.2.2. Subjective and Objective Scope of the Duty. – 6.2.3. Minimum Content: From Template to Reasoned File. – 6.2.4. Comparative Methodological Lessons. – 6.2.5. Avoiding Cosmetic Compliance. – 6.3. Special Categories of Data and the Paradox of Bias Testing. – 7. Treatment Measures and Standardisation. – 7.1. Implementation Challenges in the Public Sector. – 8. Conclusion.

1. *Introduction*

Artificial intelligence systems are increasingly involved in decisions that distribute opportunities, burdens and social recognition. In recruitment, credit, education, access to public services and biometric identification, a model may decide, recommend, rank, or filter in ways that affect the practical enjoyment of equality. The question posed by this article – can inequality be measured? – is therefore not merely technical. It is a legal and institutional question about the kind of evidence that should count when a system is suspected of reproducing discrimination or creating new forms of disadvantage.[1]

The European Union has now moved from soft-law exhortation to binding obligations. As of April 2026, Artificial Intelligence Act is in force, although most requirements for high-risk systems apply from 2 August 2026 and some obligations linked to Article 6(1) apply from 2 August 2027.[2] The Regulation does not prescribe a single fairness metric. Instead, it

*Article submitted to double-blind peer review.
This article develops and expands the author's contribution to the workshop 'Transparencia algorítmica 2.0: a propósito del caso BOSCO' (Universidad Complutense de Madrid, Faculty of Law, 12 December 2025), organised within the research project 'Fortalecimiento del Estado de Derecho a través de la Inteligencia Artificial' (PID2024-157561NB-I00), funded by the Spanish Ministry of Science (2024 Call for 'Proyectos de Generación de Conocimiento', 'Proyectos de Investigación No Orientada', State Plan for R&D Projects 2024).

[?] European Parliament and Council, Regulation (EU) No. 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), in Official Journal of the European Union, L 2024/1689, 12 July 2024, arts. 9, 10, 13, 14, 27 and Annex III.

[2] Regulation (EU) No. 2024/1689, cit., art. 113.

requires risk management, data governance, technical documentation, transparency, human oversight and, for certain deployers, a fundamental rights impact assessment. This regulatory design matters: it treats bias as a quality, safety and rights risk that must be examined, documented and mitigated, but it leaves the choice of concrete tests to context-sensitive technical and legal judgment.

Recent European legal scholarship has approached these duties not as isolated technical requirements, but as parts of an emerging architecture of risk governance, data governance, transparency, human oversight and fundamental-rights control. The collective Treatise on the European Union Artificial Intelligence Regulation[3] is useful in this respect because it is structured around thematic blocks – including governance, fundamental rights, risk management, and accountability – offering cross-cutting insights into the Regulation's architecture and legal implications. Mantelero situates the Regulation within a risk-based and human-centred regulatory response;[4] Simón Castellano's chapters on unacceptable artificial intelligence systems and on Article 9 show how the prohibition layer and the high-risk risk-management system operate as complementary guarantees;[5] Loza Corera, Gutiérrez David and Cerrillo i Martínez connect data governance and transparency duties with the information that deployers need in order to assess bias in context;[6] and Lazcoz Moratinos warns that human oversight under Article 14 cannot be reduced to a merely formal requirement.[7]

The present article builds on that broader literature, but narrows the inquiry to one operational question: how can algorithmic bias be translated into structured evidence capable of supporting a fundamental rights impact assessment, a risk-management file and, eventually, legal accountability?

The central thesis is captured in a deliberately compact formula: algorithmic inequality is measurable as evidence, but not decidable as compliance by metric alone. This formula is intended to capture the double movement that runs through the article. On the one hand, unequal selection rates, unequal error rates, unequal calibration and unequal data coverage are measurable facts. They can be tested, compared, monitored and documented. On the other hand, the legal meaning of those facts depends on the protected ground, the right affected, the severity of the consequence, the possibility of justification and the adequacy of safeguards.

The article's original contribution lies in translating that formula into a framework that links metrics and governance. The bridge connects four languages that often remain separate: the statistical language of fairness metrics, the lifecycle language of technical standards, the risk-management language of Artificial Intelligence Act, and the rights language of equality law. The proposed framework does not ask lawyers to become data scientists, nor data scientists to decide discrimination law. It asks both communities to produce an auditable chain of reasoning: why this metric, why this threshold, why this group, why this mitigation, and why the residual risk can or cannot be accepted.

This distinction is critical because the most consequential disputes will rarely concern whether a calculation was arithmetically correct. Rather, they will concern whether that calculation was legally and normatively relevant. A model may satisfy demographic parity while imposing unacceptable false positive rates on a vulnerable group. It may equalise true positive rates while embedding

[3] L. Cotino Hueso and P. Simón Castellano (dirs.), *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Cizur Menor, Aranzadi, 2024.

[4] A. Mantelero, *El Reglamento de inteligencia artificial: la respuesta del legislador europeo a los retos de la inteligencia artificial*, in *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea,* 2024*,* 53.

[5] P. Simón Castellano, *El resto de los sistemas de inteligencia artificial prohibidos o inaceptables en el Reglamento*, in *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea,* 2024, 275; P. Simón Castellano, *Los sistemas de gestión de riesgos como obligación específica para los sistemas de inteligencia artificial de alto riesgo en el artículo 9 del Reglamento*, in *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea,* 2024, 535.

[6] M. Loza Corera, *Datos y gobernanza de datos y conexiones con principios protección de datos en el artículo 10 del Reglamento*, in *Tratado*, cit., 565; M.E. Gutiérrez David, *Transparencia y comunicación de información a los responsables del despliegue en el artículo 13 Reglamento de inteligencia artificial*, in *Tratado*, cit., 627; A. Cerrillo i Martínez, *El artículo 50 del Reglamento y las obligaciones de transparencia de los proveedores y responsables del despliegue de determinados sistemas de inteligencia artificial*, in *Tratado*, cit., 793.

[7] G. Lazcoz Moratinos, *La vigilancia o supervisión humana en el artículo 14 del Reglamento de inteligencia artificial: ¿un mero requisito obligatorio para los sistemas de alto riesgo?*, in *Tratado*, cit., 681.

unjustified group-specific thresholds. It may reveal no statistically significant disparity simply because the protected group was underrepresented in the test data. It may meet a vendor's benchmark while failing in the local deployment context. The contribution of a rights-sensitive methodology is to render such questions visible – before harm is standardised, legitimised, or overlooked.

For that reason, the spreadsheet model should not be understood as a mere compliance checklist. Its stronger function is evidentiary: it turns bias assessment into a reasoned administrative file. Each row of a test plan, risk register, or control library answers a future question that a regulator, judge, auditor, works council, data protection officer or affected person may ask. The document is not simply proof that the organisation filled in a template. It is proof of what the organisation knew, when it knew it, which alternatives it considered and how it justified the decision to deploy, restrict, redesign or stop the system.

The article develops that judgment through a lifecycle methodology. It draws on the AESIA Guide on data and governance, especially Annex C, which distinguishes sources of bias, techniques for evaluating bias and measures for treating bias.[8] It also relies on UNE-CEN/CLC ISO/IEC/TS 12791:2024, which addresses the treatment of unwanted bias in machine-learning tasks of classification and regression.[9] This combination allows the legal analysis to avoid two errors. The first is formalism: assuming that compliance can be inferred from the existence of a policy, a human-in-the-loop label or a general accuracy score. The second is metric absolutism: assuming that a numerical parity result is equivalent to legal fairness.

The argument proceeds in six steps. Section 2 defines algorithmic bias as a relation between data, models, institutional context and legally protected interests. Section 3 explains how bias can be measured through disaggregated performance, adverse impact and error analysis. Section 4 translates the methodology into acceptance criteria inspired by the impact-assessment workbook accompanying this article. Section 5 discusses two empirical illustrations: automated recruitment and facial recognition. Section 6 connects measurement to Artificial Intelligence Act, Spanish equality law and data protection. Section 7 concludes by proposing that bias assessment should be understood as a form of structured evidence, not as an automated verdict.

2. *Bias as a Lifecycle Risk*

2.1. *Why Bias Cannot Be Reduced to Bad Data*

Bias in artificial intelligence is often described as a problem of biased data. That description is useful but incomplete. Data may be unrepresentative, historically contaminated, unevenly labelled or incomplete; yet a system may also become biased because the objective function is poorly chosen, because thresholds are calibrated for one population and deployed in another, because users over-rely on automated recommendations, or because feedback loops transform an initially small disparity into a durable institutional pattern. ISO/IEC TR 24027 and the AESIA Guide both adopt this broader understanding of bias as a phenomenon that can enter at multiple points of the system lifecycle.[10]

Annex C.1 of the AESIA Guide is particularly useful because it separates data-related sources of bias from human cognitive and social sources. Selection bias, coverage bias, participation bias, labelling bias, missingness, data-preparation bias and outliers can all affect whether the dataset represents the population for whom the system will be used. At the same time, automation bias, implicit bias, in-group bias, out-group homogeneity and social bias explain how human decisions around data collection, labelling, interpretation and deployment can reproduce inequality even when no explicit discriminatory variable is present.[11]

This lifecycle view has a direct legal consequence. A provider cannot defend a high-risk system merely by showing that a

[8] Agencia Española de Supervisión de Inteligencia Artificial, *Guía de datos y gobernanza de datos*, A Coruña, AESIA, 2025, 57-63.

[9] Asociación Española de Normalización, UNE-CEN/CLC ISO/IEC/TS 12791:2024. *Information technology - Artificial intelligence - Treatment of unwanted bias in classification and regression machine learning tasks* (ISO/IEC TS 12791:2024), Madrid, UNE, 2024.

[10] ISO/IEC, ISO/IEC TR 24027:2021. *Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making*, Geneva, ISO, 2021.

[11] AESIA, *Guía de datos y gobernanza de datos*, cit., Annex C.1-C.3, 57-63.

protected attribute was removed from the dataset. In many domains, protected status is encoded by proxies: postcode, education path, career interruptions, working patterns, language variety, disability-related accommodations or historical interactions with public authorities. Conversely, a deployer cannot rely on a global accuracy score if a subgroup experiences systematically higher error rates. Bias must therefore be assessed in relation to the intended purpose, the affected persons, the operating context and the decision flow in which the system is embedded.

A lifecycle account also prevents a common category error. Bias is sometimes treated as a property of the model, as if the question were exhausted once the classifier has been tested in isolation. In practice, bias can be introduced by the problem definition, by the target variable, by the training data, by the feature set, by the interface through which humans see the recommendation, by the incentives of human reviewers, by the complaint mechanism or by post-deployment drift. Measuring inequality only at model-validation stage is therefore under-inclusive. It is comparable to assessing the safety of a medical device without considering how it will be used in the hospital.

The lifecycle proposed here has five phases. Inception defines the purpose of the system, the affected population, the protected groups and the acceptance criteria. Design and development select variables, labels, modelling techniques and mitigation strategies. Verification and validation test data and outputs by group, including intersectional groups where sample size permits. Operation and monitoring examine whether subgroup performance changes in real use. Retirement or replacement preserves the documentation necessary to understand historical harm, investigate incidents and prevent recurrence.

Phase	Function in the bias assess-ment
Inception	Define purpose, decision context, af-fected groups, legal require-ments, an-ti-bias criteria and acceptance thresh-olds before technical development.
Design and deve-lopment	Select features, prepare and label data, train the model, document feature ra-tionality and assess possible demo-graphic proxies.
Verification and validation	Measure global perfor-mance and dis-aggregated perfor-mance; test outputs by group; verify acceptance criteria and document failures.
Operation and monitoring	Monitor drift, complaints, human oversight, subgroup perfor-mance and continuing effectiveness of mitigation controls.
Retirement or re-placement	Preserve relevant documen-tation, manage deletion or retention of data and investigate historical incidents or latent discriminatory effects.

Table 1. Lifecycle phases for discriminatory-impact assessment

This phase structure is not merely descriptive. It distributes responsibility. The legal or compliance team cannot wait until validation to ask whether a protected group is affected; the data team cannot wait until deployment to ask whether the test data were representative; the human resources team cannot wait until a complaint to ask whether the model is excluding qualified persons. A lifecycle method gives each actor a time and a record in which to act.

Given the technical nature of certain fairness metrics referenced throughout the article, a brief glossary is provided to ensure conceptual clarity. The objective is not to replicate the language of engineering, but to clarify the normative relevance of each metric within legal and regulatory reasoning, particularly for readers unfamiliar with the specifics of machine-learning evaluation methods.

Term	Full form	Legal relevance in this article
TPR	True positive ra-te	Percentage of persons who genuinely satisfy the criterion and are correctly identified by the system; central to equality of opportunity.
FPR	False positive ra-te	Percentage of persons who do not satisfy the criterion but are accepted or flagged by error; important where false accusa-tions or undue burdens are at stake.
FNR	False negative rate	Percentage of persons who sat-isfy the criterion but are reject-ed or missed by error; crucial in recruitment and access-to-opportunity contexts.

Term	Full form	Legal relevance in this article
DI	Disparate impact	Ratio comparing positive-outcome rates between groups; useful for screening indirect discrimi-nation but not a com-plete fairness test.
AUC	Area under the curve	Global measure of a classifier's ability to distinguish classes; useful for model quality but not sufficient for equality anal-ysis.
RMSE	Root mean squa-red error	Regression error metric that penalises larger errors; should be compared by group where predictions affect opportunities or burdens.
MAE	Mean absolute error	Average absolute regression error; easier to interpret and useful for comparing predic-tion error across groups.

Table 2. Essential metrics and terms used in the assessment

The glossary also underlines a substantive point. Accuracy, AUC, RMSE and MAE are quality metrics; they do not by themselves measure discrimination. TPR, FPR, FNR and DI are closer to equality analysis because they reveal who receives the benefit, who bears the error and whether mistakes are distributed unevenly. The legal assessment begins when those technical differences are connected to protected grounds and rights.

The metrics, then, function not as ends in themselves but as evidence. They help trace the structural effects of algorithmic decisions on access, recognition, and burdens. Their legal significance emerges only when interpreted through the lens of fundamental rights, non-discrimination, and proportionality. The article's methodology treats them as part of an integrated assessment framework – one that supports, rather than replaces, legal judgment.

2.2. *The Legal Object of Measurement*

The legal object of measurement is not abstract unfairness. It is the risk that an artificial intelligence system will impair equality before the law, non-discrimination, dignity, privacy or access to essential opportunities. The Charter of Fundamental Rights prohibits discrimination and guarantees equality before the law.[12] Spanish Law 15/2022 expressly includes artificial intelligence and massive data management within its material scope and promotes impact assessments to determine possible discriminatory bias in algorithmic decision-making.[13] EU equality directives, in turn, prohibit direct and indirect discrimination in employment and other protected areas.[14]

That legal framework explains why a bias metric should not be chosen simply because it is mathematically convenient. In employment, the relevant harm may be exclusion of qualified candidates, unequal access to promotion or a false negative that prevents a worker from being considered. In biometric identification, the relevant harm may be a false match that exposes a person to police scrutiny, or a false non-match that denies access to a service. In credit, the relevant harm may be denial, worse conditions or unjustified risk labelling. Each context changes which error matters most, which threshold is tolerable and which mitigation is proportionate.

3. *Techniques for Measuring Algorithmic Inequality*

3.1. *Disaggregated Performance as the Minimum Baseline*

The minimum condition for measuring algorithmic bias is disaggregation. A model that performs well on average may perform poorly for a protected or vulnerable group if that group is small, underrepresented or more difficult to classify under the available features. Overall accuracy is therefore an insufficient compliance indicator. For classification systems, the evaluator should

[12] European Union, Charter of Fundamental Rights of the European Union, in Official Journal of the European Union, C 326, 26 October 2012, arts. 20 and 21.

[13] Spain, Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación, in *Boletín Oficial del Estado*, no. 167, 13 July 2022, arts. 3, 4 and 23.

[14] Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation, in Official Journal of the European Communities, L 303, 2 December 2000; Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, in Official Journal of the European Communities, L 180, 19 July 2000; European Parliament and Council, Directive 2006/54/EC of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation, in Official Journal of the European Union, L 204, 26 July 2006.

compute the confusion matrix separately for each relevant group: true positives, true negatives, false positives and false negatives. From that matrix, accuracy, precision, recall, true positive rate, false positive rate and false negative rate can be calculated by group.

This approach does not require the assumption that every observed difference is unlawful. It only requires the evaluator to make the distribution of errors visible. Once visible, the difference can be tested for statistical significance, compared with predefined tolerances and interpreted in light of the decision's legal and social consequences. Without that step, bias remains a governance blind spot: the system may pass a general performance review while failing the people for whom the decision is most consequential.

3.2. *Group Fairness Metrics*

Group fairness metrics translate equality concerns into statistical comparisons. Demographic parity compares the rate of positive outcomes across groups. Adverse impact analysis, often expressed through the disparate impact ratio, compares the selection rate of a protected group with that of a reference group. Equal opportunity compares true positive rates: among persons who genuinely satisfy the relevant criterion, do different groups have the same probability of being selected? Equalized odds compares both true positive and false positive rates. Calibration asks whether a score has the same empirical meaning across groups.[15]

These metrics are indispensable but mutually tense. In many real datasets, particularly where base rates differ across groups for reasons that may themselves reflect structural inequality, it may be impossible to satisfy all fairness criteria simultaneously.[16] The leading technical literature therefore treats fairness metrics as limited instruments that must be connected to the system's objective, decision threshold and social context.[17] The impossibility results do not make fairness measurement futile. Rather, they require an explicit choice about the equality value that should govern the system. In a recruitment tool, equal opportunity and false negative control may be more legally salient than equal false positive rates, because the primary harm is exclusion of qualified candidates. In fraud detection, false positives may be more salient because they trigger investigation, stigma or interruption of benefits.

A useful way to avoid metric shopping is to distinguish between gateway metrics, harm-specific metrics and explanatory metrics. Gateway metrics ask whether the system is technically reliable enough to be considered at all: accuracy, area under the curve, mean absolute error or root mean squared error. Harm-specific metrics ask whether the error or outcome distribution maps onto a legally relevant harm: adverse impact for unequal access, true positive rate differences for unequal opportunity, false negative rate differences for wrongful exclusion, false positive rate differences for wrongful burdening or undue preference. Explanatory metrics then help identify why a disparity occurs, for example through feature importance, proxy correlation, missingness analysis or counterfactual tests.

This hierarchy matters because a single table of fairness metrics can be misleading if the metrics are treated as interchangeable. In recruitment, a false negative means a person who should be considered is filtered out. In welfare fraud detection, a false positive may lead to investigation, stigma or interruption of payments. In biometric identification, a false match and a false non-match have different rights consequences. A rights-sensitive assessment therefore begins with the harm and moves to the metric, not the other way around.

The evaluator should also resist the temptation to interpret the absence of a threshold breach as the absence of bias. A system may remain formally within a ten-percentage-point tolerance while still producing unacceptable consequences for a small group if the decision is irreversible or concerns a fundamental entitlement. Conversely, a larger difference may be explainable where the sample is small, the confidence interval is wide, the system is used only as a non-decisive support tool, and

[15] M. Hardt, E. Price and N. Srebro, *Equality of Opportunity in Supervised Learning*, in *Advances in Neural Information Processing Systems*, vol. 29, 2016, 3315.

[16] J. Kleinberg, S. Mullainathan and M. Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, presented at *Innovations in Theoretical Computer Science*, Berkeley, CA, 2017.

[17] S. Barocas, M. Hardt and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, Cambridge, MA, MIT Press, 2023.

meaningful human review corrects the effect. Thresholds are governance triggers; they are not miniature courts.

For legal purposes, the most valuable output is not the metric in isolation but the comparison between metric, threshold, justification and control. If the adverse impact ratio falls below 0.80, the register should show whether the dataset is unbalanced, whether the relevant protected attribute is directly or indirectly represented, whether alternative features were tested, whether a less discriminatory model exists and whether deployment has been paused, limited or redesigned. That chain of reasoning is what makes measurement useful in litigation and supervision.

Legal scholarship has warned against the temptation to automate fairness by selecting one metric and treating it as a proxy for compliance.[18] More recent technical literature similarly shows that formal fairness definitions can perform poorly when detached from policy goals, institutional context and distributional consequences.[19] The proper role of metrics is therefore evidentiary and diagnostic: they identify disparities, support comparison across interventions and create an audit trail for legal accountability.

3.3. *Counterfactual and Proxy Analysis*

A second family of techniques asks whether the same individual would have received a different result if a protected attribute or a correlated proxy had changed while relevant qualifications remained constant. Counterfactual analysis is attractive because discrimination law often asks whether a protected characteristic made a difference. However, counterfactual fairness is difficult to operationalise without causal assumptions. Changing gender, disability or ethnic origin in a dataset is rarely a neutral edit: those attributes may be correlated with life histories shaped by structural inequality. The evaluator must therefore document which variables are held constant, which are allowed to vary and why.

Proxy analysis is more immediately practical. It examines whether variables that appear neutral operate as substitutes for protected attributes. In employment, examples may include gaps in employment history, school names, postcode, extracurricular activities, language markers or availability patterns. In public-sector risk scoring, proxies may include neighbourhood, household composition, receipt of benefits or interaction with administrative systems. Proxy analysis should be linked to feature rationality: if a feature cannot be justified by the intended purpose and creates group disparity, it should be removed, transformed, constrained or subject to stronger oversight.

4. *A Methodological Framework for Acceptance Criteria*

The impact-assessment workbook prepared for this article operationalises the AESIA and ISO/IEC TS 12791 approach in six elements: scope and context, protected groups, lifecycle checklist, data and group-risk documentation, dynamic testing metrics, and risk-control registration. Its most important contribution is procedural: tolerances are defined before evaluation. This prevents retrofitting the metric to the preferred result and creates a record of why a disparity was accepted, mitigated or treated as non-deployable.[20]

The workbook can be understood as a normative architecture in miniature. It begins with the definition of scope because a metric has no stable meaning without an intended purpose. It then requires the evaluator to identify protected attributes and possible proxies because non-discrimination law is concerned with effects, not only explicit variables. It asks for data provenance, representativeness and coverage because model errors are often generated before the model is trained. It finally links each finding to a control, a responsible person and a residual-risk decision. That structure is what turns a technical test into a legally usable assessment.

In methodological terms, the proposal is distinctive because it brings Article 10, Article 27, AESIA Annex C and ISO/IEC TS 12791 into the same evidentiary frame. Article 10 focuses on the provider's data governance

[18] S. Wachter, B. Mittelstadt and C. Russell, *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI*, in *Computer Law & Security Review*, vol. 41, 2021, 105567.

[19] S. Corbett-Davies, J.D. Gaebler, H. Nilforoshan, R. Shroff and S. Goel, *The Measure and Mismeasure of Fairness*, in *Journal of Machine Learning Research*, vol. 24, no. 312, 2023, 1.

[20] UNE-CEN/CLC ISO/IEC/TS 12791:2024, cit., clauses 5 and 6.

duties. Article 27 focuses on the deployer's concrete context of use and fundamental rights impact. AESIA Annex C gives a taxonomy of bias sources, evaluation techniques and treatment measures. ISO/IEC TS 12791 supplies a lifecycle treatment logic for unwanted bias in classification and regression tasks. The article's claim is that discriminatory-impact assessment is strongest when these four layers are read together rather than sequentially.

Legal requirement	**AI Act anchor**	**Technical/ methodo-logical anchor**	**Assessment artefact**	**Contribution to the article**
Risk manage-ment	Article 9	ISO/IEC TS 12791 clauses 5 and 6	Risk register; likelihood, severity, controls and residual risk	Bias is treated as a lifecycle risk, not a one-off de-fect.
Data gover-nance	Article 10	AESIA Annex C.1 and C.2	Data prove-nance, repre-setativeness, missingness and proxy analysis	Static data tests become evidence of compliance.
Transparency to deployers	Article 13	ISO/IEC TS 12791 verifica-tion and valida-tion documen-tation	Instructions for use; known limi-tations; sub-group per-formance	Deployers re-ceive infor-mation need-ed for con-text-specific assess-ment.
Human over-sight	Article 14	AESIA Annex C.3	Escalation rules, re-viewer train-ing and au-tomation-bias controls	Human re-view is mean-ingful only if designed and documented.
Accuracy and robustness	Article 15	ISO/IEC TS 12791 testing and monitoring	Global and disaggregat-ed perfor-mance, ro-bustness and drift moni-toring	Technical quality is as-sessed global-ly and by the affected group.
Fundamental rights impact assessment	Article 27	Comparative FRIA method-ologies and risk matrices	Affected groups, spe-cific risks, safeguards, complaints and redress	The deployer translates provider evi-dence into a concrete rights file.
Post-market monitoring	Articles 72 and 73	ISO/IEC TS 12791 opera-tion and moni-toring	Incident re-porting, complaints, periodic re-testing and corrective action	Bias assess-ment contin-ues after de-ployment.

Table 3. Normative correspondence for bias assessment

The table also clarifies why Article 27 deserves more attention in the bias debate. If Article 10 asks whether the provider prepared the system responsibly, Article 27 asks whether the deployer can use it responsibly in a particular institutional setting. That distinction is decisive for discrimination. A recruitment model tested across Europe may still be inappropriate for a local labour market. A biometric system validated on benchmark datasets may still be risky in a public authority's lighting conditions or population mix. The rights impact is not only in the model; it is in the situated act of deployment.

The table below translates the workbook's acceptance criteria into article form. The thresholds are not proposed as universal legal safe harbours. They are operational tolerances for an employment-related high-risk system, to be justified by the provider or deployer, validated by technical teams and reviewed by legal, human resources and affected-stakeholder representatives.

Criterion	**Concrete metric**	**Threshold / tolerance**	**Rationale**
Functional correctness	Global accuracy / AUC for classifi-cation; RMSE or MAE for re-gression	At least the validated in-ternal benchmark or sectoral benchmark	The system must satisfy its de-clared purpose before bias is as-sessed.
Disaggregated perfor-mance by protected group	Accuracy and true positive rate by group	Difference no greater than 10 percentage points	Good global per-formance is in-sufficient if a protected group receives degraded perfor-mance.
Adverse im-pact	Disparate impact ratio	At least 0.80	Useful orienta-tion for indirect-discrimination screening in re-cruitment, subject to context.
Equality of opportunity	True positive rate difference between groups	No greater than 10 per-centage points	Controls unjusti-fied exclusion of persons who sat-isfy the relevant criterion.
Critical error control	False negative rate difference between groups	No greater than 10 per-centage points	Minimises erro-neous exclusion in employment-impacting deci-sions.

Criterion	Concrete metric	Threshold / tolerance	Rationale
Undue preferential error control	False positive rate difference between groups	No greater than 5 percentage points	Avoids unjustified favouring or burden shifting without objective reason.
Statistical robustness	Statistical significance and adequate sample size	$p < 0.05$ or sustained differences in representative samples	Prevents governance decisions based on noise or insufficient subgroup data.

Table 4. Acceptance criteria for an employment-related bias impact assessment

Three features of this table deserve emphasis. First, functional correctness comes before fairness comparison: a system that fails its stated task cannot be saved by equalising its errors. Secondly, the table separates outcome disparity from error disparity. Disparate impact addresses the distribution of positive outcomes, whereas true positive rate, false negative rate and false positive rate differences address the distribution of mistakes. Thirdly, statistical robustness is not a technical luxury. It is a legal safeguard against overreacting to noise or concealing disparity behind insufficient sample sizes.

4.1. *From Inherent Risk to Residual Risk*

The workbook also uses a probability-by-severity risk matrix. Probability is high where differences are clear, sustained, statistically supported or exceed a predefined threshold; medium where results are close to thresholds or unstable; and low where differences remain within tolerances. Severity is high where the system directly affects access to employment, promotion or salary, where automation is decisive, or where exclusion is difficult to reverse. Severity is medium where the impact is relevant but correctable through meaningful review, and low where the impact is limited or readily reversible.

Dimension	Level	Opera-tional description
Likelihood	High	Clear and sustained group differences; threshold breach; statistical support or complaints.
Likelihood	Medium	Differences close to thresholds; inconsistent results; limited significance; monitor-ing required.
Likelihood	Low	Metrics within tolerances; no material threshold breach; no external indications.
Severity	High	Direct effect on access to employment, promotion or salary; hard-to-reverse exclusion.
Severity	Medium	Relevant but correctable impact; meaningful human review and correction are available.
Severity	Low	Limited or reversible impact; effective human oversight.

Table 5. Probability and severity criteria for inherent bias risk

This risk logic is consistent with Artificial Intelligence Act's process orientation. The Regulation does not say that a system is lawful because the disparate impact ratio is 0.81 or unlawful because it is 0.79. It requires the provider to identify, analyse, evaluate and mitigate risks, to document the evidence and to monitor the system after deployment. A metric is therefore a trigger for governance action: it may require retraining, rebalancing, threshold adjustment, additional data collection, removal of proxies, stronger human review, a limitation of the intended purpose or non-deployment.

5. *Empirical Illustrations*

5.1. *Automated Recruitment*

The widely reported Amazon recruitment experiment remains a useful example because it shows how historical data can convert past labour-market inequality into a predictive rule. Reuters reported that the experimental system was trained on resumes submitted over a ten-year period and that most of those resumes came from men. The model learned patterns associated with historically successful candidates and reportedly penalised terms such as ‘women's’, as in ‘women's chess club captain’, as well as graduates of some all-

women's colleges.[21]

A legally meaningful audit would not ask only whether the model used gender as an explicit input. It would ask whether the selection rate, true positive rate and false negative rate differed by gender; whether features acted as proxies for gender; whether the training data reflected a historically discriminatory baseline; and whether human reviewers were able and expected to detect systematic disparities. If the system rejected qualified women at a higher rate than qualified men, the relevant failure would be equality of opportunity and false negative control. If it produced a lower selection rate for women without objective justification, the relevant evidence would also include adverse impact.

A recruitment audit should therefore begin with a mapping of the employment decision. Is the tool used for sourcing candidates, ranking applications, filtering out candidates, recommending interview questions, predicting performance, allocating training or assessing promotion? Each use changes the harm. A sourcing tool may reproduce occupational segregation by showing opportunities to one group more than another. A screening tool may create hidden exclusion. A ranking tool may shape human attention even if it does not formally reject anyone. A performance prediction tool may influence pay, training or career progression.

The three minimum metrics in the workbook are deliberately chosen for that employment context. Disparate impact detects inequality of results. True positive rate difference detects inequality of opportunity among persons who meet the relevant criterion. False negative rate difference detects unequal erroneous exclusion. Together they create a compact employment fairness profile. They do not answer every question, but they prevent the most common evasion: claiming that a tool is fair because it is accurate overall while a protected group bears the cost of its mistakes.

The Amazon example also illustrates the limits of patching. Removing a single discriminatory token does not guarantee that the model will not learn other correlated signals. A robust mitigation plan would require a broader feature audit, balanced or reweighted training data, subgroup validation, counterfactual testing, documentation of trade-offs and post-deployment monitoring. In the language of the workbook, this is not a single control but a control system: data provenance, feature rationality, test-plan evidence, human oversight and complaint handling must operate together.

5.2. *Facial Recognition*

Facial recognition and facial analysis show why intersectional testing matters. Buolamwini and Gebru's Gender Shades study evaluated commercial gender classification systems and found substantially higher error rates for darker-skinned women than for lighter-skinned men.[22] NIST's Face Recognition Vendor Test on demographic effects likewise documented demographic differentials across many face recognition algorithms, while warning that results vary by algorithm, task and dataset.[23]

The relevant metric depends on the use case. In one-to-one verification, a false non-match may wrongly deny access to a person who is entitled to pass. In one-to-many identification, a false match may expose a person to investigation or coercive action. In either case, an average error rate conceals the legal problem if errors cluster by sex, skin tone, age or intersectional combinations. The evaluator must therefore test false match and false non-match rates by group, document whether the operational threshold was chosen with subgroup performance in mind, and assess whether the deployment context magnifies harm.

Under Artificial Intelligence Act, certain biometric identification systems and employment-related systems appear in Annex III as high-risk categories.[24] That classification does not resolve the fairness problem, but it changes the evidentiary burden: the provider must be prepared to show that data governance, testing, documentation and oversight are adequate for the intended

[21] J. Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*, in *Reuters*, 10 October 2018.

[22] J. Buolamwini and T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in *Proceedings of Machine Learning Research*, vol. 81, 2018, 77, especially 88-89.

[23] P.J. Grother, M.L. Ngan and K.K. Hanaoka, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*, Gaithersburg, MD, National Institute of Standards and Technology, 2019.

[24] Regulation (EU) 2024/1689, cit., Annex III, points 1, 4 and 5.

purpose. In high-stakes biometric contexts, a tolerance that may be acceptable for convenience access may be unacceptable for law-enforcement identification.

5.3. *Case-Law Summary: SyRI*

The SyRI judgment is not a judgment on Artificial Intelligence Act, but it is a significant European reference point for algorithmic risk governance. The District Court of The Hague reviewed Dutch legislation enabling the Systeem Risico Indicatie, a system used to detect possible social security, tax and labour-law fraud. The court held that the legislation violated Article 8 of the European Convention on Human Rights because it did not strike a fair balance between fraud prevention and private life, particularly given the system's insufficient transparency and verifiability.[25]

For the purposes of this article, SyRI is relevant for two reasons. First, it confirms that algorithmic assessment in the public sector is judged not only by accuracy but by legality, transparency, proportionality and safeguards. Secondly, it shows that opacity can itself be a rights problem. If affected persons, courts and supervisory bodies cannot understand the categories of data used, the risk logic applied and the safeguards available, then measurement cannot perform its accountability function.

SyRI also shows the importance of institutional explainability. The legal problem was not simply that the system might be inaccurate. It was that the persons affected and the court could not sufficiently assess how the system operated, which risk indicators were decisive, how data combinations were weighted and what safeguards constrained the administration. In discrimination analysis, such opacity is particularly damaging because disparate impact is often proved through patterns that require access to data, categories and decision logic.

For an Article 27 assessment, the lesson is clear: the deployer must be able to describe not only the tool but the administrative practice in which it is embedded. Who initiates the assessment? Who can see the score? Does the score trigger a decision, an investigation or only a recommendation? Can the affected person contest the result before harm occurs? Are complaint outcomes fed back into model monitoring? These questions transform fundamental rights impact assessment from a formal template into a meaningful safeguard.

6. *Measurement Under EU Regulatory Governance*

6.1. *Provider Duties: Data, Documentation and Risk Management*

Article 10 of Artificial Intelligence Act is the central provision for data and data governance in high-risk systems. It requires data governance practices concerning design choices, data collection, data preparation, assumptions, data availability, examination of possible biases, measures to detect, prevent and mitigate biases, and identification of data gaps. It also requires datasets to be relevant, sufficiently representative, and, to the best extent possible, free of errors and complete in view of the intended purpose.[26]

The legal importance of Article 10 is that it makes bias examination a compliance duty rather than a voluntary ethical practice. A provider must therefore be able to explain which groups were considered, which data gaps were identified, which metrics were selected, which thresholds were used, what mitigation was attempted and what residual risk remains. Annex IV documentation then becomes the institutional memory of the assessment: it should allow a notified body, market surveillance authority or court to reconstruct the reasoning.

Article 9 reinforces this duty by requiring a risk management system throughout the lifecycle. For bias, lifecycle means more than development. Drift, population change, new uses, changed thresholds, interface design, human reliance and feedback loops can all alter subgroup performance after deployment. A system that was acceptable at launch may become discriminatory if the operational population changes or if human operators learn to defer to the model without meaningful review.

6.2. *Deployer Duties and Fundamental Rights Impact Assessment*

Article 27 requires certain deployers of high-risk systems to perform a fundamental

[25] District Court of The Hague, NJCM c.s./The State of the Netherlands, judgment of 5 February 2020, ECLI:NL:RBDHA:2020:1878, paras. 6.86-6.95.

[26] Regulation (EU) 2024/1689, cit., art. 10(2)-(5).

rights impact assessment before first use. The assessment must describe the process in which the system will be used, the period and frequency of use, the categories of persons and groups likely to be affected, the specific risks of harm, human oversight measures and the measures to be taken if those risks materialise.[27] It complements, rather than replaces, data protection impact assessment under the General Data Protection Regulation.[28]

That reading is consistent with Simón Castellano's comparative analysis of fundamental rights impact assessment in the public sector, which presents the assessment as an *ex ante* governance instrument that extends beyond data protection impact assessment and must cover the full range of rights protected by the Charter of Fundamental Rights of the European Union.[29]

A deeper reading of Article 27 shows that it is not a generic ethics statement. It has six operational elements. First, the deployer must describe the process in which the system will be used. This prevents abstract assessments detached from the actual decision flow. Secondly, it must identify the period and frequency of use. A system used once in a pilot does not create the same accumulated risk as a system used daily to screen thousands of applicants. Thirdly, it must identify the categories of natural persons and groups likely to be affected. This is the legal opening through which protected groups, vulnerable groups and intersectional groups enter the assessment.

Fourthly, the deployer must identify specific risks of harm for those persons and groups, taking into account the information provided by the provider under Article 13. This makes the deployer dependent on provider transparency, but it does not allow the deployer to outsource responsibility. If local conditions reveal a risk not visible in provider documentation, the deployer must record it. Fifthly, the deployer must describe human oversight measures in accordance with the instructions for use. Sixthly, it must identify measures to be taken if risks materialise, including internal governance and complaint mechanisms.

These elements map naturally onto a discriminatory-impact assessment. The process description identifies where bias may affect the decision. The period and frequency help evaluate scale and cumulative harm. The affected-groups field defines the protected and vulnerable groups to be tested. The risk-of-harm field translates statistical disparity into legal consequence. Human oversight defines whether the automated output can be challenged or corrected. Governance and complaint mechanisms define whether affected persons have a route to remedy. In this sense, Article 27 is the legal home of the bias register, not an accessory to it.[30]

Recital 96 reinforces that reading by referring to the possible involvement of affected groups, independent experts and civil society organisations, and by mentioning complaint handling and redress procedures. That language matters. Bias assessment is not only a provider's laboratory exercise; it is a participatory and institutional process. In high-impact public-sector deployments, stakeholder involvement can reveal harms that are invisible in training data: disability-related barriers, linguistic exclusion, intersectional vulnerability, fear of contesting public authorities or indirect exclusion through digital channels.[31]

The relationship between Article 27 and data protection impact assessment should also be clarified. The General Data Protection Regulation asks whether processing personal data creates high risks for rights and freedoms. Article 27 asks whether the use of a high-risk artificial intelligence system affects fundamental rights in context. These assessments overlap, but they are not identical. A data protection impact assessment may focus on lawful basis, minimisation, accuracy, transparency and security. A fundamental rights impact assessment must

[27] Regulation (EU) 2024/1689, cit., art. 27.

[28] European Parliament and Council, Regulation (EU) 2016/679 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 (General Data Protection Regulation), in Official Journal of the European Union, L 119, 4 May 2016, arts. 5, 9, 22 and 35.

[29] P. Simón Castellano, *L'avaluació d'impacte en els drets fonamentals de l'ús d'intel·ligència artificial en el sector públic: models i metodologies en perspectiva comparada*, in *Revista Catalana de Dret Públic*, no. 71, 2025, 4, doi:10.58992/rcdp.i71.2025.4486.

[30] Regulation (EU) 2024/1689, cit., art. 27(1)(a)-(f), art. 27(2)-(5).

[31] Regulation (EU) 2024/1689, cit., recital 96, on stakeholder involvement, complaint handling and redress procedures in fundamental rights impact assessment.

also address equality, dignity, access to services, labour rights, good administration, effective remedy and the institutional consequences of automation.

For that reason, the Article 27 file should include at least four annexes when the system may create discriminatory impact. The first is a data and group coverage sheet. The second is a dynamic testing report with metrics by group. The third is a bias risk register linking evidence, probability, severity, controls and residual risk. The fourth is a control library showing which safeguards exist at each lifecycle phase. The tables below translate those annexes into a concise model that can be adapted to employment, public administration or other high-risk contexts.

The deployer perspective is essential because bias often arises from use rather than model architecture alone. A recruitment model may be lawful as a limited screening aid but unlawful if it becomes a de facto exclusion gate. A biometric system may be tested under laboratory conditions but deployed in lighting, camera-angle or crowd-density conditions that worsen subgroup error rates. A fundamental rights impact assessment should therefore include the actual decision flow, the authority of human reviewers, the complaint route, the possibility of correction and the consequences of an adverse decision.

6.2.1. *From Legislative Silence to the Article 27 Duty*

The present shape of Article 27 is best understood against its legislative history. The original Commission proposal did not place fundamental rights impact assessment at the centre of the regulatory model. The proposal certainly used the language of risk and contained several requirements designed to protect health, safety and fundamental rights indirectly, but it did not establish a horizontal duty for deployers to perform a specific fundamental rights impact assessment for high-risk artificial intelligence systems. Its closest reference was the relationship with data protection impact assessment, a much narrower instrument because it is triggered by personal-data processing and is structured around the rights and freedoms affected by that processing.

The European Parliament's amendments changed the normative centre of gravity. They sought to convert fundamental rights impact assessment into a general precondition for the use of high-risk systems. That broader proposal was normatively coherent: if the Regulation classifies a system as high risk because it can affect access to education, employment, welfare, justice, migration or essential services, then the deployer should be required to assess, in its own context, how that system may affect the persons subject to it. The final text, however, is more limited. It institutionalises the assessment but restricts the circle of deployers obliged to carry it out. This compromise explains why Article 27 is both a breakthrough and an incomplete instrument.[32]

The breakthrough is that fundamental rights impact assessment is no longer merely an ethical recommendation. It is a legal duty for covered deployers before first use. The incompleteness is that many private-sector uses of high-risk systems remain outside the formal Article 27 obligation, even though they may produce comparable effects on work, access to services or social participation. From the perspective of algorithmic bias, this means that the strongest legal framework for contextual rights assessment is not coextensive with the full map of possible discriminatory harms. The result is an asymmetrical architecture: provider duties apply broadly to high-risk systems, while deployer-side fundamental rights assessment applies more selectively.

This asymmetry is important for the article's main thesis. Bias measurement by a provider can identify general performance differentials, but it cannot exhaust the rights analysis because rights are affected in context. A model may be technically the same and legally different depending on who deploys it, for what purpose, over which population, with what human oversight and with what remedy. Article 27 is valuable precisely because it asks the deployer to move from generic system properties to situated fundamental rights effects. Its limitation is that the law does not require that situated assessment in every high-risk private deployment.

6.2.2. *Subjective and Objective Scope of the Duty*

Article 27 combines a subjective and an objective trigger. The first trigger is

[32] P. Simón Castellano, *L'avaluació d'impacte en els drets fonamentals de l'ús d'intel·ligència artificial en el sector públic*, cit., 4-24.

institutional: bodies governed by public law must carry out the assessment when they deploy high-risk systems. The same logic applies to private entities providing public services, because they exercise functions that affect citizens in contexts of general interest. The second trigger is sectoral: deployers using high-risk systems for creditworthiness assessment of natural persons, or for risk assessment and pricing in life and health insurance, are covered irrespective of whether they are public or private. The legal intuition is clear. The duty follows areas where artificial intelligence can affect basic social access, not merely commercial preference.[33]

The public-sector emphasis is particularly significant. When a municipality uses an algorithm to prioritise social assistance, when a public employment service ranks candidates for training, when a health administration introduces a triage or waiting-list tool, or when a migration authority uses a risk score to allocate scrutiny, the system is not merely optimising an internal workflow. It is structuring the exercise of public power. A fundamental rights impact assessment then becomes part of administrative legality: it is connected to reason-giving, transparency, proportionality, equality, good administration and effective remedy.

6.2.3. *Minimum Content: From Template to Reasoned File*

Article 27's minimum content can be read as a sequence of legal questions. The description of the process asks: where, exactly, will artificial intelligence enter the decision chain? The period and frequency of use ask: how often will the system affect persons and for how long? The categories of affected persons ask: whose rights may be implicated, including vulnerable and intersectional groups? The identification of specific risks asks: which rights may be harmed and through which causal pathway? The assessment of provider information asks: what does the deployer know about system performance, limitations and technical safeguards? The governance measures ask: who will supervise, correct, explain and remedy?

This sequence should prevent the assessment from degenerating into a generic form. A meaningful file should contain a map of the administrative or organisational process, a description of the decision points, the intended purpose of the system, the population affected, the protected attributes or proxies to be considered, the evidence used, the metrics chosen, the tolerance thresholds, the risk rating, the controls applied and the residual-risk decision. It should also identify the person or body authorised to stop or redesign the project when risk becomes unacceptable.

The reference to first use should not be understood as a permission to freeze the assessment. The Regulation allows previous assessments to be reused in similar cases, but similarity must be interpreted strictly. A deployment is not similar merely because the product is the same. The relevant similarity concerns purpose, population, institutional context, data, thresholds, human oversight, consequences and remedies. If any of those factors changes materially, the assessment should be updated. Otherwise the reuse clause can turn into a vehicle for cosmetic compliance: an old assessment follows the technology while the real rights context changes.

A deployer-side assessment should also interact with Article 26(9), which connects information gathered for fundamental rights impact assessment with compliance with data protection impact assessment where applicable. The point is not to duplicate paperwork. The point is to integrate evidence. A data protection impact assessment can address lawful basis, minimisation, transparency, accuracy, security and rights of data subjects. The fundamental rights assessment adds equality, dignity, access to public services, non-discrimination, good administration, effective remedy and the distributional effects of automation. The best practice is therefore a single governance dossier with distinct analytical layers.

6.2.4. *Comparative Methodological Lessons*

Recent comparative practice points toward a hybrid model. The Catalan approach, as analysed in the Simón Castellano's article, is especially strong in legal granularity and rights traceability. It works with risk matrices, mitigation plans and a structured relation between probability, severity and action. The Dutch AIIA model is strong in governance by phases: it separates necessity, public values

[33] Regulation (EU) 2024/1689, cit., art. 27(1)-(5), and art. 26(9).

and proportionality from more technical questions of design, robustness and monitoring. The Italian approach is strong in institutional integration: it seeks to coordinate the technical-organisational profile, the fundamental rights impact assessment and the data protection impact assessment within public procurement and lifecycle management.[34]

Model	**Main contribution**	**Usefulness for algorithmic bias**
Catalan model	Rights-risk matrix, legal traceability, mitigation plan and proportionality reasoning.	Useful for linking bias metrics to concrete fundamental rights and to an action plan.
Dutch AIIA model	Stage-gates, necessity and proportionality analysis, public values and technical lifecycle controls.	Useful for deciding whether to continue, redesign, restrict or stop a use case at each phase.
Italian AgID approach	Integrated dossier connecting technical profile, fundamental rights assessment and data protection assessment.	Useful for public procurement, organisational ownership and avoiding duplicated assessments.
Hybrid proposal	Rights matrix plus stage-gates plus integrated governance file.	Best fit for Article 27 because it is rigorous, decision-oriented and administratively usable.

Table 6. Comparative lessons for a rights-sensitive bias assessment

The hybrid model is particularly apt for bias assessment. The rights-risk matrix prevents the analysis from becoming a purely technical validation exercise. Stage-gates prevent inertia: after each phase, the responsible body must decide whether to continue, add safeguards, redesign or stop. The integrated dossier prevents fragmentation between legal, technical, procurement and data protection teams. In practice, the Article 27 file should not be a separate island. It should travel with procurement documents, provider documentation, data protection analysis, test reports, control registers and monitoring records.

6.2.5. *Avoiding Cosmetic Compliance*

The main risk of Article 27 is not that organisations will refuse to create a document. The main risk is that they will create a document that does not change any decision. Cosmetic compliance can take several forms: copying the provider's assessment without local analysis; applying a generic checklist without rights reasoning; defining thresholds after seeing the results; omitting affected groups because data are inconvenient; treating human oversight as a label rather than a practice; or archiving the assessment without linking it to procurement, monitoring and complaint handling.

A robust assessment therefore needs institutional teeth. It should identify who owns the file, who validates the methodology, who approves residual risk, who can suspend deployment, who receives complaints, who reviews drift, and who reports to senior management or political leadership. Where the system affects public services, it should also identify what information can be published without compromising security, trade secrets or personal data. Transparency is not an ornament. It creates incentives for seriousness and allows affected persons, civil society, parliamentary bodies, ombuds institutions and courts to examine the quality of the assessment.

The most demanding version of Article 27 is therefore cultural rather than procedural. It requires public administrations and service providers to move from a project mentality to a lifecycle mentality. A technological project does not end when a system goes live. It continues through monitoring, complaints, model updates, changes in population, changes in law, changes in organisational use and eventual retirement. Bias assessment is a continuous governance function, not a launch condition that expires once the tool is deployed.

6.3. *Special Categories of Data and the Paradox of Bias Testing*

Bias testing often requires information about protected attributes. This creates a paradox: data minimisation and sensitivity rules discourage processing of such data, but non-discrimination analysis may require it.

[34] P. Simón Castellano, *L'avaluació d'impacte en els drets fonamentals de l'ús d'intel·ligència artificial en el sector públic*, cit., 12-18.

Artificial Intelligence Act addresses this by allowing providers of high-risk systems, under strict conditions, to process special categories of personal data when strictly necessary for bias detection and correction. That permission is narrow. It requires safeguards, restrictions on reuse, security measures, access controls, deletion and documentation of necessity.

The normative point is that equality cannot be audited in the dark. If an organisation refuses to collect or process any group-related information, it may be unable to know whether its system disadvantages protected groups. But the answer is not indiscriminate collection. The answer is a controlled testing environment, lawful basis analysis, privacy-preserving methods where possible, aggregation, role-based access, retention limits and clear separation between bias testing data and operational decision data.

7. *Treatment Measures and Standardisation*

Once bias is detected, treatment can occur at three levels. Pre-processing measures modify data: rebalancing, enrichment, better labelling, removal of low-quality records, careful imputation, stratified sampling or additional data collection. In-processing measures modify the model: fairness constraints, regularisation, adversarial approaches, cost-sensitive learning or group-aware thresholding. Post-processing measures modify outputs: threshold adjustment, calibrated decision rules, human review triggers or abstention where the model is uncertain.

The choice among these techniques is not purely technical. Rebalancing may improve representation but distort prevalence. Threshold adjustment may equalise errors but create different treatment rules. Removing proxies may reduce discrimination but also remove genuinely relevant information. Human review may correct errors but can introduce automation bias or inconsistent discretion. The mitigation decision should therefore be documented as a trade-off, with reasons, affected interests and residual risks.

Standardisation is both necessary and risky. It is necessary because providers and deployers need reproducible methods, common documentation, test-plan expectations and benchmark language. It is risky because fairness is not a physical property like voltage or tensile strength. A standard that hard-codes a single metric or tolerance may create a misleading impression of legal safe harbour. The better approach is process-based standardisation: identify stakeholders, define intended purpose, document data provenance, select justified metrics, test by group and intersection, set acceptance criteria before evaluation, mitigate, monitor and re-evaluate.[35]

Lifecycle stage	**Control**	**Methodo-logical/ legal anchor**
Inception	Stakeholder and affected-group identification, including intersectional groups	AESIA Annex C.1; ISO/IEC TS 12791 clause 5.1
Inception	Predefined anti-bias requirements and acceptance criteria	Article 10 AI Act; ISO/IEC TS 12791 clause 5.1.6
Design and development	Data provenance, representa-tiveness and proxy review	Article 10 AI Act; ISO/IEC TS 12791 clauses 5.1.4 and 5.2
Verification and validation	Static data testing and dynamic model/system testing by group	AESIA Annex C.2; ISO/IEC TS 12791 clause 5.3
Mitigation	Documented data, model and post-processing treatments with trade-off analysis	AESIA Annex C.3; ISO/IEC TS 12791 clause 6
Operation	Logging, drift monitoring, complaint channel and periodic reassessment	Articles 9, 26 and 27 AI Act; ISO/IEC TS 12791 clause 5.4

[35] See also H. Janssen, S.A. Lee and J. Singh, *Practical fundamental rights impact assessments,* in *International Journal of Law and Information Technology*, vol. 30, no. 2, 2022, 200, for a practice-oriented methodology; and B.C. Stahl et al., *A systematic review of artificial intelligence impact assessments*, in *Artificial Intelligence Review*, vol. 56, 2023, for a comprehensive review of AI-specific impact assessment frameworks.

Lifecycle stage	Control	Methodo-logical/ legal anchor
Retirement	Retention or deletion rules for data and documentation after decommissioning	ISO/IEC TS 12791 clause 5.5

Table 7. Lifecycle controls for bias measurement and treatment

This is where UNE-CEN/CLC ISO/IEC/TS 12791:2024 can be especially useful. Its focus on unwanted bias in classification and regression tasks fits the type of systems most likely to appear in employment and eligibility decisions. Its lifecycle clauses can be mapped directly onto a legal governance model: inception, design and development, verification and validation, operation and monitoring, and retirement. The standard does not answer the equality question by itself, but it supplies a disciplined vocabulary for asking it.

The following control library expands the shorter lifecycle table used above. Its purpose is not to present twenty bureaucratic boxes, but to show how discriminatory-impact assessment can be made operational. Each control answers a recurrent failure mode: no one identified the affected groups; the organisation tested only global accuracy; the provider refused to disclose bias-relevant information; the model used proxies; the test data were not separate from the training data; human reviewers rubber-stamped outputs; complaints were not connected to model monitoring; or the system was retired without preserving evidence.

ID	Measure / safeguard	Type	Lifecycle phase	Bias-control function
CTRL01	Stakeholder and affected-group identification, including intersectional groups, with recorded feedback.	Governance	Inception	Defines who must be protected and consulted.
CTRL02	Approval of anti-bias require-ments, legal require-ments and internal policies with traceability.	Governance	Inception	Creates a documented benchmark before testing.
CTRL03	Documen-tation of data provenance, selection criteria and representa-tiveness by subgroup.	Data	Inception	Identifies coverage and selection bias.
CTRL04	Data quality by group: diversity, missingness, consistency and plan for rebal-ancing or additional data.	Data	Verification and validation	Prevents hidden subgroup degradation.
CTRL05	Quantitative acceptance criteria, including group thresholds, defined and justified before evaluation.	Governance	Inception	Prevents ex post metric shopping.
CTRL06	Procure-ment clauses requiring trans-parency on bias-relevant aspects and coopera-tion in audits.	Procurement	Inception	Reduces vendor opacity.
CTRL07	Feature review to detect demographic proxies, with documented rationale and alternatives.	Model	Design and development	Controls indirect discrimi-nation through variables.
CTRL08	Labelling controls: clear instructions, ground-truth samples, checks, annotator agreement and dispute resolution.	Data	Design and development	Reduces labelling bias.
CTRL09	Test Plan and Test Completion Report including metrics, justification and monitor-ing criteria.	Governance	Verification and validation	Creates an auditable testing file.
CTRL10	Dynamic tests with separate, representative data and metrics disaggregated by group.	Testing	Verification and validation	Measures output bias directly.
CTRL11	Extreme-input tests by group and robustness review for vulnerability-related bias.	Testing	Verification and validation	Detects failures at distribution edges.
CTRL12	Production monitor-ing: logging, drift detection, periodic re-evaluation and effectiveness review of controls.	Operation	Operation	Keeps assess-ment alive after deploy-ment.
CTRL13	Meaningful human oversight and escalation procedure; training against automation bias.	Operation	Operation	Prevents formal or rubber-stamp oversight.
CTRL14	Complaint, review and rectification channel for adverse outcomes, with incident logging.	Governance	Operation	Connects affected persons to correction.
CTRL15	Documented trade-offs between perfor-mance and equality; safeguards against metric gaming.	Model	Mitigation	Makes fairness choices explicit.

ID	Measure / safeguard	Type	Lifecycle phase	Bias-control function
CTRL16	Data mitigation: up-sampling, down-sampling, augmentation, additional data and feedback-loop prevention.	Data	Mitigation	Treats representational sources of bias.
CTRL17	Model mitigation: regularisation, constraints, dual objectives and decoupled classifiers where justified.	Model	Mitigation	Treats model-level disparity.
CTRL18	Periodic discriminatory-impact assess-ment and internal or external trans-parency.	Gover-nance	Operation	Supports accounta-bility and review.
CTRL19	Retention or deletion management for data and documen-tation after retirement.	Data	Retirement	Preserves evidence and manages residual risk.
CTRL20	Independent audit or external review where risk is high, with findings integrated into improvement.	Gover-nance	Verification and validation	Adds assurance for serious rights risks.

Table 8. Detailed library of controls for discriminatory-impact assessment

The control library is intentionally modular. Not every control will have the same intensity in every system. A low-impact internal triage tool may not require the same external audit as a public-sector fraud detection system. But the burden should be reversed for high-risk uses: if a control is absent, the assessment should explain why it is unnecessary, disproportionate or replaced by an equivalent safeguard. Silence should not count as compliance.

Testing documentation should be equally structured. ISO/IEC TS 12791 points toward testing and treatment across the lifecycle, while ISO/IEC/IEEE 29119-3 supplies the vocabulary of Test Plans and Test Completion Reports. For bias, the test plan should identify the protected attributes, the reference and comparison groups, the independent test dataset, the operating conditions simulated, the metrics selected and the criteria for repeating the test. The completion report should record not only results but also interpretation and action.[36][37]

[36] AESIA, Guía de datos y gobernanza de datos, cit., Annex C.2-C.3, 60-63.

[37] ISO/IEC/IEEE, ISO/IEC/IEEE 29119-3:2021. Software and systems engineering - Software testing - Part 3: Test documentation, Geneva, ISO, 2021.

Block	Field / metric	What it detects	Gover-nance output
Test refe-rence	Test Plan ID; Test Completion Report ID; version of model and dataset.	Shows audita-bility and re-peatability.	Documented be-fore de-ploy-ment or material change.
Data separa-tion	Confirmation that test data are inde-pendent of training data.	Avoids opti-mistic va-li-dation and leakage.	Required unless justified and risk-accepted.
Operating conditions	Geography, lan-guage, population, threshold, interface and human-review setting.	Connects model testing to de-ploy-ment context.	Must match in-tended use or explain differ-ences.
Minimum employment metrics	Disparate Impact, TPR Difference and FNR Differ-ence.	Detects une-qual results, unequal op-portunity and unequal ex-clusion errors.	Compared with predefined tole-rances.
Classifica-tion log	Attribute, reference group, compared group, metric, ref-erence value, com-pared value, differ-ence or ratio, threshold, com-pli-ance result and action.	Creates the core evidence table for bina-ry or mul-ticlass deci-sions.	Action required when threshold is breached or sample is insuf-ficient.
Regression log	Attribute, groups, MAE or RMSE by group, ratio or dif-ference, threshold, interpretation and action.	Extends bias testing to scores, rank-ings and pre-dictions.	Residual error by group must be justified.
Monitoring criteria	Triggers for re-testing: drift, com-plaints, model up-date, new popula-tion or threshold change.	Prevents one-off compli-ance.	Linked to opera-tion controls and complaint channels.

Table 9. Dynamic testing record for output bias

Data field	Evidence to record	Bias question	Decision output
Dataset identifi-cation	Train, vali-dation, test or production dataset; version and date.	Which evi-dence source is being as-sessed?	Versioned da-taset register.

Data field	Evidence to record	Bias question	Decision output
Origin and provider	Internal source, vendor, public dataset or mixed source.	Who controls data quality and disclosure?	Provider contract or provenance record.
Sensitive variables and proxies	Direct attributes and variables correlated with protected grounds.	Could the system learn protected status indirectly?	Proxy review and feature rationale.
Representativeness	Coverage of the target population and relevant subgroups.	Is a protected group underrepresented or absent?	Rebal-ancing, enrichment or use limitation.
Feature and label imbalance	Distribution of inputs and outcomes by group.	Does historical inequality appear in labels or features?	Label review or mitigation plan.
Missingness and corruption	Missing, inconsistent or corrupted values by group.	Does one group suffer poorer data quality?	Cleaning, imputation or exclusion decision.
Diversity and intersec-tionality	Counts and percentages by protected and intersectional groups.	Are compounded vulnerabilities visible?	Minimum sample assess-ment and uncertainty note.
Actions and decisions	Accepted, mitigated, escalated, retested or stopped.	What happened after the finding?	Residual-risk decision.

Table 10. Static data sheet for provenance, quality and group coverage

A final register links the preceding tables. The register should not be confused with the technical report. Its function is managerial and legal. It converts a finding into a risk entry, assigns a responsible person, connects the finding to controls and records residual risk. In a later dispute, this register will often be more important than a long technical appendix, because it shows how the organisation governed the evidence it had.

Register field	Content	Legal function
ID	Unique identifier for the bias risk.	Allows traceability between tests, controls and decisions.
Bias description	Short description of the disparity, proxy or process failure.	States the risk in legal and technical terms.
Lifecycle phase	Inception, design, verifi-cation, opera-tion, mitigation or retirement.	Shows where the risk entered or was detected.
Protected attribute and affected group	Protected ground, comparison group and intersectional category where relevant.	Connects metric to equality law.
Evidence	Metric, data-quality finding, complaint, audit result or expert review.	Identifies the factual basis.
Probability and severity	Low, medium or high, with justification.	Produces inherent risk.
Action required	No action, monitor, mitigate, restrict, suspend or stop.	Transforms evidence into gover-nance.
Controls and vulnera-bility	Relevant CTRL identifiers and control effectiveness.	Shows whether safeguards reduce risk.
Responsible person and date	Owner, approval body and review date.	Creates accounta-bility.
Residual risk	Remaining risk after controls and mitigation.	Records the deploy-or-not decision.

Table 11. Bias risk register model

The four operational instruments - data sheet, dynamic test log, control library and bias register - also make the article more citable because they translate a doctrinal claim into a replicable method. A reader can disagree with a specific tolerance, but still use the structure. That is the point. The method is not a universal answer to fairness. It is a disciplined way of asking the right questions in a form that legal actors can inspect.

The method also helps reconcile the language of technical standards with the language of rights. Technical standards tend to ask whether a process is documented, reproducible and controlled. Fundamental rights law asks whether persons are protected against unjustified interferences and unequal treatment. The proposed assessment sits between those languages. It uses standards to make the process reproducible, and rights to decide what the process is for.

This point is especially important for public administrations. Public-sector artificial intelligence is rarely a private transaction between provider and user. It is an exercise of power. Where the system ranks, flags, screens or predicts persons in access to benefits, services, inspections or sanctions, the assessment must also address good administration, reason-giving, contestability and effective remedy. Bias measurement is therefore part of administrative legality, not merely digital governance.

The same reasoning applies to employment, although through a different institutional lens. In recruitment and workplace analytics, affected persons may not know that an automated tool has shaped the opportunity structure. They may never see the rejected application, the ranking score or the proxy that reduced their chances. A discriminatory-impact assessment therefore has a preventive function: it creates a record before individual proof becomes impossible. In this sense, measurement protects not only equality but evidentiary fairness.

7.1. *Implementation Challenges in the Public Sector*

The methodological proposal would remain incomplete without an account of implementation. Public administrations are not laboratories with unlimited technical capacity. Many local and regional bodies lack stable teams combining data science, constitutional law, equality law, procurement, ethics, cybersecurity and sector-specific expertise. Yet a serious assessment requires all of those languages. It must understand the model architecture, the data, the administrative procedure, the affected rights, the practical consequences of an error and the institutional channels for correction.

This creates a capacity dilemma. External consultants can provide immediate expertise, but excessive dependence on them weakens public control and institutional learning. Internal capacity is normatively preferable, but it takes time, training and resources. A realistic strategy is mixed: specialised external support for complex cases, combined with internal teams capable of owning the assessment, asking the right questions, understanding the evidence and making the final decision. The assessment should never become a vendor document with a public logo attached.

The second challenge is organisational integration. The assessment must occur early enough to matter. If it is carried out after procurement, after technical design or after political commitment to deployment, it will be pressured to validate a decision already taken. The better model places fundamental rights impact assessment at project inception and procurement design. It should shape specifications, transparency clauses, data-access obligations, audit rights, acceptance criteria, human oversight design and exit clauses. In high-risk systems, no procurement file should treat bias evidence as an optional annex.

The third challenge is decision consequence. What happens if the assessment identifies unacceptable discriminatory risk? A credible methodology must make room for negative outcomes: non-deployment, redesign, narrower use, stronger safeguards, additional data collection, independent audit or suspension. If every assessment ends in deployment, the assessment is not an assessment but a ritual. The governance body approving the file should therefore record why the residual risk is acceptable and what would trigger reconsideration.

The fourth challenge is supervision. Article 27 requires notification of results to the market surveillance authority through a completed template. That notification should not become a mere registry act. The authority should be able, at least in high-impact cases, to examine methodology, evidence, thresholds and controls. Public-sector supervision may need a dialogic form: not only punishment, but iterative improvement, recommendations, requests for clarification and, where necessary, orders to mitigate or stop. This is particularly important where monetary sanctions are not the central accountability tool.

The fifth challenge is participation. Algorithmic bias is often visible to affected communities before it is visible in an institutional dashboard. Persons with disabilities, linguistic minorities, racialised communities, older persons, welfare recipients or jobseekers may perceive exclusion patterns that designers did not anticipate. Participation can take several forms: stakeholder interviews, targeted consultation, publication of executive summaries, civil society review, workshops, ombuds involvement or complaint analytics. Participation should be

proportionate, but in sensitive public-sector uses it should be treated as a source of evidence, not merely as democratic decoration.[38]

8. *Conclusion*

Inequality in algorithmic systems can be measured, but not in the simplistic sense that one metric can settle the legal question. Measurement can reveal whether a system distributes outcomes or errors unevenly across groups. It can identify whether a recruitment model excludes qualified women, whether a biometric system misidentifies darker-skinned persons at higher rates, whether a public-sector risk model concentrates burdens in vulnerable neighbourhoods, or whether an apparently neutral feature acts as a proxy for protected status.

Yet measurement is not adjudication. A disparity must be interpreted through the purpose of the system, the protected interest, the decision context, the availability of alternatives, the seriousness of harm and the safeguards in place. Artificial Intelligence Act is best read in this light. It does not constitutionalise a fairness metric. It creates an evidentiary and governance architecture in which bias must be examined, mitigated, documented and monitored.

The methodological contribution of this article is therefore practical. Define tolerances before testing. Measure global performance and subgroup performance separately. Use adverse impact to detect unequal outcomes, true positive rate differences to detect unequal opportunity, false negative rate differences to detect exclusion harms, and false positive rate differences to detect unjustified burdens or preferential errors. Require statistical robustness. Record residual risk. Link each metric to a control. Repeat the assessment when data, population, model or use changes.

The answer to the title is thus a qualified yes. Inequality can be measured when measurement is treated as structured evidence within a broader legal and ethical process. It cannot be measured responsibly when the evaluator seeks a single number that converts contested values into automatic compliance. The future of trustworthy artificial intelligence in Europe will depend on keeping both truths in view: without metrics, bias remains invisible; without law and judgment, metrics become a false comfort.

Five final implications follow. First, methodological guidance is not a secondary matter. Artificial Intelligence Act creates duties, but the effectiveness of those duties depends on models, templates, standards and institutional routines capable of translating legal requirements into operational evidence. Without a common methodological floor, Article 27 risks uneven application. With an excessively rigid model, it risks formalism. The desirable path is standardisation with room for contextual justification.

Secondly, discriminatory-impact assessment must be multidimensional. Privacy, equality, dignity, access to services, good administration, labour rights and effective remedy can be affected at the same time. A data protection impact assessment is indispensable where personal data are processed, but it cannot absorb the whole rights analysis. Fundamental rights impact assessment is broader because it asks how an automated system changes the distribution of power, opportunity, explanation and contestability.

Thirdly, cosmetic compliance is the enemy of trustworthy artificial intelligence. A beautifully formatted assessment that does not influence procurement, design, deployment, monitoring or redress is worse than useless because it creates a false appearance of diligence. The assessment must be connected to an internal algorithmic governance system: authorised-use registers, risk committees, audit trails, complaint channels, human oversight protocols, monitoring dashboards and escalation rules.

Fourthly, the assessment must remain adaptive. Artificial intelligence systems change, populations change, thresholds change and institutional uses change. A finding of acceptable residual risk at first use cannot be treated as permanent. Bias assessment should be repeated when the model is updated, when the dataset changes, when the intended purpose is extended, when complaints accumulate, when subgroup performance drifts or when the legal context changes. Continuous improvement is not a managerial slogan; it is the only way to keep the assessment aligned with reality.

Fifthly, Article 27 can catalyse a new administrative culture. The goal is not to

[38] P. Simón Castellano, *L'avaluació d'impacte en els drets fonamentals de l'ús d'intel·ligència artificial en el sector públic*, cit., 18-23.

paralyse innovation but to change the order of questions. The old logic asks whether a technological tool works and then asks, sometimes too late, whether it harms rights. The new logic asks from the beginning whether the tool is necessary, proportionate, explainable, contestable and fair enough for the public or social function it will perform. That shift is the true promise of fundamental rights impact assessment.[39]

For the specific problem of algorithmic bias, the conclusion is therefore sharper than a generic call for ethics. Bias must be measured, but it must be measured inside a rights file. The metric without the file is blind to legal meaning; the file without the metric is blind to empirical reality. The most defensible model is one in which statistical evidence, lifecycle controls and fundamental rights reasoning discipline each other.

This is why the proposed formula should guide the article as a whole: algorithmic inequality is measurable as evidence, but not decidable as compliance by a metric alone. The sentence is intentionally compact because it captures the central legal lesson. Numbers are necessary to reveal inequality, compare alternatives and monitor improvement. Law is necessary to decide what those numbers mean, which harms matter, which justifications are acceptable and which residual risks cannot be imposed on affected persons.

[39] P. Simón Castellano, *L'avaluació d'impacte en els drets fonamentals de l'ús d'intel·ligència artificial en el sector públic,* cit., 23-24.