



AN EQUINET
HANDBOOK

How to use the Artificial Intelligence Act to investigate AI bias and discrimination: A guide for Equality Bodies

by Brent Mittelstadt



Federal
Anti-Discrimination
Agency

2024

How to use the Artificial Intelligence Act to investigate AI bias and discrimination: A guide for Equality Bodies is published by Equinet, European Network of Equality Bodies. Equinet brings together 47 organisations from across Europe which are empowered to counteract discrimination as national Equality Bodies across the range of grounds including age, disability, gender, race or ethnic origin, religion or belief, and sexual orientation.

Equinet members: Commissioner for the Protection from Discrimination, **Albania** | Austrian Disability Ombudsperson, **Austria** | Ombud for Equal Treatment, **Austria** | Unia (Interfederal Centre for Equal Opportunities), **Belgium** | Institute for the Equality of Women and Men, **Belgium** | Institution of Human Rights Ombudsman of Bosnia and Herzegovina, **Bosnia and Herzegovina** | Commission for Protection against Discrimination, **Bulgaria** | Ombudswoman of the Republic of Croatia, **Croatia** | Gender Equality Ombudsperson, **Croatia** | Ombudsman for Persons with Disabilities, **Croatia** | Office of the Commissioner for Administration and the Protection of Human Rights, **Cyprus** | Public Defender of Rights, **Czech Republic** | Danish Institute for Human Rights, **Denmark** | Gender Equality and Equal Treatment Commissioner, **Estonia** | Ombudsman for Equality, **Finland** | Non-Discrimination Ombudsman, **Finland** | Defender of Rights, **France** | Public Defender (Ombudsman) of Georgia, **Georgia** | Federal Anti-Discrimination Agency, **Germany** | Greek Ombudsman, **Greece** | Office of the Commissioner for Fundamental Rights, **Hungary** | Irish Human Rights and Equality Commission, **Ireland** | National Office Against Racial Discrimination, **Italy** | Ombudsperson Institution, **Kosovo*** | Ombudsman's Office of the Republic of Latvia, **Latvia** | Office of the Equal Opportunities Ombudsperson, **Lithuania** | Centre for Equal Treatment, **Luxembourg** | National Commission for the Promotion of Equality, **Malta** | Commission for the Rights of Persons with Disability, **Malta** | Equality Council, **Moldova** | Protector of Human Rights and Freedoms (Ombudsman), **Montenegro** | Netherlands Institute for Human Rights, **Netherlands** | Commission for Prevention and Protection against Discrimination, **North Macedonia** | Equality and Anti-Discrimination Ombud, **Norway** | Commissioner for Human Rights of the Republic of Poland, **Poland** | Commission for Citizenship and Gender Equality, **Portugal** | Commission for Equality in Labour and Employment, **Portugal** | National Council for Combating Discrimination, **Romania** | Commissioner for Protection of Equality, **Serbia** | Slovak National Centre for Human Rights, **Slovakia** | Advocate of the Principle of Equality, **Slovenia** | Council for the Elimination of Ethnic or Racial Discrimination, **Spain** | Institute of Women, **Spain** | Equality Ombudsman, **Sweden** | Ukrainian Parliament Commissioner for Human Rights, **Ukraine** | Equality and Human Rights Commission, **United Kingdom – Great Britain** | Equality Commission for Northern Ireland, **United Kingdom – Northern Ireland**

*This designation is without prejudice to positions on status and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo declaration of independence.

Equinet Secretariat | Place Victor Horta, 40 | 1060 Brussels | Belgium |
info@equineteurope.org | www.equineteurope.org

ISBN 978-92-95112-87-2

© Equinet 2024 - Reproduction is permitted provided the source is acknowledged.

This is a publication commissioned by Equinet and financed by Germany's Federal Anti-Discrimination Agency (FADA). The veracity of the information provided is the responsibility of the author and other contributors. Equinet is co-funded by the European Union. The views expressed in this publication belong to the author(s), and neither Equinet, FADA, nor the European Commission are liable for any use that may be made of the information contained therein. This information does not necessarily reflect the position or opinion of individual members of Equinet, FADA, or the European Commission.

Acknowledgements

Author

Brent Mittelstadt (University of Oxford)

Editorial and Publication Coordination

Milla Vidina (Equinet, European Network of Equality Bodies)

Formatting

Louis De Visscher (Kreora Communication)



Contents

1. Introduction	5
2. Standards and the AIA	8
2.1. Harmonised standards	10
2.2. Will standards decide when AI is too biased or discriminatory?	13
3. Bias in the AIA	15
3.1. Unwanted bias and fairness	18
4. Powers of Equality Bodies under the AIA	22
4.1. Technical documentation for high-risk AI systems	25
4.2. Technical documentation for general-purpose AI models	43
4.3. Fundamental rights impact assessments	45
4.4. Post-market surveillance	47
4.5. Risk management system	47
5. Open challenges for Equality Bodies	49
5.1. Ambiguity in thresholds	50
5.2. Using statistical evidence in legal cases	51
5.3. Algorithmic discrimination is unintuitive and remote	52
5.4. Discrimination against new, unprotected groups	53
5.5. Gaps in data needed to measure bias and fairness	54
5.6. Aligning fairness measures with legal foundations	55
6. Conclusion and key recommendations	56



1. Introduction

How to use the Artificial Intelligence Act to investigate AI bias and discrimination: A guide for Equality Bodies



Bias, discrimination, and other risks to fundamental rights, health, and safety posed by emerging AI systems have increasingly become a priority in recent years for EU policymakers. In August 2024 the first international framework dedicated to the regulation of AI, the Artificial Intelligence Act (AIA), came into force.

The AIA is part of the EU’s overhaul of product safety regulation called the New Legislative Framework (NLF). The NLF requires manufacturers to self-assess, in some cases with third-party oversight by notified bodies, whether their products meet essential legal requirements before they enter the market. Following this, self-assessment is the primary way that AI systems with implications for fundamental rights risks will be assessed under the AIA. This is a weaker form of oversight than third-party or external assessment by a regulatory body. As a result, it is all the more important that Equality Bodies are able to play a meaningful role in translating the AIA into practice, for example by exercising their right to access self-assessment and technical documentation and raising any concerns with relevant national authorities. The main objective of this Guide is to explain how they can play this role effectively.

The AIA is a risk-based regulatory framework. This means that the Act requires providers of AI systems to assess the level of risks their systems pose to health, safety, and fundamental rights before placing them on the market.¹ Systems can be classified as (1) low- or no-risk, (2) high-risk, or (3) unacceptable risk. Each classification carries different legal requirements, with the majority focusing on high-risk systems.

The AIA’s inclusion of fundamental rights risks alongside health and safety makes it a unique framework in the context of product safety regulation. Relevant fundamental rights include democracy, the rule of law, environmental protection, and non-discrimination.² Equality bodies thus have a critical role to play to ensure AI providers and deployers are fully and fairly accounting for the risks their systems pose to the right to non-discrimination and taking appropriate steps to mitigate them in practice.

Providers of high-risk systems must fulfil a variety of “essential requirements” concerning risk assessment, data governance, technical documentation and records, transparency, human oversight, accuracy, cybersecurity, and robustness.³ Additionally, high-risk systems must be registered in a public database and subject to post-market monitoring, and certain application areas require fundamental rights impact assessments. Notably, in alignment with the NLF, AI providers will be presumed to fulfil these “essential requirements” if they implement harmonised technical standards (more on this below).⁴

¹ Alessandro Mantelero, ‘The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, Legal Obligations and Key Elements for a Model Template’ (2024) 54 Computer Law & Security Review 106020, 2.

² AI Act Article 1(1), Article 77(1).

³ Comparatively speaking, deployers of high-risk systems have much lower obligations. They have duties to ensure human oversight, recordkeeping, monitoring, and likewise must conduct fundamental rights impact assessments in certain cases (see: Section 4.3).

⁴ This is formally called a “presumption of conformity.”



Through these essential requirements and harmonised standards, the AIA provides new opportunities and tools to help Equality Bodies identify and investigate bias and discrimination caused by AI systems. Providers of high-risk AI systems will be required, by default, to undertake testing and create technical documentation. Under Article 77, Equality Bodies and other national fundamental rights authorities (NFRA) are given a right to access this documentation,⁵ or indeed “any documentation created or maintained under this Regulation in accessible language and format” insofar as it is necessary to fulfil “their mandates within the limits of their jurisdiction.”⁶ This “right of access” seeks to help Equality Bodies in their legal casework and investigations, and also other fundamental rights authorities with relevant mandate and powers, in carrying out work on AI. Should the documentation prove insufficient in this regard, fundamental rights authorities can also exercise their “right to testing.” Through this right they can make a request to national supervision and enforcement bodies which, under the AIA, are called market surveillance authorities (MSA) for further testing to be carried out,⁷ and collaborate with the MSA in the evaluation.⁸

The “right to access documentation” and “right to testing” will be the two key mechanisms for Equality Bodies to detect and monitor algorithmic discrimination and to that end they need to engage with harmonised technical standards. Testing and documentation requirements are discussed in the AIA itself, but their details will largely be defined in a series of harmonised technical standards currently being written by European standards settings organisations. These will determine what information and level of detail are included in this documentation. It is thus essential for Equality Bodies to be aware of the range and scope of standards currently being prepared, and how they link with AIA requirements and their new powers.

This report provides guidance to Equality Bodies on how to effectively use their new AIA powers and harmonised technical standards to investigate AI bias and discrimination. Section 2 introduces the concept and purpose of technical standards in product safety regulation and their likely content based on historical lessons. Section 3 then examines the concepts of bias and discrimination and how they are used in the AIA. Section 4 introduces the new documentation and testing powers granted to Equality Bodies and explains how they connect with requirements faced by providers and deployers of high-risk AI systems and general-purpose AI models. Section 5 presents a series of reflections on challenges for Equality Bodies using the AIA and technical standards to investigate AI bias and discrimination. Section 6 concludes with concrete recommendations about how Equality Bodies can use their new powers most effectively.

⁵ Not all fundamental right authorities are granted powers under Article 77. Rather, bodies must be named by national governments. The deadline for naming authorities was October 31, 2024. A full list has not yet been published by the European Commission, but individual lists can be found. For example, for a list of the Republic of Ireland’s nine named authorities, see: <https://enterprise.gov.ie/en/what-we-do/innovation-research-development/artificial-intelligence/eu-ai-act/>.

⁶ AIA Article 77(1).

⁷ AIA Article 77(3).

⁸ AIA Article 79(2).





2. Standards and the AIA

How to use the Artificial Intelligence Act to investigate AI bias and discrimination: A guide for Equality Bodies



In May 2023 the European Commission sent a standardisation request to the European Committee for Standardisation (“CEN”) and the European Committee for Electrotechnical Standardisation (“CENELEC”). The request includes a list of harmonised standards to be developed to aid in implementation of the AIA.⁹ These standards will fill in many of the practical details left open in the AIA itself which are essential for the framework to be operationalised and enforced (see: Box 1). They will thus play a key regulatory role for AI in the EU, fundamentally shaping the products, services, and organisational practices of AI providers and deployers.

But what exactly are technical standards? Standardization is a process undertaken by standards setting organisations (SSO), often at the request of policy-makers, to define voluntary technical or quality specifications for products and services.¹⁰ In turn, standards are technical documents “designed to be used as a rule, guideline or definition...a consensus-built, repeatable way of doing something,”¹¹ or more specifically documents that list “requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose.”¹²

Simply put, technical standards set self-defined expectations for industry to ensure the safety and quality of their products and services. They create a set of rules and requirements to be voluntarily followed, for example by AI providers or deployers. For industry, the incentive to follow standards comes from the reputational and regulatory benefit of adherence, and potential lower costs and efficiency gains from following pre-defined rules.¹³ Aligning how a product or service is designed, implemented, governed, or used with a set of common rules demonstrates a commitment to self-governance, product safety, and regulation.

Reflecting this, using a standard will often be enough to show that specific products or services are legally compliant. In effect, standards are often the easiest way for a company to show they are following the law. Regulators likewise find standards useful, for example to consolidate expert knowledge to address risks (as often happens in safety regulation),¹⁴ or promote

⁹ ‘Register of Commission Documents - C(2023)3215’ <[https://ec.europa.eu/transparency/documents-register/detail?ref=C\(2023\)3215&lang=en](https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en)> accessed 17 November 2024.

¹⁰ Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council 2012 para (1). *ibid.*

¹¹ CEN/CENELEC, ‘European Standards’ (CEN-CENELEC) <<https://www.cencenelec.eu/european-standardization/european-standards/>> accessed 22 November 2024.

¹² International Organization for Standardization, ‘ISO - Standards’ (ISO) <<https://www.iso.org/standards.html>> accessed 22 November 2024.

¹³ CEN/CENELEC (n 11).

¹⁴ On standards as a means of global governance, cf. Dieter Kerwer, ‘Rules That Many Use: Standards and Global Regulation’ [2005] 18 *Governance* 611, 616. On standards as a means of global governance, cf. *ibid.*

adherence to ethical norms and legal requirements.¹⁵ This is not to say that regulators have direct control in settings or enforcing standards (as they are self-defined by industry-led SSOs), but rather that they benefit from their existence and adherence by filling in gaps in knowledge and lowering their burden of enforcement by enabling regulated entities to follow a common set of practices.

Standards are consensus-based and in principle could involve all interested and affected stakeholders including manufacturers, consumers, and regulators of particular products, services, or materials. In practice, manufacturers play a central role in SSOs, while civil society is also occasionally involved but in a secondary capacity, for example as “observers” (as is the case with the AIA), due to a lack of resources or representation across Member States.¹⁶

2.1. Harmonised standards

In the context of the AIA, AI providers are not legally required to follow standards; companies can freely choose how they self-assess. However, the easiest path to comply with the law is to implement “harmonised standards.” These are technical standards created by European SSOs i.e., CEN/CENELEC) at the direction of the European Commission that, if followed, create a “presumption of conformity” with the law. This is critically important for businesses because they must demonstrate compliance before making AI products and services available to the EU market.

This is a highly advantageous approach for AI providers. Using harmonized standards reduces their legal uncertainty because they will not need to interpret or translate the “essential requirements” for high-risk systems themselves.

Harmonised standards are effectively the “preferred default” of the European Commission; manufacturers who follow them enjoy a lower regulatory burden¹⁷ than those who do not. Harmonised standards are expected to ease the implementation of the AIA by filling in essential practical details and requirements for AI providers and deployers. Standardisation is thus a critical aspect of putting the AIA into force by 2026.

For Equality Bodies, harmonised standards are useful in several regards because they specify how providers can meet the AIA’s essential requirements in practice. They will specify the

¹⁵ Martin Ebers, ‘Standardizing AI: The Case of the European Commission’s Proposal for an “Artificial Intelligence Act”’ in Larry A DiMatteo, Cristina Poncibò and Michel Cannarsa (eds), *The Cambridge Handbook of Artificial Intelligence* (1st edn, Cambridge University Press 2022) 331.

¹⁶ The EU does not provide resources for civil society organisations (CSO) to be involved with CEN/CENELEC standardisation processes. Equinet’s participation in JTC-21 is therefore something of an outlier among CSOs because they have members across the EU and resources to participate through an academic project. For more, see: Equality-compliant Artificial Intelligence: Can AI Technical Standards Protect Equality?, Equinet, Available at: <https://equineteurope.org/latest-developments-in-ai-equality/>.

¹⁷ In other words, the effort required to demonstrate their products and services adhere to the law.



types and degrees of technical documentation and statistical evidence AI providers must create and maintain by default. They will likewise determine the type of risk management and post-market monitoring to be carried out by providers to anticipate, identify, and mitigate risks to fundamental rights. Both procedures will enable Equality Bodies to monitor for possible discrimination throughout the lifecycle of high-risk AI systems by working with their national MSAs. Equality bodies have a right to request access to all documentation within the scope of the AIA throughout a system's lifecycle (see: Section 4). As discussed below, these resources can be extremely helpful for investigating AI-based discrimination.

Equality bodies should be aware of the existence of relevant harmonised standards published by CEN/CENELEC because they determine the content and requirements of the technical documentation they can access via Article 77. While Equality Bodies will typically not have access to SSOs or the standardisation process itself, they can review the content of harmonised standards once published.¹⁸ At this stage it is unclear whether harmonised standards will be published behind a paywall; a recent decision from the European Court of Justice requires four non-AIA harmonised standards cited in the Official Journal of the European Union to be made freely available due to an overriding public interest.¹⁹ It remains to be seen whether the precedent set in this case is extended to cover AIA harmonised standards.²⁰

The European Commission has requested CEN/CENELEC to develop standards in ten areas for AI systems: risk management, governance and quality of datasets, record keeping, transparency, human oversight, accuracy, robustness, cybersecurity, quality management, and conformity assessment.²¹ CEN/CENELEC has established the Joint Technical Committee 21 "Artificial Intelligence" (JTC-21) to develop harmonised standards for the AIA.

JTC-21 will not be starting from scratch, but rather plan to take existing international standards under consideration when drafting the AIA harmonised standards.²² Equality bodies can therefore also familiarise themselves with these pre-existing standards to prepare themselves for the eventual publication of the AIA harmonised standards. As with the harmonised AIA standards, international standards are typically not free of charge and are subject to intellectual property protections, both of which may prove to be major barriers to access for Equality Bodies and fundamental rights authorities. These limitations aside, standards and related technical documentations relevant to bias which are being considered by JTC-21 include:

¹⁸ References to the standards will be published in the Official Journal of the European Union, which is public and freely available, but the standards themselves will not be available free of charge.





¹⁹ *PublicResourceOrg, Inc and Right to Know CLG v European Commission* [2024] ECJ Case C-588/21 P.

²⁰ Rossana Ducato, 'Why Harmonised Standards Should Be Open' [2023] 54 IIC - International Review of Intellectual Property and Competition Law 1173.

²¹ 'Register of Commission Documents - C(2023)3215' (n 9).




²² Joint Research Centre and others, *Analysis of the Preliminary AI Standardisation Work Plan in Support of the AI Act* (Publications Office of the European Union 2023) <<https://data.europa.eu/doi/10.2760/5847>> accessed 22 November 2024.



-  ISO/IEC 22989 “Artificial Intelligence concepts and terminology”;
-  ISO/IEC 23894 “AI Risk Management”;
-  ISO/IEC TR 24027 “Bias in AI systems and AI aided decision making”;
-  ISO/IEC TS 12791 “Treatment of unwanted bias in classification and regression machine learning tasks.”

One critical difference to note between CEN/CENELEC harmonised standards and those from ISO/IEC is that the former must address risks to fundamental rights, whereas the latter do not. Overlap should thus be expected but CEN/CENELEC standards are anticipated to cover a broader range of risks and methods to measure and mitigate them.

The following standards being developed by JTC-21 will be particularly important for setting the requirements of technical documentation available to Equality Bodies and FRAs²³:

-  JT021036 “Artificial intelligence – Concepts, measures and requirements for managing bias in AI systems”: The purpose of this standard is to define “concepts, measures and requirements for assessment and treatment of bias in AI systems. This includes bias unwanted by the AI Provider and AI Deployer according to their specification of the AI system, in the context of the AIA.”
-  JT021008 “AI trustworthiness framework”: The purpose of this standard is to define “a framework for AI systems trustworthiness which contains terminology, concepts, high-level horizontal requirements, guidance and a method to contextualize those to specific stakeholders, domains or applications.”
-  JT021024 “AI risk management”: The purpose of this standard is to define “requirements on risk management for AI systems” and provide “clear and actionable guidance on how risk can be addressed and mitigated throughout the entire lifecycle of the AI system.”

Once drafted, these standards may specify requirements for bias testing and documentation, covering for example the types of metrics to be used, documentation to be created and maintained, possible mitigations including debiasing methods and model constraints, or even how to define and compare appropriate groups in evaluating potential performance gaps. The types of tools available for measuring and mitigating bias which may be included in future harmonised standards are reviewed in Section 4.

²³ A list and description of all standards currently being developed by JTC-21 is available at: are CEN/CENELEC, ‘CEN/CLC/JTC 21 Work Programme’ <https://standards.cencenelec.eu/dyn/www/f?p=205:22:0:::FSP_ORG_ID,FSP_LANG_ID:2916257,25&cs=1827B89DA69577BF3631EE2B6070F207D> accessed 22 November 2024.

2.2. Will standards decide when AI is too biased or discriminatory?

Technical standards are rarely purely objective or value neutral. Rather, the standardisation process, and the choice of who gets a “seat at the table”, means they inevitably reflect certain commercial and societal interests, political norms, and moral values.²⁴ At the same time they play a critical role in filling in key conceptual and practical gaps in enforcement of the law (see: Box 1).

AIA harmonised standards will address abstract normative concepts such as bias, robustness, performance, representation, and security. Implementing these concepts inevitably involves addressing challenging normative questions (e.g., When is a system safe enough? Which data biases are acceptable or “unwanted”, and why?) because these concepts mean different things to different parties.²⁵

Historically, when drafting standards with normative content (e.g., ISO 26000 which aims to define ‘social responsibility’), SSOs have monitored existing normative consensus as reflected in international laws and regulatory frameworks.²⁶ Normative consensus refers to agreement on the meaning of normative concepts such as accountability, safety, or bias, or the best way to turn norms into practical requirements, as reflected in pre-existing laws, case law, government policies, regulations, or other frameworks with more democratic legitimacy

BOX 1

Conceptual gaps in the AIA

The AIA does not provide definitions or requirements for many key concepts. The meaning of these concepts will be filled in through harmonised standards.

Discrimination

The AIA provides no definition of a discrimination risk, and no guidance on how to comply with non-discrimination law even though developers of systems are required to examine the data for “biases likely to lead to [...] discrimination prohibited under Union law” (Art 10(2)(f)).

Explanation

The AIA grants individuals a “right to explanation” (Article 86) but does not specify what constitutes an explanation.

²⁴ On this, see, for example: Raymund Werle and Eric J Iversen, ‘Promoting Legitimacy in Technical Standardization’ (2006) 2 Science, Technology & Innovation Studies 19, 21–23. On this, see, for example: *ibid*.

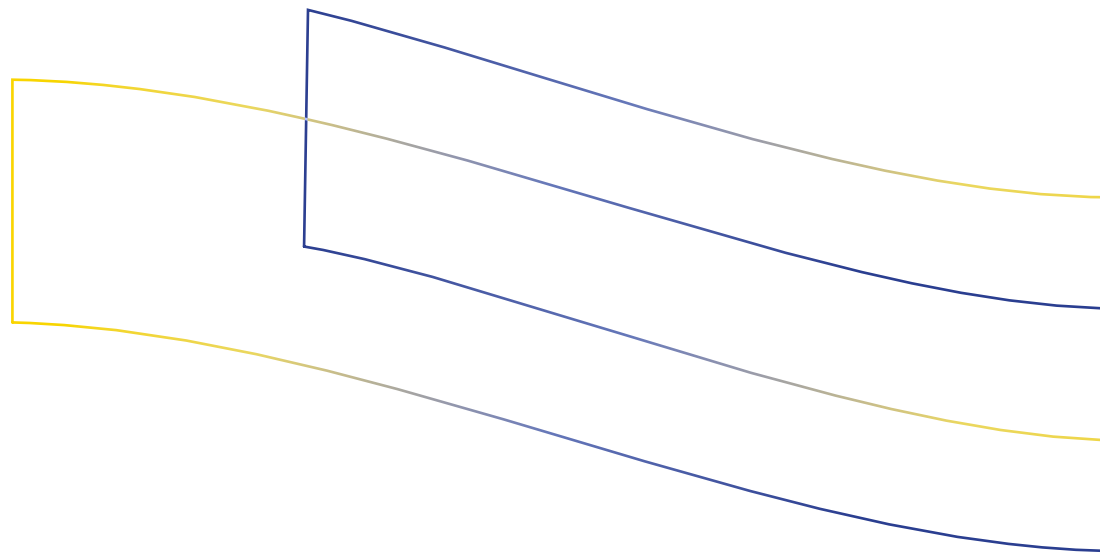
²⁵ WB Gallie, ‘Essentially Contested Concepts’ (1955) 56 Proceedings of the Aristotelian Society 167. Specifying these concepts would involve endorsing specific interpretations or theoretical frameworks for normative ideas (e.g., equality, transparency, dignity). It would likewise require setting acceptable or preferred trade-offs between competing interests, for example between equal treatment (i.e., formal equality) or levelling the playing field (i.e., substantive equality). Answering such questions is difficult and highly context-sensitive; see: Section 5.1.

²⁶ Stephanie Bijlmakers and Geert van Calster, ‘You’d Be Surprised How Much It Costs to Look This Cheap! A Case Study of ISO 26000 on Social Responsibility’, *The Law, Economics and Politics of International Standardisation* (Cambridge University Press 2015); Johann Laux, Sandra Wachter and Brent Mittelstadt, ‘Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act’ (2024) 53 Computer Law & Security Review 105957.

than SSOs. Tracking consensus defers the need to address critical normative questions for SSOs, instead allowing AI developers and users to define concepts like fairness themselves, for example through fundamental rights impact assessments or risk management frameworks (see: Sections 4.3 and 4.5).

In theory, CEN/CENELEC could tackle these normative questions directly. In practice, answers will likely come from other sources, especially impact assessments and risk management systems. SSOs have long faced concerns about democratic legitimacy,²⁷ with standardisation processes often excluding impacted stakeholders with limited resources (e.g., time, expertise, staff), civil society, and the general public.²⁸ These concerns contribute to the reluctance among SSOs to answer normative questions directly in standards.

This means that Equality Bodies should not expect AIA standards to answer the sort of questions they address day-to-day. For example, standards are unlikely to indicate whether a particular bias is intentional or unwanted, or whether a gap in performance is discriminatory or acceptable. They can provide methods and processes to measure bias or inequality in AI systems but are unlikely to say whether these biases and inequalities are (il)legal or (un)ethical.²⁹



²⁷ Standardization as a governance tool has faced significant criticism for its perceived lack of legitimacy. Primarily a technical discourse, it often excludes non-expert stakeholders and the general public. Industry representatives hold considerable sway within standard-setting organizations (SSOs). However, the involvement of technical experts in standardization can be highly political. For more, see: Laux, Wachter and Mittelstadt (n 26); Bijlmakers and van Calster (n 26).

²⁸ Laux, Wachter and Mittelstadt (n 26).

²⁹ *ibid.*



3. Bias in the AIA

How to use the Artificial Intelligence Act to investigate AI bias and discrimination: A guide for Equality Bodies



AI systems are typically trained real-world data, meaning they learn the biases and inequalities that exist in society. These biases may advantage or disadvantage specific groups, objects, concepts, or outcomes. Even when working with data that appears unbiased, unintended or unwanted biases can still emerge in the resulting models. Mitigating biases can be particularly frustrating because eliminating one type can lead to the emergence or reinforcement of another; there is no such thing as an “unbiased” AI model or system.

Bias can be understood in both a neutral and normative sense. Neutrally, bias is simply an expression of preference for one state of affairs or class of things over another: for example, a positive preference towards, or negative prejudice against, certain groups of people, brand of cars, species of animals, or other classes of phenomena. AI systems that assign labels to data, classify cases, or predict outcomes (e.g., the likelihood of repaying a loan), cannot by definition operate without bias.

This is not the sense in which “bias” is typically discussed in the context of AI and regulation or ethics. Rather, it is the normative sense of the word. Biases become normative when a particular preference or prejudice is found problematic for social, ethical, legal, or other relevant reasons. AI systems that perform worse when classifying skin cancer with patients for darker skin tones, for example, would be expressing a problematic normative bias, owing perhaps to an imbalance of skin tone in their training data.³⁰

The usage of the term “bias” in the AIA aligns with the normative definition. In the AIA the concept appears in connection to “unfair biases,” “discriminatory impacts,” “biased results and discriminatory effects,” and impacts on fundamental rights (such as equality).³¹ Biases, as with equality, tend to be measured according to legally protected attributes (e.g., ethnicity, gender, age, religion).³²

Bias is mentioned in many of the Recitals of the AIA,³³ but only appears in two legally binding Articles. Article 10 on data and data governance requires providers of high-risk AI systems to examine their training, testing, and validation data sets for “possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to

³⁰ Brent Mittelstadt, Sandra Wachter and Chris Russell, ‘The Unfairness of Fair Machine Learning: Leveling Down and Strict Egalitarianism by Default’ (2024) 30 Michigan Technology Law Review <<https://repository.law.umich.edu/mltr/vol30/iss1/3>>; Joy Buolamwini and Timnit Gebru, ‘Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification’, *Conference on Fairness, Accountability and Transparency* (2018) <<http://proceedings.mlr.press/v81/buolamwini18a.html>> accessed 31 July 2020.

³¹ See for example AIA Recital 27, Recital 32, Article 27, and other mentions of the term.

³² Sandra Wachter, Brent Mittelstadt and Chris Russell, ‘Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI’ (2021) 41 Computer Law & Security Review 105567.

³³ Recital 67 on data quality management points towards several potential sources of bias in training, validation, and testing datasets: “Data sets for training, validation and testing, including the labels, should be relevant, sufficiently representative, and to the best extent possible free of errors and complete in view of the intended purpose of the system... Biases can for example be inherent in underlying data sets, especially when historical data is being used, or generated when the systems are implemented in real world settings. Results provided by AI systems could be influenced by such inherent biases that are inclined to gradually increase and thereby perpetuate and amplify existing discrimination, in particular for persons belonging to certain vulnerable groups, including racial or ethnic groups.”



Key definitions in the AIA (Article 3)

Training data: Data used for training an AI system through fitting its learnable parameters.

Validation data: Data used for providing an evaluation of the trained AI system and for tuning its non-learnable parameters and its learning process in order, inter alia, to prevent underfitting or overfitting.

Testing data: Data used for providing an independent evaluation of the AI system in order to confirm the expected performance of that system before its placing on the market or putting into service.

Input data: Data provided to or directly acquired by an AI system on the basis of which the system produces an output.

Performance: Ability of an AI system to achieve its intended purpose.

Risk: Combination of the probability of an occurrence of harm and the severity of that harm.

Bias: Not defined.

Accuracy: Not defined.

Robustness: Not defined.

discrimination prohibited under Union law.”³⁴ If identified, providers must also describe “appropriate measures to detect, prevent and mitigate possible biases.”³⁵ These may include using special powers granted in Article 10(5) to process sensitive data for the purposes of bias detection and correction.

Bias is also linked to other essential concepts and requirements of accuracy, robustness, and cybersecurity through Article 15 (see: Box 2), which identifies the possibility of biases emerging in systems once they are placed onto the market. This is a question of system robustness, or systems being designed in such a way as to be resilient against “errors, faults or inconsistencies that may occur within the system or the environment in which the system operates.”³⁶ As with bias, JTC-21 standards on robustness and other essential requirements are still forthcoming. At a minimum the risk management system created for AI risk systems will be a key mechanism to ensure robustness over the system’s lifecycle (see: Section 4.5).

While the precise content of the forthcoming CEN/CENELEC bias-related standards remains unknown at this stage, existing research, development, and technical standards published by international SSOs with which CEN/CENELEC are collaborating can be used to predict their content. Two recent publications by from ISO/IEC on AI bias are particularly helpful:

³⁴ AIA Article 10.

³⁵ AIA Article 10(2)(g).

³⁶ AIA Article 15(4).

BOX 3

Example of unwanted bias

Imagine a CV screening AI system being used by an employer to decide who to invite for an interview. The system has learned that career breaks or short-term contracts are correlated with poor future job performance. This learned bias could inadvertently disadvantage candidates who have taken maternity or paternity leave. The unwanted negative association between parental status and merit (i.e., future job performance) is the type of harm of bias that debiasing and fairness methods aim to correct (see: Sections 4.1.2 and 4.1.3).

1. **ISO/IEC TR 24027:2021**: A technical report on bias in AI systems and AI aided decision making;
2. **ISO/IEC TS 12791:2024**: A technical specification on the treatment of unwanted bias in machine learning.³⁷

Both documents will likely form the basis of a future international standard on AI bias from the International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). Bias has not yet been given a precise definition by CEN/CENELEC, but it is anticipated that the definitions of such key concepts (e.g., bias, performance, unfairness) will align with existing international standards. For example, ISO/IEC 24027 defines bias as the “systematic difference in treatment of certain objects, people or groups in comparison to others.”

3.1. Unwanted bias and fairness

Bias-related standards make a distinction between “unwanted” and intentional or acceptable biases (see: Box 3).³⁸ There are many different types and sources of biases that can affect AI systems, each of which can be desired or unwanted to different degrees (see: Box 4). Unwanted biases are those which are not intentionally included in the system’s design or “intended purpose” as defined in a system’s technical documentation (which Equality Bodies can access), promotional materials, or instructions for deployers.³⁹ They are undesirable or unintentional according to the AI provider or deployer’s preferences. Harmonised standards cannot define unwanted biases at a high level for precisely this reason—they are unwanted according to the provider or deployer of specific systems (see: Section 2.2). Efforts to make AI systems less biased and fairer tend to focus on mitigating unintended negative impacts in a model’s design or training, test, or validation data (see: Box 5).

³⁷ Technical reports and technical specifications are not standards themselves, but rather early components in the standards setting process. Technical reports discuss the state of the art in research and development in a given topic area, but are not prescriptive, meaning they do not list specific requirements, recommendations, or permissions. Technical specifications are a pre-cursor to standards and are more prescriptive but are used in areas where consensus has not yet been achieved by the SSO, or where the topic area is still under active development. See: ‘ISO - Deliverables’ (ISO) <<https://www.iso.org/deliverables-all.html>> accessed 17 November 2024.

³⁸ ISO 24027 and JTC-21 bias management standard both refer to “unwanted biases.”

³⁹ The intended purpose describes “the use for which an AI system is intended by the provider, including the specific context and conditions of use...” AI Act Article 3(12).

Categories of bias from ISO 24027

Bias: Systematic difference in treatment of certain objects, people, or groups in comparison to others.

Human cognitive bias: Bias that occurs when humans are processing and interpreting information. Can affect design decisions about data collection and labelling, system design, model training, and others.

Types: Automation bias; Group attribution bias; Implicit bias; Confirmation bias; In-group bias; Out-group homogeneity bias; Societal bias; Rule-based system design; Requirements bias.

Data bias: Data properties that if unaddressed lead to AI systems that perform better or worse for different groups. Data bias arises from technical design decisions and constraints and it can be caused by human cognitive bias, the training methodology chosen and variances in training infrastructure.

Types: Statistical bias including selection bias, sampling bias, coverage bias, non-response bias, confounding variables, non-normality; Data labels and labelling process bias; Non-representative sampling; Missing features and labels; Data processing bias; Simpson's paradox; Data aggregation; Distributed training.

Engineering decisions bias: Biases in machine learning model architectures, including all model specifications, parameters, and manually designed features.

Types: Feature engineering; Algorithm selection; Hyperparameter tuning; Informativeness; Model bias; Model interaction bias including model expressiveness.

Reflecting this, JTC-21's draft bias management standard refers to "bias unwanted by the AI Provider and AI Deployer," again suggesting that the types of biases relevant to the AIA are those that link to inequality, discrimination, or other fundamental rights impacts.⁴⁰ One of the standards likely to influence the future harmonized bias standards, namely ISO 24027, likewise makes a distinction between the neutral and social senses of bias, saying that in a social context, bias has a clear negative connotation as one of the main causes of discrimination and injustice." It refers to the issues caused by this type of socially problematic bias as issues of "fairness" and "unfairness." Specifically, a fair AI output would be "a treatment, a behaviour or an outcome that respects established facts, beliefs and norms and is not determined by favouritism or unjust discrimination."

Bias is about differences in treatment, whereas fairness is about the impact that treatment has on individuals, groups, organisations, and societies.⁴¹ This distinction between neutral

⁴⁰ CEN/CENELEC (n 23).

⁴¹ ISO 24027, Clause 5.3.

Types of unfair impact of AI (ISO 24027)

Unfair allocation: The system unfairly extends or withholds opportunities or resources in ways that have negative effects on some parties as compared to others.

Unfair quality of service: The system performs less well for some parties than for others, even if no opportunities or resources are extended or withheld.

Stereotyping: The system reinforces existing societal stereotypes.

Denigration: The system behaves in ways that are derogatory or demeaning.

“Over” or “under” representation and erasure: The system over-represents or underrepresents some parties as compared to others, or even fails to represent their existence.

and problematic (or “unwanted”) bias and its social impacts is helpful, but the definition of fairness leaves much room for interpretation. How, for example, can it be determined whether a particular instance of favouritism or discrimination caused by AI is acceptable or unjust? Methods for bridging the gap between non-discrimination law and statistical measures of bias are discussed below (see: Section 4.1.2.2).

The JTC-21 harmonised standards relevant to bias are unlikely to tell providers which biases are wanted or unwanted. Rather, whether a specific bias is desirable or unwanted must be decided on a context-specific basis. In practice, FRIAs and risk management systems for high-risk AI systems are intended to help identify and evaluate biases and their potential link to unfair or discriminatory impacts (see: Sections 4.3 and 4.5). Risk management systems may define expected distributions of data or outputs (e.g., errors, positive or negative outcomes) across relevant groups of people impacted by the system. They are also expected to decide which mitigations are necessary to correct

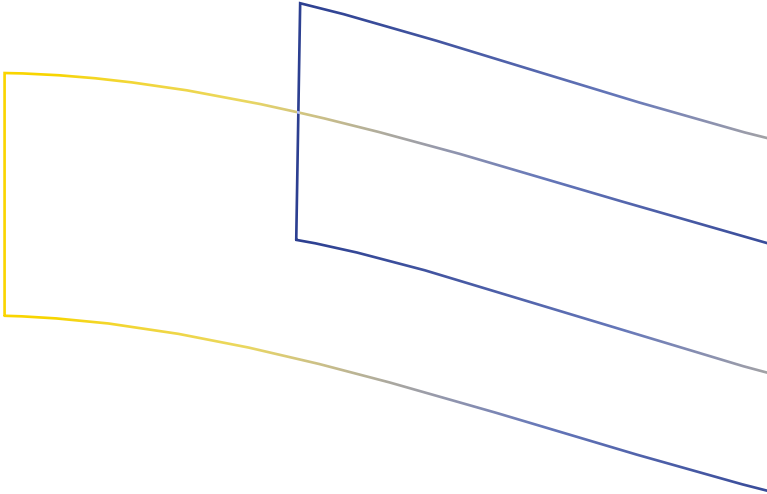
unwanted impacts (e.g., forcing positive decisions to be distributed equally between male and female applicants).

Equality bodies should pay close attention to how AI providers decide whether a particular bias is desirable or unwanted through FRIAs and risk management systems. This process may itself be biased or not aligned with expectations of equality law, for example by ignoring biases impacting intersectional groups. Prior research on the difficulty and complexity of translating the concept of “equality” into technical bias measures and methods is highly relevant here to show how contextual and subjective the definition of “unwanted” bias can be.⁴²

Equality bodies will likewise want to examine datasheets and model cards provided as part of a system’s technical documentation (see: Section 4.1). Both types of documentation will describe the provenance (i.e., source, history, methods for collection, cleaning, labelling) of training, testing, and validation data. This information is crucial to identify possible biases, in particular

⁴² Wachter, Mittelstadt and Russell, ‘Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI’ (n 32).

historical and representational biases related to social prejudices or data gaps impacting particular demographic groups.⁴³ Any such problems or gaps in a data's provenance will have knock-on effects on AI systems that learn from the data and should be a starting point for investigations by Equality Bodies into potential discrimination caused by the system in question.



⁴³ Harini Suresh and John V Guttag, 'A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle' [2021] Equity and Access in Algorithms, Mechanisms, and Optimization 1.



4. Powers of Equality Bodies under the AIA

How to use the Artificial Intelligence Act to investigate AI bias and discrimination: A guide for Equality Bodies



The AIA creates obligations for providers and deployers of high-risk AI systems to create and maintain various types of documentation throughout the system’s lifecycle (see: Box 6). Equality Bodies and other NFRAs nominated as Article 77 authorities have several important rights in this context.

First, they have a **right to access** the documentation in “accessible language and format” by making a request to the MSA of the relevant Member State.⁴⁴ It is worth noting that at this stage of implementation of the AIA it is unclear what type of language or format may qualify as sufficiently accessible. These terms could be interpreted narrowly as referring strictly to accessibility for readers with disabilities, or machine readability. However, a broader interpretation would follow Recital 72, which discusses requirements for instructions for

use prepared by AI providers for AI deployers. Here, “accessible” documentation should include illustrative examples and information that is meaningful, comprehensible, and understandable to deployers based on their level of knowledge.⁴⁵

It is also worth noting the AIA stipulates NFRAs have access to “any documentation created or maintained” under the AIA that is “necessary for effectively fulfilling their mandates within the limits of their jurisdiction.”⁴⁶ While the scope of “any documentation” is unclear at this stage of implementation of the AIA, given the broad formulation it is sensible to conclude that NFRAs will have access to all the types of data, analysis, and related documentation discussed below.

Second, Equality Bodies have a **right to request technical testing** if they find that the documentation provided is insufficient to determine whether fundamental rights have been

AIA documentation requirements

Article 11: Technical documentation for high-risk AI systems, including instructions for use by deployers.

Article 53: Technical documentation for general-purpose AI models.

Article 27: Fundamental rights impact assessments.

Article 72: Post-market monitoring data.

Article 73: Incident reporting.

Articles 40-49: Evidence in support of conformity assessments (may be identical to Article 11 documentation).

⁴⁴ AIA Article 77(1).

⁴⁵ AIA Recital 72. This Recital is not about Article 11 technical documentation, but rather transparency requirements of AI providers to AI deployers (Article 13), and thus this definition may not apply in this context. Nonetheless, it is the only explanation of what “accessible” means in relation to documentation in the AIA. It states: “In order to enhance legibility and accessibility of the information included in the instructions for use, where appropriate, illustrative examples, for instance on the limitations and on the intended and precluded uses of the AI system, should be included. Providers should ensure that all documentation, including the instructions for use, contains meaningful, comprehensive, accessible and understandable information, taking into account the needs and foreseeable knowledge of the target deployers. Instructions for use should be made available in a language which can be easily understood by target deployers, as determined by the Member State concerned.”

⁴⁶ AIA Article 77(1).

or will likely be violated and have submitted a reasoned request to the MSAs.⁴⁷ This could be the case, for example, where the documentation does not report results of a sufficiently broad range of fairness tests (see: Section 0), or fails to be written in an accessible format and language.⁴⁸ Equality bodies may likewise have an opportunity to participate in any subsequent evaluation carried out by the MSA made on the basis of their request.⁴⁹

Third, Equality Bodies have the right to be informed and request “full cooperation” by MSAs whenever risks to fundamental rights, including equality and non-discrimination, are identified. In this way, Equality Bodies potentially possess unique powers to influence decisions of MSAs to call for corrective actions to AI systems based on post-market monitoring data.

As discussed above, Equality Bodies can collaborate with MSAs to access documentation and evaluate AI systems that require further testing. In this context, if a MSA has reason to believe a system poses a risk to fundamental rights, they are obliged to inform and cooperate with relevant NFRAs including Equality Bodies to evaluate the system. The MSA and NFRAs then cooperatively evaluate whether the high-risk system has fallen out of compliance with the AIA’s essential requirements (Chapter III). For non-compliant systems MSAs are required to compel the operator to “take all appropriate corrective actions to bring the AI system into compliance, to withdraw the AI system from the market, or to recall it within a period the MSA may prescribe.”⁵⁰

The rights to access and to request testing hold the potential to be much more robust than they may first appear. Equality Bodies can, through collaboration with MSAs, identify discriminatory or otherwise harmful AI systems and call for them to be brought into compliance with the AIA or withdrawn from the market.⁵¹ Imagine, for example, a case in which an Equality Body uses technical documentation to show that a system performs very poorly for certain intersectional demographic groups. If this risk was not noted and mitigated in the provider’s risk management plan, the Equality Body could potentially make a request to the relevant MSA to conduct further evaluation of the system under Article 79. Furthermore, if a risk to equality has

⁴⁷ AIA Article 77(3).

⁴⁸ AIA Article 77(1).

⁴⁹ AIA Article 79(2).

⁵⁰ AIA Article 79(2). These powers are not limited to individual member states; rather, where the non-compliance spans other Member States, the MSA can notify the European Commission of its evaluation and actions requested of the operator [Article 79(3)].

⁵¹ AIA Article 79(5).

been identified this also triggers an obligation by deployers and developers to cooperate “as relevant” with the Equality Bodies and other Article 77 authorities.⁵²

Article 79 thus provides perhaps the strongest enforcement mechanism available to Equality Bodies and other NFRAs under Article 77. Close collaboration with MSAs should thus be a key priority for Equality Bodies under the AIA.

Together, the rights to access documentation, request testing and compel collaboration by MSAs could be very powerful rights to help hold AI providers accountable for the impact of their high-risk systems on fundamental rights like equality.⁵³ A broad range of documentation will be available to Equality Bodies and NFRAs, including (1) technical documentation for high-risk AI systems, (2) technical documentation for general-purpose AI models, (3) fundamental rights impact assessments, and (4) post-market surveillance data. Information provided in this documentation will be based on the harmonised technical standards discussed above, and kept up to date through (5) risk management systems AI providers must deploy along with their high-risk systems.

4.1. Technical documentation for high-risk AI systems

Providers of high-risk AI systems are required to publish and update technical documentation including instructions for use in a clear and comprehensive form describing how the system complies with the law and is intended to be used by deployers.⁵⁴ This documentation will

⁵² AIA Article 79(2) which states “Where risks to fundamental rights are identified, the market surveillance authority shall also inform and fully cooperate with the relevant national public authorities or bodies referred to in Article 77(1). The relevant operators shall cooperate as necessary with the market surveillance authority and with the other national public authorities or bodies referred to in Article 77(1). Where, in the course of that evaluation, the market surveillance authority or, where applicable the market surveillance authority in cooperation with the national public authority referred to in Article 77(1), finds that the AI system does not comply with the requirements and obligations laid down in this Regulation, it shall without undue delay require the relevant operator to take all appropriate corrective actions to bring the AI system into compliance, to withdraw the AI system from the market, or to recall it within a period the market surveillance authority may prescribe, and in any event within the shorter of 15 working days, or as provided for in the relevant Union harmonisation legislation.”




⁵³ The extent to which each of these documentation types will be available to NFRAs will depend on enforcement of the AIA in the future, as well as the willingness and capacities of MSAs to facilitate access.

⁵⁴ These documents will be shared with national competent authorities and notified bodies in the first instance. National competent authorities are (1) market surveillance authorities and (2) notifying authorities appointed by Member States. Market surveillance authorities are established under Regulation 2019/1020 and are responsible for ensuring compliance of products placed on the EU market with relevant EU harmonisation legislation. Notifying authorities are bodies responsible for notifying, assessing, and designating notified bodies who carry out conformity assessment. Notified bodies are conformity assessment bodies, or independent organisations responsible for conformity assessment in cases where third party assessment is required under the AIA or EU harmonisation legislation (see Annex I of the AIA for a list of relevant legislation). All Member States must appoint at least one notifying authority and market surveillance authority. Member States are allowed to name any public entity as a national competent authority, including for example data protection, cybersecurity, or competition authorities, according to their local needs and capacities. As an example, Germany has named its Federal Accreditation Body (“Deutsche Akkreditierungsstelle”) as notifying authority and its Federal Network Agency (“Bundesnetzagentur”) as market surveillance authority for the purposes of enforcing the AIA.



be an essential resource for Equality Bodies seeking to identify or assess potential cases of algorithmic discrimination.




Requirements for technical documentation for high-risk AI systems will be based on the requirements of Annex IV of the AIA (i.e., minimum required information). The documentation will contain information derived from a variety of testing and tuning methods and tools, reflected in technical standards and developed by researchers in recent years to analyse and reduce bias in AI systems. These tools and methods can help providers of high-risk AI systems meet their requirements to measure, document, and mitigate bias, and report how they did so in the system's technical documentation. General categories of the tuning and testing methods and tools Equality Bodies can expect to see in the technical documentation include:

-  Bias tests and de-biasing methods which focus on how individuals or groups are treated by a system. They can target training, test, and validation datasets or change model outputs with pre-, in-, and post-processing methods;⁵⁵
-  Fairness metrics and enforcement methods which focus on measuring and mitigating the impact of biases (or differences in treatment) on people and society. These include individual, group, and counterfactual approaches, which are implemented in open-source toolkits;⁵⁶
-  Transparency and explainability methods that produce explanations of how AI turns inputs into outputs globally (i.e., how the model behaves overall) or at a local level (e.g., how the model decides specific cases or classifies specific groups), as well as model inspection methods, interpretable models, and post hoc explanations addressing the causes or prevalence of biases in a model;⁵⁷

⁵⁵ Dino Pedreshi, Salvatore Ruggieri and Franco Turini, 'Discrimination-Aware Data Mining', *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM 2008) <<http://doi.acm.org/10.1145/1401890.1401959>> accessed 27 September 2017; Rachel KE Bellamy and others, 'AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias' [2018] arXiv:1810.01943 [cs] <<http://arxiv.org/abs/1810.01943>> accessed 15 March 2019; Salvatore Ruggieri, Dino Pedreschi and Franco Turini, 'Data Mining for Discrimination Discovery' (2010) 4 ACM Transactions on Knowledge Discovery from Data (TKDD) 9.

⁵⁶ Sahil Verma and Julia Rubin, 'Fairness Definitions Explained', *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (IEEE 2018); Moritz Hardt, Eric Price and Nati Srebro, 'Equality of Opportunity in Supervised Learning', *Advances in Neural Information Processing Systems* (2016); Cynthia Dwork and others, 'Fairness Through Awareness' [2011] arXiv:1104.3913 [cs] <<http://arxiv.org/abs/1104.3913>> accessed 15 February 2016; Matt J Kusner and others, 'Counterfactual Fairness' in I Guyon and others (eds), *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc 2017) <<http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf>> accessed 17 July 2019; Bellamy and others (n 55).

⁵⁷ Christoph Molnar, *Interpretable Machine Learning* (2020) <<https://christophm.github.io/interpretable-ml-book/>> accessed 31 January 2019; Tim Miller, 'Explanation in Artificial Intelligence: Insights from the Social Sciences' (2019) 267 *Artificial Intelligence 1*; Brent Mittelstadt, 'Interpretability and Transparency in Artificial Intelligence' in Carissa Véliz (ed), *The Oxford Handbook of Digital Ethics* (Oxford University Press 2022) <<https://doi.org/10.1093/oxfordhb/9780198857815.013.5>> accessed 13 September 2022.

-  Standardised documentation such as datasheets for datasets, model cards, nutrition labels (i.e., a simplified datasheet), and fact sheets for users;⁵⁸
-  Impact assessments including human rights, fundamental rights, privacy, and equality impact assessments, as well as Algorithmic Impact Assessments;⁵⁹
-  Documentation of ethics procedures within an organisation, for example reports from internal or external ethics review committees, model or dataset selection criteria, content moderation policies, and other relevant elements of requirements analysis.

Technical documentation under the AIA will likely be organized based on universal and sector-specific documentation standards developed by industry in recent years, including references to “datasheets” and “model cards” in Annex IV. For Equality Bodies it would be helpful to be aware of these two types of documentation in particular because they are well-established in the context of AI research and commercial AI development. Research is already underway on their utility, limitations, and effectiveness at improving the quality of AI research, products, and services, including identifying and mitigating biases.⁶⁰

Dataset documentation methods are designed to assist users and organisations considering adopting a specific AI system, model, or dataset to evaluate the suitability and limitations of a dataset for training models for specific tasks. This typically involves providing information on how the datasets were created and structured, including details about features, data sources, and the processes of data collection, cleaning, and distribution.⁶¹ Some methods also include standardized disclosures and statistical tests related to ethical and legal aspects,⁶² addressing issues like biases, known proxies for sensitive attributes (such as ethnicity and gender), and data gaps. Documenting these characteristics can help uncover problematic biases that machine learning systems might learn from the data, which developers and analysts might otherwise overlook.⁶³

⁵⁸ Timnit Gebru and others, ‘Datasheets for Datasets’ <<https://arxiv.org/abs/1803.09010>> accessed 1 October 2018; Margaret Mitchell and others, ‘Model Cards for Model Reporting’ [2019] Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* ’19 220; Sarah Holland and others, ‘The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards’ [2018] arXiv:1805.03677 [cs] <<http://arxiv.org/abs/1805.03677>> accessed 1 October 2018.

⁵⁹ Dillon Reisman and others, ‘Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability’ [2018] AI Now Institute 1; L Edwards, D McAuley and L Diver, ‘From Privacy Impact Assessment to Social Impact Assessment’, *2016 IEEE Security and Privacy Workshops (SPW)* (2016).

⁶⁰ See for example Abhishek Wadhvani and Priyank Jain, ‘Machine Learning Model Cards Transparency Review: Using Model Card Toolkit’, *2020 IEEE Pune Section International Conference (PuneCon)* (IEEE 2020) <<https://ieeexplore.ieee.org/abstract/document/9362382/>> accessed 3 December 2024; José Luiz Nunes and others, ‘Using Model Cards for Ethical Reflection: A Qualitative Exploration’, *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems* (ACM 2022) <<https://dl.acm.org/doi/10.1145/3554364.3559117>> accessed 3 December 2024; Carolina AM Heming and others, ‘Benchmarking Bias: Expanding Clinical AI Model Card to Incorporate Bias Reporting of Social and Non-Social Factors’ [arXiv, 2 July 2024] <<http://arxiv.org/abs/2311.12560>> accessed 3 December 2024; Timnit Gebru and others, ‘Datasheets for Datasets. Documentation to Facilitate Communication between Dataset Creators and Consumers’ [2021] Communications of the ACM <<https://cacm.acm.org/research/datasheets-for-datasets/>> accessed 3 December 2024; Karen L Boyd, ‘Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data’ (2021) 5 Proceedings of the ACM on Human-Computer Interaction 1.

⁶¹ Gebru and others (n 58). *ibid.*

⁶² Holland and others (n 58). *ibid.*

⁶³ Gebru and others (n 58); Holland and others (n 58). Gebru and others (n 58); Holland and others (n 58).

Similar initiatives exist for trained machine learning models. “Model reporting” documentation is intended to accompany models when they are deployed in environments that differ from their training contexts. For instance, the “model cards for model reporting” initiative created a documentation standard that outlines various performance attributes and intended use cases, including how performance may vary across different cultural, demographic, phenotypic, and intersectional groups.⁶⁴ User-oriented model documentation has also been proposed to boost user trust and adoption. One example is “FactSheets,” which would require AI providers to offer a standardized statement of conformity regarding the purpose, performance, safety, security, and provenance of their models in a user-friendly format.⁶⁵

The AIA has adopted the language of “datasheets” and “model cards” without explicitly endorsing any of the initiatives discussed above. The required structure and content of these types of documentation will be set through JTC-21 standards drawing on the technical documentation requirements set out AIA (see: Sections 4.1.1 and 4.2). They may nonetheless be a good fit to fulfil documentation requirements under the AIA because they are aimed at broad and diverse audiences and thus should be accessible in language and format.

4.1.1. Scope of the technical documentation

Technical documentation for high-risk AI systems will need to cover many aspects, from general descriptions of the data and methods used to build, train, and validate them, to their intended uses and interactions with other technologies, as well as relevant human oversight, accountability, and cybersecurity measures. Table 1 points Equality Bodies towards the required elements of technical documentation that are relevant to bias and discrimination assessment and the respective components of the related technical standards.

Specific aspects of the documentation requirements laid down in Annex IV will be particularly relevant for Equality Bodies in measuring the types of AI biases and disparities discussed above (see: Section 2), in particular:

General description of the AI system: Providers must describe the system’s intended purpose, expected levels of accuracy, relevant hardware and software components, forms of deployment, documentation of physical products in which the AI system is embedded, user interface, and instructions for use. It is worth noting here that the “intended purpose,” “expected level of accuracy, and “reasonably foreseeable misuses” of the system are defined by the AI provider. This form of self-regulation presents a clear risk of “ethics washing” which Equality Bodies should monitor.⁶⁶

⁶⁴ Mitchell and others (n 58). *ibid.*

⁶⁵ M Arnold and others, ‘FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity’ [2019] 63 *IBM Journal of Research and Development* 6:1. *ibid.*

⁶⁶ Ben Wagner, ‘Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping’ [2018] *Being profiling*. *Cogitas ergo sum* 1; Wachter, Mittelstadt and Russell, ‘Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI’ (n 32).



Effectively, the AIA only holds providers liable against standards they have self-defined. This creates a significant enforcement loophole which will likely severely limit provider liability for the downstream impact of their systems. If providers define a broad intended purpose, low expected levels of accuracy for particular groups, or a narrow range of potential misuses, their system may appear to function as intended and yet cause significant harm (see: Box 7). How these concepts are defined is thus a key consideration for Equality Bodies. Setting the “expected level of accuracy” low for historically underrepresented groups, for example, would mean a system with significant performance gaps between demographic groups would be “working as intended” because the expectation of disparity was built into its design. The intended purpose of the system can also narrow the scope of “foreseeable unintended outcomes and risks” and thus acts as a limitation on provider liability.

Performance measures: Providers will report the results of tests to measure various aspects of a system’s performance such as accuracy (i.e., how often the system makes the correct prediction), robustness (i.e., resilience against errors and unanticipated behaviours), and compliance with other essential requirements under Chapter III, including potential discriminatory impacts. These include measures that reveal performance gaps between demographic groups can assist Equality Bodies in investigating bias.⁶⁷ More information is provided in Section 4.1.2.2.

Robustness measures may prove particularly relevant to Equality Bodies wishing to monitor how biases are reinforced by AI systems over time. The AIA notes a particular risk that “learning” systems, or those which continue to learn from the environment they are deployed in and thus change how they make predictions or decisions over time, may create bias-reinforcing “feedback loops.” Feedback loops occur when a system learns from its outputs; if these outputs are biased, that bias can be reinforced by using them as a source for further training. Providers are required to report on the possibility of such feedback loops and the steps taken to mitigate them.⁶⁸

Data requirements: Providers must provide datasheets describing training methods, techniques, and datasets used in building the AI system. These will include information about the data’s provenance, scope, characteristics, collection and selection, labelling, and cleaning, all of which can introduce, reduce, or amplify biases.⁶⁹ Datasheets may prove to be a particularly accessible type of documentation because they present the above information in a consistent and concise format. More information on the possible content and utility of information on data requirements is provided in Section 4.1.3.

Human oversight information: Information on how the system’s functionality will be monitored and controlled over time, including how system outputs will be made understandable for deployers. The documentation should also discuss foreseeable

⁶⁷ AIA Annex IV 2(g).

⁶⁸ AIA Article 15(4).

⁶⁹ AIA Annex IV 2(d).

unintended outcomes and risks to fundamental rights and discrimination (based on the intended purpose of the system).⁷⁰

Human oversight is intended to provide iterative assessment and control of a system, including monitoring and mitigation of emergent risks to health, safety, and fundamental rights.⁷¹ As discussed above, the “intended purpose” and list of “reasonably foreseeable misuses” created by the AI providers take on renewed importance in this context because they set the scope of emergent risks targeted by human oversight. Human oversight may in practice have a very narrow scope and miss emergent or novel risks to fundamental rights as a result.

Of particular interest for Equality Bodies is the requirement for providers to set an “expected level of accuracy” for specific groups of people. Deployers are required to oversee the system and report on how accuracy compares over time to the expected level of accuracy as originally defined by the AI provider. This would appear to mean that deployers will need to provide statistics to providers as part of their risk management and post-market monitoring procedures that can be directly used to measure group fairness (see: Section 4.1.2.1).⁷²

Applied standards: Providers must list the harmonised standards applied to the AI system. Where harmonised standards have not been used, providers must explain how the requirements of the AIA (Chapter III, Section 2) have been otherwise met, for example through usage of other relevant standards and technical specifications. This list will enable Equality Bodies to determine when a relevant bias, trustworthiness, or risk management harmonised standard has been followed.⁷³

BOX 7

Ethics washing biased AI systems

Imagine an AI system designed for skin cancer screening that has high accuracy for patients with lighter skin tones and low accuracy for patients with darker skin tones. The AI provider can hide or “wash” this disparity in two ways. First, they can describe the intended purpose as “skin cancer screening for lighter skinned patients.” The system’s high accuracy for this group would mean it is fulfilling its intended purpose and would not appear as biased. Second, the manufacturer could deploy the system for both groups but define the “expected level of accuracy” as high for the lighter skinned patients and low for darker skinned patients. From an equality perspective, the relevant question is whether a system with comparably high accuracy has been developed and deployed for patients with darker skin tones. The intended purpose and expected level of accuracy do not answer this question.

⁷⁰ AIA Annex IV 3.

⁷¹ AIA Article 14[2].

⁷² AIA Annex IV 3.

⁷³ AIA Annex IV 7.

4.1.2. Understanding bias in performance measures

The most critical component of technical documentation that Equality Bodies should look at are the so-called “performance measures.” In this context “performance” is a generic term used as a placeholder for other attributes that measure some aspect of system functionality, for example, the rate at which the model makes correct (i.e., accuracy) or incorrect predictions (i.e., error rate). They are most commonly used for AI systems performing classification tasks, meaning the system is classifying or predicting something about a person or data point (e.g., labelling a person as a “high risk borrower”, predicting whether a person will commit a crime in the future).

Performance measures are relevant for monitoring discrimination because they can reveal gaps in how a system treats different groups of people. These gaps, or disparity in treatment or outcomes between groups, may be discriminatory for systems used in sectors regulated by non-discrimination law. They can show, for example, that a system is more likely to rank male applicants highly than female applicants, or Black pedestrians are more likely to be incorrectly identified and stopped by police than white pedestrians.⁷⁴

A variety of tests have been created in recent years that use performance measures to identify and mitigate biases in AI systems. Broadly, these tools can be separated into two types: fairness metrics that use statistics to measure performance gaps affecting individuals or groups of people, and fairness enforcement methods that correct bias by forcing a model to behave “fairly.” These methods are explored in the following two sections.

⁷⁴ ‘Automated Police Tech Contributes to UK Structural Racism Problem | Computer Weekly’ (*ComputerWeekly.com*) <<https://www.computerweekly.com/news/366603173/Automated-police-tech-contributes-to-UK-structural-racism-problem>> accessed 3 December 2024.



TABLE 1: TECHNICAL DOCUMENTATION REQUIREMENTS FOR HIGH-RISK AI SYSTEMS

AIA Technical documentation requirement (Annex IV)	Relevant JTC 21 harmonised technical standards for AI systems ⁷⁵	Components relevant to Equality Bodies
General description of the AI system, including: Intended purpose; Planned interactions with other AI systems; A list of software, firmware, and hardware used; Forms of deployment on the EU market; Documentation of physical products in which the AI system is embedded; Description of the user interface; Instructions for use for the deployer.	Risk management systems.	Model cards, datasheets.
Performance measures, including: Level of accuracy, including metrics, robustness, and cybersecurity measures used for testing and validation, and known or foreseeable circumstances that could impact these aspects; Known or foreseeable circumstances that pose a risk to health and safety or fundamental rights; Capabilities of the system to provide information to explain outputs; Performance for specific individuals and demographic groups.	Governance and quality of datasets; Transparency and information provisions for users; Accuracy specifications; Robustness specifications; Cybersecurity specifications.	Fairness metrics, model cards, datasheets, FRIAs and other impact assessments.
Data requirements including datasheets describing training methods, techniques, and datasets used in building the AI system, as well as the data's provenance, scope, characteristics, collection and selection, labelling, and cleaning.	Governance and quality of datasets.	Datasheets, debiasing methods for models and training, test, and validation datasets.
Information to aid in interpretation of system outputs by deployers.	Transparency and information provisions for users.	Model cards, datasheets fairness metrics.
Risk management system.	Risk management systems.	Risk register.
Planned changes to the system and its performance.	Quality management systems for providers of AI systems, including post-market monitoring processes.	Generally relevant.
Human oversight information.	Human oversight.	Generally relevant.
Required computational and hardware resources, including details of the system's anticipated lifetime and maintenance.	Quality management systems for providers of AI systems, including post-market monitoring processes.	Generally relevant.
Mechanisms to collect, store, and interpret event logs.	Record keeping through logging capabilities.	Generally relevant.
Post-market monitoring plan.	Quality management systems for providers of AI systems, including post-market monitoring processes.	Generally relevant.

⁷⁵ These categories of standards are taken from the initial standardisation request issued by the European Commission to CEN/CENELEC. Multiple harmonised standards may be developed or adapted within these categories. JTC-21's current work programme is available here: CEN/CENELEC (n 23).

Before discussing them, it is worth noting that both fairness metrics and enforcement methods have been implemented in a range of open-source fairness toolkits. Equality Bodies that have in-house technical expertise can use these toolkits to measure performance gaps in AI models or datasets made available to them. Some toolkits such as Fairness Measures, TensorFlow Fairness Indicators, FAT Forensics, and FairTest focus solely on measuring bias.⁷⁶ Other toolkits include methods for “debiasing” datasets and models, including Themis-ML, Aequitas, and OxonFair, the latter of which provides recommendations for matching tests to specific use cases.⁷⁷ By far the most popular toolkits are IBM AI Fairness 360 and Microsoft Fairlearn, but both have significant drawbacks and limitations,⁷⁸ meaning there is not a single “one size fits all” toolkit that should be used in practice by AI providers, deployers, or Equality Bodies.⁷⁹

4.1.2.1. Fairness metrics to statistically measure performance gaps

In recent years researchers have created many statistical measures to measure unwanted biases and the fairness of AI systems.⁸⁰ “Fairness metrics” are a mathematical formula to measure differences in performance affecting certain individuals or groups. Terminology in the field is inconsistent so these types of metrics are also sometimes called “bias tests.”

A variety of types of metrics have been proposed, including those that measure differences between individuals and groups, counterfactual measures (i.e., a measurement in which a person’s protected attribute is “flipped” to see the impact it has on the model), and fairness through unawareness which involves hiding sensitive attributes from the model to determine whether similar individuals receive similar outcomes when their only difference is a sensitive attribute (e.g., different genders).

By far the most popular approach to measuring fairness is a category of metrics called “group fairness.” These metrics compare performance between demographic groups according to one or more characteristics such as accuracy, recall, precision, or others (see: Box 8). They define a

⁷⁶ ‘Fairness Measures - Detecting Algorithmic Discrimination’ <<https://fairnessmeasures.github.io/>> accessed 23 November 2024; ‘Fairness Indicators | TFX’ (*TensorFlow*) <https://www.tensorflow.org/tfx/guide/fairness_indicators> accessed 23 November 2024; ‘FAT Forensics — FAT Forensics 0.1.2 Documentation’ <<https://fat-forensics.org/index.html>> accessed 23 November 2024; ‘Columbia/Fairtest’ <<https://github.com/columbia/fairtest>> accessed 23 November 2024.

⁷⁷ ‘A Fairness-Aware Machine Learning Library — Themis-ML 0.0.2 Documentation’ <<https://themis-ml.readthedocs.io/en/latest/>> accessed 23 November 2024; Pedro Saleiro and others, ‘Aequitas: A Bias and Fairness Audit Toolkit’ [2019] arXiv:1811.05577 [cs] <<http://arxiv.org/abs/1811.05577>> accessed 10 May 2022; Eoin Delaney and others, ‘OxonFair: A Flexible Toolkit for Algorithmic Fairness’ [arXiv, 5 November 2024] <<http://arxiv.org/abs/2407.13710>> accessed 17 November 2024.





⁷⁸ Rachel KE Bellamy and others, ‘AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias’ (2019) 63 *IBM Journal of Research and Development* 4: 1; Sarah Bird and others, ‘Fairlearn: A Toolkit for Assessing and Improving Fairness in AI’ [2020] Microsoft, Tech. Rep. MSR-TR-2020-32.

⁷⁹ These limitations are discussed here: Delaney and others (n 80).

⁸⁰ ISO 24027, Clause 7.1.

“fair” AI model as one in which a chosen performance gap(s) between groups are reduced or eliminated, typically while keeping the system as accurate as possible.⁸¹

Consider a hypothetical example: we are using an AI system for cancer screening. The system predicts whether a person currently has cancer. Each person can receive either a positive or negative prediction, meaning the system predicts they either have cancer (positive) or do not have cancer (negative). This prediction may or may not be correct, leading to four possible outcomes:

-  True positive: Patients with cancer correctly predicted to have cancer.
-  False positive: Healthy patients incorrectly predicted to have cancer.
-  True negative: Healthy patients correctly predicted to be healthy.
-  False negative: Patients with cancer incorrectly predicted to be healthy.

At their most basic, many group fairness metrics are simply different ways of balancing how frequently different groups of people experience true/false positives and true/false negatives measured against the “ground truth.” Here, “ground truth” means how things are in reality; for example, if a system is predicting cancer, the “ground truth” is whether or not the person being screened actually has cancer.⁸² A common way of presenting the system’s performance in this regard is with a “confusion matrix” which reports how frequently different groups of people experience true/false positives and true/false negatives.

⁸¹ Mittelstadt, Wachter and Russell (n 30). *ibid*.

⁸² Ground truth is a highly contested concept, and this is an oversimplistic definition. Ground truth data is not available when using AI to predict something about people, for example whether a person is likely to be good at a particular job, do well at university, or default on a loan. Reflecting this, it is often impossible to perfectly measure fairness because of a lack of data. For some classification tasks, “ground truth” data is available to validate results. A system used to predict the age of a person, for example, can be validated on data listing each person’s actual age. For others, “ground truth” data is impossible or incomplete, meaning the classifications or predictions cannot be validated against real world data. This is the case for many prediction tasks. Take for ex-ample a system used to predict a person’s future likelihood of repaying a loan: the predictions can be validated in the future for people who are actually given loans but cannot be validated for people that did not receive the loan (and thus never had a chance to repay it).

Measuring performance

Fairness metrics measure many types of performance characteristics, such as:

- **Accuracy:** The rate at which the model makes the correct prediction.
Example: A model that correctly labelled 6 out of 10 pictures of dogs as “Dog” would have 60% accuracy.
- **Recall:** The rate at which the model correctly predicts true positives.
Example: A cancer screening model would have perfect recall if it correctly identified all people who eventually developed cancer in the future.
- **Precision:** The rate at which the model correctly predicts true positives divided by the total number of positive predictions.
Example: A model that labelled 10 pictures as “Dog”, but only 7 pictures were of dogs and 3 were of cats would have a precision rate of 70%.

It would be impossible to review all known fairness metrics. Hundreds have been developed in recent years that focus on different performance characteristics or balance performance between groups in different ways.⁸³ When Equality Bodies receive information on fairness metrics, it would be useful to know why the AI provider chose to use particular metrics over others. Different metrics serve different purposes and answer different questions and can be used both to highlight and hide unfairness, or gaps in performance between groups (e.g., how more frequently incorrect predictions are made by an AI system about patients of a certain ethnicity). Different fairness metrics will have different justified uses, based for example on the type of harm or error we might be looking for, such as overtreatment or undertreatment of particular patient groups.⁸⁴ The AIA technical documentation requires providers to comment on the appropriateness of the chosen performance metrics, for example by drawing on harmonised standards.⁸⁵ In this context, it is helpful to review some of the most popular group fairness metrics to understand the different types of performance gaps they measure and under what conditions they can be justifiably used in practice (see: Table 2).

⁸³ Sahil Verma and Julia Rubin, ‘Fairness Definitions Explained’, *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (IEEE 2018); Sandra Wachter, Brent Mittelstadt and Chris Russell, ‘Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law’ [2021] 123 W. Va. L. Rev. 735. Verma and Rubin; Wachter, Mittelstadt and Russell, ‘Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law’.

⁸⁴ For more discussion of the types of harms targeted by different fairness metrics, see Mittelstadt, Wachter and Russell (n 30).

⁸⁵ AIA Annex IV 4.

4.1.2.2. Using fairness metrics to assess equality

Fairness metrics will be very useful to investigate discrimination under the AIA. Providers of high-risk systems are required to report evidence of performance gaps between protected groups at various points (see: Sections 4.1, 4.4 and 4.5). Fairness metrics can be used by Equality Bodies to determine the severity of these gaps and identify those which may constitute illegal disparity. They are primarily used for testing models to identify performance gaps between groups of people. They help identify potentially discriminatory gaps, for example in how often different demographic groups receive positive decisions, but do nothing to correct them. Any gap in performance between groups can indicate potential illegal disparity requiring further investigation.

TABLE 2 - COMPARISON OF GROUP FAIRNESS METRICS. ADAPTED WITH PERMISSION. ⁸⁶			
Fairness metric	Description	Justified use	Example
Demographic Parity	Different groups should receive positive outcomes at the same rate.	Situations where historic data is expected to be prejudicial, and there is no agreed upon ground-truth.	Hiring, offering loans, access to education, representation in the media.
Equal Opportunity	Different groups should experience false negatives at the same rate.	Situations where there is agreed up on ground-truth and the overwhelming harm comes from false negatives.	Cancer or other serious illness screening.
Predictive Parity	Different groups should have the same precision rate.	Situations where there is agreed up on ground-truth and the overwhelming harm comes from false positives.	Misidentification as a known person of interest to the police.
False positive error rate balance	Different groups should receive false positives at the same rate.	Situations where there is agreed up on ground-truth and the overwhelming harm comes from false positives.	Misidentification as a known person of interest to the police.
Equalized odds	True positive and false positive rates should be the same for different groups.	Combination of Equal Opportunity and False positive error rate balance.	Treatment of illness by performing risky surgery.
Overall accuracy equality	Accuracy rates are the same across different groups.	Situations where there is agreed up on ground-truth and the harm of misclassification is the same regardless of how people are situated.	Offering someone left- or right-handed scissors.

⁸⁶ Mittelstadt, Wachter and Russell (n 30). *ibid.*

Equality bodies will typically not be able to use fairness metrics directly, as doing so requires access to the AI system or a set of labelled model outputs (e.g., all of the positive and negative loan decisions made by a system). However, Equality Bodies can use test results reported in technical documentation to assess equality in specific use cases and for demographic groups and request additional testing with specific metrics when discrimination is suspected. To understand how, we must first look at the implementation of non-discrimination law to see (1) how disadvantaged and comparator groups are defined in practice, (2) how courts recommend groups be compared to identify *prima facie* discrimination, and (3) how these concepts and method can be translated for equality in AI.

Non-discrimination law distinguishes between direct and indirect discrimination. Direct discrimination occurs when an individual or group is treated unfairly explicitly due to a particular protected characteristic; for example, a rule to not admit female applicants to a particular university programme would be directly discriminatory. Indirect discrimination, on the other hand, occurs when a seemingly neutral rule, practice, or criterion nonetheless has a discriminatory impact on a particular demographic group. Intent is not required to prove indirect discrimination; rather, it is only required to demonstrate a substantial and disproportionate impact on a particular group, regardless of the intent of the discriminating individual or organisation.

Of the two types, indirect discrimination will be far more common in AI systems.⁸⁷ Helpfully, prior research has shown that statistical measures (see: Section 4.1.2.1) can be broadly divided between those that align with “formal equality” or equal treatment (i.e., “bias preserving” metrics) which are relevant for protection against direct discrimination, and those that align with “substantive equality” or levelling the playing field (i.e., “bias transforming” metrics) which are relevant for protection against indirect discrimination.⁸⁸

To bring a case of indirect discrimination under EU non-discrimination law a claimant must establish that *prima facie* discrimination has occurred. Evidence must be provided to show that (1) a particular harm has occurred or is likely to occur; (2) the harm manifests or is likely to

⁸⁷ Wachter, Mittelstadt and Russell, ‘Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law’ (n 86); Wachter, Mittelstadt and Russell, ‘Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI’ (n 32). Direct algorithmic discrimination will be simpler to identify in practice because it requires an AI system to explicitly use a protected characteristic to make a prediction or decision. It will also be less common, as protected characteristics are often removed from AI training datasets, or otherwise obscured to ensure they are not used explicitly by the trained model. Indirect algorithmic discrimination will be more common and yet much harder to identify because it requires a claimant to establish that a significant disproportionate impact has occurred to a particular protected group. AI systems are trained on features and find correlations between people and cases that are not human interpretable, meaning they will not map neatly onto legally protected characteristics. Instead, people claiming algorithmic discrimination has occurred will need to show that the impact in question has affected a protected group.

⁸⁸ Wachter, Mittelstadt and Russell, ‘Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law’ (n 86). In the simplest terms, bias preserving metrics take the status quo (i.e. the training data) as a neutral starting point from which to measure inequality, and seek only to not make things more unequal than is currently the case. If, for example, female applicants receive public housing 20% less often than male applicants, a bias preserving metric might seek only to ensure that the rate does not increase beyond 20%. In contrast, bias transforming metrics do not use the status quo as a neutral starting point and can thus help fix structural and historical inequalities faced by particular groups.



manifest significantly within a protected group of people; and (3) the harm is disproportionate when compared with others in a similar situation. If these requirements can be met the burden of proof shifts to the alleged offender, such as the AI deployer.⁸⁹ They can then provide counterevidence and arguments to show that the contested AI decision (i.e. the “contested rule or practice”) is actually justified, or otherwise refute the claim’s basis.⁹⁰

Defining the disadvantaged group(s) and comparator groups is a key aspect of bringing a case. These groups are typically defined by protected traits such as age, ethnicity, gender, sexual orientation or religious beliefs, disability, or others. Critically, they can be defined in a broad (e.g., all female applicants) or narrow way (e.g., female applicants based in London), and may be intersectional (e.g., Black female applicants under the age of 40).⁹¹ Discrimination affecting broad, narrow, or intersectional groups can disappear if the size or scope of the groups being compared changes; imagine, for example, discrimination experienced by Black female applicants which disappears when focusing on Black or female applicants only. Equality bodies should be vigilant about how AI providers define the groups they are measuring as reported in the system’s technical documentation; single characteristic groups (e.g., women) are often the default, which may hide illegal disparity affecting only intersectional groups (e.g. Black women).⁹²

With test results in hand and appropriate groups defined, the question then becomes how to use the results of fairness metrics to determine whether an AI system is causing discrimination. The European Court of Justice (ECJ) has previously described their “gold standard” approach to assessing *prima facie* discrimination. There are three general options to assess whether a particular rule, practice, or decisions made by an AI model is discriminatory: examine its effect on (1) the disadvantaged group, (2) the comparator (advantaged) group, or (3) both groups. The ECJ’s preferred approach is the third option.⁹³

Assuming sufficient and broad ranging evidence about performance gaps impacting specific groups have been provided, Equality Bodies can carry out the analysis described above. However, Equality Bodies may find that the evidence provided is deficient because it (1) does

⁸⁹ Wachter, Mittelstadt and Russell, ‘Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI’ (n 32).

⁹⁰ Lilla Farkas and others, ‘Reversing the Burden of Proof: Practical Dilemmas at the European and National Level’ (Publications Office of the European Union 2015) 9 <<http://dx.publications.europa.eu/10.2838/05358>> accessed 9 February 2020.

⁹¹ Timo Makkonen, *Measuring Discrimination Data Collection and EU Equality Law* (Office for Official Publications of the European Communities 2007) 36.

⁹² Wachter, Mittelstadt and Russell, ‘Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI’ (n 32). *ibid.*

⁹³ The ECJ describes its “gold standard” approach to assessing potential discrimination in *Seymour-Smith*: “the best approach to the comparison of statistics is to consider, on the one hand, the respective proportions of men in the workforce able to satisfy the requirement of two years” employment under the disputed rule and of those unable to do so, and, on the other, to compare those proportions as regards women in the workforce. It is not sufficient to consider the number of persons affected, since that depends on the number of working people in the Member State as a whole as well as the percentages of men and women employed in that State.” See: *Regina v Secretary of State for Employment, ex parte Nicole Seymour-Smith and Laura Perez* 1999 E.C.R. I-60 [59].



not report the results of multiple fairness metrics reporting on different performance gaps (e.g., precision, recall, error rates) or (2) narrowly defines impacted groups meaning intersectional discrimination can be hidden. In such a case, they can make a request to the relevant national MSA to carry out further specific tests with additional fairness metrics or groups. Further, it may be possible for Equality Bodies to participate fully in any evaluation carried out by the MSA after the request has been made.⁹⁴

4.1.2.3. Fairness enforcement methods to make AI systems behave “fairly”

A variety of fairness enforcement methods have been proposed to fix the performance gaps identified by fairness metrics. They do so by enforcing a particular fairness metric on the model. In practice, this means changing how the model makes predictions, assigns labels, or classifies cases, or hiding protected attributes from it entirely (i.e., “fairness through unawareness”)⁹⁵; for example, if we want to ensure men and women receive positive loan decisions at the same rate in the interest of equality (i.e., “Demographic Parity”), we could force our model to give out positive decisions equally between these groups. This change would likely reduce the overall accuracy of the model which can be harmful in itself.

Equality bodies should thus be aware that using fairness enforcement methods comes at a cost, and often increases avoidable harms to health, safety, and fundamental rights while reducing a system’s overall accuracy.⁹⁶ Fairness metrics translate the concept of equality in simplistic terms to mean simply reducing or eliminating performance gaps between groups of people. The easiest way to eliminate such gaps is to “level down,” or make the system perform equally badly for all affected groups.⁹⁷

This approach can needlessly harm people by reducing the system’s accuracy. In the case of cancer screening, for example, it would mean misdiagnosing more cases of cancer than is strictly necessary solely to ensure all groups of patients are diagnosed with the same accuracy. Imagine our cancer screening system is more accurate for men than women. If we want to make the “recall” rate even between men and women (see: Box 8) while keeping accuracy as high as possible, which means we would catch more cases of cancer in women than would otherwise be the case, we can “level down.” This means forcing the model to be more likely to diagnose women as having cancer and men as not having cancer, even if it has low confidence in its prediction. By doing so we can balance the recall rate between men and women

⁹⁴ AIA Article 79[2].

⁹⁵ Cynthia Dwork and others, ‘Fairness through Awareness’, *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (ACM 2012).

⁹⁶ For example, in the case of hiring the harm would be lower overall hiring rates, or in the case of cancer detection, an increased failure to correctly identify people who have cancer.

⁹⁷ Mittelstadt, Wachter and Russell (n 30).



while losing as little accuracy as possible. The cost is that we harm more male patients by misdiagnosing them more frequently than would have otherwise been the case.⁹⁸

While levelling down problematically ignores the contextual and non-quantifiable aspects of equality and pursues equal treatment while failing to achieve substantive equality, it is nonetheless an appealing approach for AI providers and deployers. It can “solve” fairness problems without needing any additional resources or data.⁹⁹ To protect citizens, Equality Bodies should be vigilant in identifying such “cheap” or “simple” solutions to questions of equality in AI.

Thankfully, in recognition of these challenges, the usage of fairness enforcement methods in real world uses of AI is relatively rare. Nonetheless, if an AI provider reports that they are using methods to enforce fairness on their model, Equality Bodies should be ready to investigate the costs and harms of doing so to determine whether levelling down has occurred. Certain fairness toolkits such as OxonFair can help identify levelling down and enforce fairness through “levelling up” as an alternative, which prevents introducing avoidable harms in the pursuit of equality in AI.¹⁰⁰

4.1.3. Understanding bias in data requirements

Data requirements reported in the technical documentation are essential to enable Equality Bodies to effectively investigate algorithmic discrimination. Bias in data needs to be addressed throughout an AI system’s lifecycle, from the inception stage where organisational needs are identified, through “design and development, verification and validation, to operations and retirement.”¹⁰¹ In the context of the AIA, organisations procuring AI systems or models from third party providers have an interest in understanding how bias has been addressed throughout the pre-usage lifecycle. Information about their approach to bias measurement and mitigation may be reported in relevant model cards or datasheets, as well as the instructions for use and technical documentation created by the AI provider. Overlap should thus be expected between information on performance measures and data requirements.

Figure 1 shows a simplified sequential view of an AI system’s lifecycle. In principle AI providers can examine and correct for data bias at each phase of the lifecycle, but in practice their investigation may be limited to one or more phases. Equality Bodies should be aware of the types of information and testing that may be produced at each stage and reported in the system’s technical documentation, and how this information can help investigate AI biases and discrimination arising from the data used to build and run high-risk AI systems.

⁹⁸ For an in-depth discussion of this phenomenon and alternative solutions, see Mittelstadt, Wachter and Russell (n 30).

⁹⁹ *ibid.*

¹⁰⁰ Delaney and others (n 80).

¹⁰¹ ISO 24027, Clause 8.1.





Figure 1 - A simplified sequential view of the AI system lifecycle. In reality, these phases are iterative rather than sequential. Bias can be introduced, measured, and mitigated at all phases.

In the inception phase (Phase 1), providers can assess potential data biases in the earliest phases of a system’s development. Specifically, decisions about the requirements the system should fulfil, the resources that should be used to build it (e.g., datasets, algorithms), and its “intended use” can all be biased. The types of information produced in this phase that will be relevant to Equality Bodies include model cards and datasheets that document known biases in selecting appropriate data sources for completeness, accuracy, timeliness, consistency, and representativeness; methods and criteria to be used by the provider for labelling, cleaning, and filling in missing data; choice of training, testing, and validation datasets; and choice of algorithm to train any models.

Design and development (Phase 2) focuses on building the system. Once choices are made about which datasets, models, or algorithms to use, decisions must still be made about how precisely to use these to build the system. For example, feature engineering, or the process through which training data are transformed into features that can be processed by an AI system, involves key decisions about how the model will learn from the chosen datasets. For example, which specific features or characteristics included in the dataset should be used? How should “good” and “bad” individual datapoints be distinguished in practice?¹⁰² How should incomplete datapoints be handled? For Equality Bodies, answers to these and similar questions will likewise be reported in model cards, datasheets, and other documentation, and can indicate potential performance gaps and biases if, for example, incomplete records are more common among certain groups in the dataset.

¹⁰² ISO 24027 Clause 8.3.2.1. Most modern AI systems require substantial volumes of labelled data for training purposes. A label is a piece of information appended to a specific record which tells the model what type of example it is (e.g., picture of a dog). This data is created at scale by data annotators or “ghost workers” who themselves introduce a range of human cognitive biases. Developers also introduce data biases by choosing how the labelling should be performed and according to which criteria. See: Mary L Gray and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Eamon Dolan Books 2019) <<https://books.google.com/books?hl=en&lr=&id=8AmXDwAAQBAJ&oi=fnd&pg=PP1&dq=ghost+workers+AI&ots=WVJ-OQ0Q3p&sig=mFhltwsngVLPuwvSByJvJlwwUw>> accessed 22 November 2024; Will Hawkins and Brent Mittelstadt, ‘The Ethical Ambiguity of AI Data Enrichment: Measuring Gaps in Research Ethics Norms and Practices’, *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2023) <<https://dl.acm.org/doi/10.1145/3593013.3593995>> accessed 14 August 2023.

Training and tuning (Phase 3) can involve removal of unwanted biases from training data or changes to how the model uses specific datapoints or features to reduce biased behaviour. The simplest approach to debias a system is by removing relevant protected attributes (e.g., gender, ethnicity) from the training data. This technique is unreliable because it does not remove proxy features that are strong indicators of protected attributes, such as home ownership or music taste as proxies for age.¹⁰³ Proxy variables mean protected attributes (e.g., gender, ethnicity) can still influence the predictions or decisions made by a system even when the protected attribute is not recorded or available to the system. Other methods target how we choose which data and features to use. A variety of “re-weighting” methods can raise or lower the importance of certain records in the training dataset to reduce the significance of gaps, for example when we have less or worse data from specific age or ethnicity groups. Equality Bodies should be aware of any specific methods used to debias datasets or models, as these may be used by providers to reduce potentially discriminatory performance gaps or promote positive equality by improving system performance for underrepresented or marginalised groups.

Verification and validation (Phase 4) investigate potential defects and biases in a system before deployment. Validation is performed by setting aside some data during the training and tuning phase, what is called a “hold-out” or validation dataset, meaning it is data the system has not seen before. This “hold-out” data is then used to test the generalizability and robustness of the system once built and identify implicit biases that were previously missed in earlier phases. Equality Bodies should pay attention to whether validation occurred, which data it used, and who performed it. Ideally, validation will involve adversarial testing with a variety of stakeholders that will be affected by the system to identify the broadest possible range of biases and performance gaps.

Deployment (Phase 5) involves making deployers, users, and people affected by AI systems aware of its limitations and biases through documentation and training. In practice this can occur by sharing “data requirements” elements of the Article 11 technical documentation with relevant stakeholders (see: Section 4.1.1). Additionally, the AIA requires instructions for use to be prepared by AI providers and given to deployers. Ideally, these instructions will include information about the performance of the system across a variety of use scenarios, and will make clear any known performance gaps, biases, or limitations that could affect real-world performance and lead to discrimination against specific groups of people affected by the system. All these aspects of deployment documentation will be directly and self-evidently relevant to Equality Bodies investigating AI bias and discrimination.

¹⁰³ Anupam Datta and others, ‘Proxy Non-Discrimination in Data-Driven Systems’ [2017] arXiv preprint arXiv:1707.08120 <<https://arxiv.org/abs/1707.08120>> accessed 22 September 2017.

4.2. Technical documentation for general-purpose AI models

Independent of the requirements discussed above for high-risk AI systems, providers of general-purpose AI models (GPAI) and general-purpose AI models with systemic risks must also draw up separate technical documentation.¹⁰⁴ Requirements are described in Article 53 and Annex XI of the AIA. It must be noted that at this stage of implementation of the AIA it is unclear whether Equality Bodies and NFRAs will have access to documentation for GPAI models. The documentation will be handled by the AI Office and national competent authorities rather than the MSAs that facilitate access to documentation for high-risk systems.¹⁰⁵

GPAI models are large models capable of competence performance on a wide range of distinct tasks. They are distinct from AI systems that can perform well only on a narrow range of well-defined tasks. Examples include generative AI systems such as ChatGPT, Google Gemini, Stable Diffusion, or DALL-E which can generate text, images, video, or audio from text prompts.

GPAI technical documentation is designed to explain their capabilities and limitations for providers planning to integrate the model in an AI system. It needs to include information on “methods to detect identifiable biases” used by the model provider.¹⁰⁶ In this context it is important to note the distinction made in the AIA between providers of (1) general-purpose AI models and (2) high-risk AI systems, and AI deployers, each of whom face different requirements under the law. Understood hierarchically in terms of size and scope, general-purpose models can be integrated into high-risk AI systems, which are then deployed in specific use cases (see: Figure 2).

As with high-risk AI systems, Equality Bodies should pay special attention to documentation on data requirements. Many GPAI models are trained on large corpuses

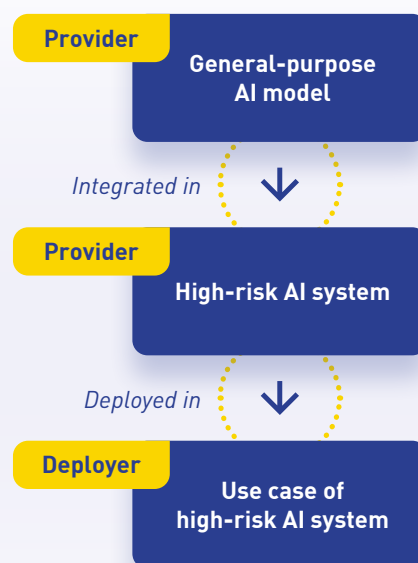


Figure 2 - Relationship between providers and deployers, and AI models and systems.

¹⁰⁴ This documentation must be provided upon request to the AI Office and national competent authorities (i.e., market surveillance authorities and notifying authorities).

¹⁰⁵ AIA Article 53.

¹⁰⁶ AIA Annex XI(2)(c).

BOX 9

Technical documentation of GPAI models

- General description of the general-purpose AI model, including:
 - » Intended purpose and types of AI systems it can be integrated in;
 - » Acceptable use policies;
 - » Date of release and methods of distribution;
 - » Architecture and number of parameters;
 - » Modalities (e.g., text, image, audio) of its inputs and outputs;
 - » Licence (e.g., open-source, proprietary);
- Technical means required for integration into an AI system;
- Design specifications of the model and training process;
- Data requirements (as above);
- Computational resources used for training;
- Known or estimated energy consumption.

of text from the open Internet containing a broad variety of problematic biases. Information about data sources can be helpful in identifying potential biases, and linking specific models to prior research on bias, disinformation, and similar topics on particular Internet platforms (e.g., Reddit, Google, Twitter). The impact of these biases can be seen, for example, in generated text that features hate speech or biased language (e.g., associating certain jobs with gendered language),¹⁰⁷ or generated images displaying gender or ethnicity biases (e.g., associating “beauty” with light skin tones).¹⁰⁸

Technical documentation for GPAI models with systemic risks includes further requirements to describe the evaluation strategies and criteria, measures in place for adversarial testing, and a detailed description of the system architecture.¹⁰⁹ This information may differ substantially from technical documentation provided for high-risk AI systems because GPAI models often perform fundamentally different tasks which are affected by different types of biases. Whereas traditional predictive AI systems are often used to *classify* cases, apply *labels* to data, or make *predictions* about people, GPAI models are often used to *generate* content such as text, images, audio, or video. Bias in this context may thus be

less about the *distribution* of resources or errors between different groups of people, and more about their *representation* in the generated text, audio, image, video, or other formats, for example whether certain groups described using gendered language, outdated stereotypes, or

¹⁰⁷ Aylin Caliskan, Joanna J Bryson and Arvind Narayanan, ‘Semantics Derived Automatically from Language Corpora Contain Human-like Biases’ [2017] 356 Science 183; Xiao Fang and others, ‘Bias of AI-Generated Content: An Examination of News Produced by Large Language Models’ [2024] 14 Scientific Reports 5224. Caliskan, Bryson and Narayanan; Fang and others.

¹⁰⁸ Federico Bianchi and others, ‘Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale’, *2023 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2023) <<https://dl.acm.org/doi/10.1145/3593013.3594095>> accessed 17 November 2024. *ibid.*

¹⁰⁹ GPAI models with systemic risks are defined according to their computational resources. Models requiring 10²⁵ floating point operations (or FLOPs) for training are deemed to have systemic risks. The European Commission can also classify GPAI models as having systemic risks independently of this requirement (AIA Article 51).

social prejudices. Theories to understand how bias emerges in GPAI models, as well as methods to measure and mitigate it, are still at a relatively early stage of development.¹¹⁰

Regardless, Equality Bodies may be particularly interested in the adversarial testing details which describe how the system was tested to identify when, how, and why it fails or produces biased or harmful outputs (e.g., hate speech, incorrect answers, non-human language). Methods and standards for adversarial testing of GPAI models are still in a relatively early stage of development, but Equality Bodies should ideally stay abreast of developments in this field to assess whether the testing or “red teaming” approach used by providers is robust.

4.3. Fundamental rights impact assessments

The fundamental rights impact assessment (FRIA) is another highly relevant documentation requirement for Equality Bodies. According to Article 27 of the AIA, deployers of certain high-risk AI systems are required to carry out a fundamental rights impact assessment (FRIA) prior to the system being placed onto the EU market.¹¹¹ FRIAs describe (1) the categories of people affected by the system, (2) the specific risks they face of harm to their fundamental rights, and (3) the measures taken by the deployer to identify and mitigate those risks.¹¹² Anticipated harms to equality and non-discrimination are among those that should be identified.

FRIAs are required of deployers, not providers. The impacts identified may thus differ from those identified by the provider through Conformity Assessment and as reported in a system’s technical documentation. In particular, it is assumed that deployers will better be able to account for contextual and case-specific factors posing potential impact on fundamental rights which cannot be captured by providers in the development phase.¹¹³

It is worth noting that the range of AI deployers required to conduct FRIAs is very limited.¹¹⁴ FRIAs only apply to deployers involved in the public sector.¹¹⁵ Companies in the private sector are excluded, except in cases where they are (1) providing a public service, (2) creditworthiness checks, or (3) risk assessment or pricing for life or health insurance.¹¹⁶ AI for critical infrastructure is also excluded.¹¹⁷ The range of deployers which must carry out a FRIA is

¹¹⁰ Philipp Hacker and others, ‘Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It’ (arXiv, 26 June 2024) <<http://arxiv.org/abs/2407.10329>> accessed 3 December 2024.

¹¹¹ Mantelero (n 1).

¹¹² AIA Article 27(1).

¹¹³ Mantelero (n 1). *ibid.*

¹¹⁴ Mantelero (n 1).

¹¹⁵ Sandra Wachter, ‘Limitations and Loopholes in the E.U. AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond’ (2024) 26 *Yale Journal of Law and Technology*.

¹¹⁶ Specifically, “deployers that are bodies governed by public law, or are private entities providing public services, and deployers of high-risk AI systems referred to in points 5 (b) and (c) of Annex III.” AIA Article 27(1).

¹¹⁷ Article 27(1).

thus rather limited, excluding key sectors where pre-existing biases and inequalities are widely documented such as employment, education, and much of the financial sector.¹¹⁸

A further limitation is also worth mentioning. Even in cases where deployers would normally be required to carry out a FRIA, they can avoid this obligation if another actor (i.e., the system provider or another deployer) has already carried out a FRIA (or equivalent impact assessment) on the system. This may be the case, for example, where a provider has conducted a self-assessment to argue that their system should not be classified as high-risk, even if it falls within the scope of high-risk systems listed in Annex III.¹¹⁹ The range of cases in which FRIAs will be conducted may thus be very limited in practice.¹²⁰

Following standard procedures for human rights impact assessment (HRIA) and risk management, FRIAs will need to include at least three phases: “(i) a planning and scoping phase, focusing on the main characteristics of the product/service and the context in which it will be placed; (ii) a data collection and risk analysis phase, identifying potential risks and estimating their potential impact on fundamental rights; (iii) a risk management phase, adopting appropriate measures to prevent or mitigate these risks and testing their effectiveness.”¹²¹ FRIAs are not intended to only describe the risks posed by the system, but rather show (with evidence) how the measures proposed to mitigate them will be effective in practice.¹²²

Even though conformity assessments carried out by AI providers also address a system’s potential impact on fundamental rights, they are distinct from FRIAs carried out by deployers. For one, providers and deployers will have different contextual knowledge about the proposed use case and environment, and thus may come to different conclusions about the nature and degree of risks to fundamental rights. Conformity assessment must also follow harmonised standards established by CEN/CENELEC, whereas FRIAs are intended to be self-guided and require reporting to MSAs based on a template questionnaire to be developed by the AI Office.¹²³

The expected effort and depth to conduct a FRIA appears to be low. The AI Office has been tasked with developing a FRIA template consisting of a questionnaire that can be implemented in an automated tool. This approach raises concerns that, as has happened historically with many other impact assessment and ethics procedures, the FRIA may become a mere checklist “tick box” exercise rather than an in-depth, context-sensitive, critical form of assessment.¹²⁴

¹¹⁸ Wachter, ‘Limitations and Loopholes in the E.U. AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond’ (n 121).

¹¹⁹ AIA Article 6(3).

¹²⁰ This is a significant enforcement loophole that grants AI providers substantial power to avoid the requirements of the AI Act. See: Wachter, ‘Limitations and Loopholes in the E.U. AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond’ (n 121).

¹²¹ Mantelero (n 1) 9.

¹²² *ibid.*

¹²³ Mantelero (n 1).

¹²⁴ *ibid.*; T Pöder and T Lukki, ‘A Critical Review of Checklist-Based Evaluation of Environmental Impact Statements’ (2011) 29 *Impact Assessment and Project Appraisal* 27; Brent Mittelstadt, ‘Principles Alone Cannot Guarantee Ethical AI’ (2019) 1 *Nature Machine Intelligence* 501.

Equality bodies appear to have access to the results of FRIAs. Deployers must report results of FRIAs to relevant MSAs, from whom Equality Bodies and other NFRAs can request access to any documentation prepared within the scope of the AIA (see: Section 4).¹²⁵ FRIAs will be valuable for Equality Bodies because they will provide information about proposed real-world uses of the system, including biases and potentially discriminatory performance gaps that deployers assess as likely to arise in practice. Deployers are duty-bearers under non-discrimination law, so FRIAs are the quickest way for Equality Bodies to understand how AI systems are likely to be used by organisations that fall within their remit as NFRAs. However, a key limitation must be noted. FRIAs are conducted before AI systems are deployed, meaning they will not provide evidence of how fundamental rights are impacted “in practice” once a system is being used. Rather, post-deployment impact is intended to be measured through post-market surveillance.

4.4. Post-market surveillance

Article 72 of the AIA stipulates that providers of high-risk AI systems are required to conduct post-market monitoring to identify risks and harmful impacts of their AI products and services once they have been deployed on the market. Post-market monitoring originates in medicine and pharmaceuticals where the aim is to assess the long-term efficacy of a product on patient groups and identify any emergent risks, impacts on health, and unforeseen interactions. In the context of AI, post-market monitoring will involve gathering data about the performance and safety of AI products and services after being placed on the market to assess their long-term compliance with AIA requirements and identify any emergent risks, including new or reinforced biases against specific groups of people.

This type of monitoring is extremely important for Equality Bodies. Providers must report adverse events, such as harms to fundamental rights, health, and safety, to their national MSAs. Article 73 requires MSAs to subsequently inform Equality Bodies and other NFRAs listed under Article 77. This means that Equality Bodies will receive up-to-date information about actual harms created by high-risk AI systems on the EU market. This information will be essential for Equality Bodies to identify use cases of AI requiring further investigation to determine whether a particular use case or adverse event amounts to discrimination under non-discrimination law.

4.5. Risk management system

There is one final documentation requirement relevant to Equality Bodies. Article 9 of the AIA requires providers of high-risk AI systems to implement a risk management system, understood as “a continuous iterative process planned and run throughout the entire lifecycle.” Risk management seeks to identify, analyse, evaluate the likelihood, and mitigate known

¹²⁵ AIA Article 27(3).

and foreseeable risks for health, safety, or fundamental rights.¹²⁶ This analysis covers the system’s entire lifecycle, meaning it also considers information gathered through post-market surveillance. Risk management is an umbrella process covering the following aspects of AI system governance:

- Technical documentation for high-risk AI systems, including instructions for use by deployers (Article 11 and Annex IV);
- Technical documentation for general-purpose AI models (Article 53 and Annex XI);
- Fundamental rights impact assessments (Article 27);
- Post-market monitoring data;
- Incident reporting;
- Evidence in support of conformity assessments (Chapter 3, Section 5).

Equality bodies should be particularly interested in the testing aspects of the risk management system. Once risks have been identified by a provider, any proposed management strategy must first be tested before the system is placed onto the market.¹²⁷ In cases where biases and discriminatory outputs are identified by providers, fairness and debiasing methods will likely be among the proposed mitigations. Equality Bodies should request results of bias tests and fairness mitigations under their right to access documentation (see: Section 4). This may prove to be invaluable information to evaluate whether *prima facie* discrimination has occurred in particular uses of high-risk AI systems.

Risk management requirements are relevant to Equality Bodies because risk management processes will monitor how risks to fundamental rights, health, and safety emerge and change post-deployment. Documentation related to risk management systems will be accessible by Equality Bodies. Assessment of bias across a system’s lifecycle is essential given the possibility of feedback loops and biases emerging or being reinforced as systems learn and change over the course of their deployment. New risks can emerge over time, and risk management is the process through which they are identified, categorized, and communicated to NFRAs.

¹²⁶ AIA Article 9. The scope of known or foreseeable risks is limited by the actions available to providers when building their systems. They are defined as risks which “may be reasonably mitigated or eliminated through the development or design of the high-risk AI system, or the provision of adequate technical information” (AIA Article 9(3)).

¹²⁷ AIA Article 9(8).



5. Open challenges for Equality Bodies

How to use the Artificial Intelligence Act to
investigate AI bias and discrimination: A guide
for Equality Bodies



Equality bodies seeking to use technical and organisational tools to measure algorithmic bias and discrimination face a variety of challenges and ambiguities which cannot be resolved directly by technical standards. These challenges nonetheless indicate future directions of valuable work to be undertaken by Equality Bodies to ensure discrimination can be consistently detected in high-risk AI systems.

5.1. Ambiguity in thresholds

While concepts such as direct and indirect discrimination are well-established and have clear precedents in case law, other concepts and details crucial to establishing the existence of *prima facie* discrimination and proving illegal discrimination are far more ambiguous and difficult to quantify. Thresholds to measure the “particular disadvantage” suffered by individuals and groups are rarely explicitly set by the judiciary in non-discrimination case law. Even when defined, they tend to be ambiguous, measured on a case-by-case basis, and dependent on intuition or rough measures.¹²⁸ Imprecise phrases have historically been used to describe illegal disparity between comparator and disadvantaged groups, such as “considerably more,”¹²⁹ “far more,”¹³⁰ “far greater number,”¹³¹ “almost exclusively women,” “significantly greater proportion

¹²⁸ European Union Agency for Fundamental Rights and Council of Europe, *Handbook on European Non-Discrimination Law* (2018 edition, Publications Office of the European Union 2018) 242–243 <https://fra.europa.eu/sites/default/files/fra_uploads/1510-fra-case-law-handbook_en.pdf>. *ibid.*

¹²⁹ *Z v A Government department, The Board of management of a community school* 2014 E.C.R. I–159 [53] In this case the Court held that “[t]he Court has consistently held that indirect discrimination on grounds of sex arises where a national measure, albeit formulated in neutral terms, puts considerably more workers of one sex at a disadvantage than the other.» The Court also cited the following cases in support; *Hellen Gerster v Freistaat Bayern* 1997 E.C.R. I-05253 [30]; *Waltraud Brachner v Pensionsversicherungsanstalt* 2011 E.C.R. I-10003 [56]; *Nadežda Riežņiece v Zemkopības ministrija and Lauku atbalsta dienests* 2013 ECLI:EU:C:2013:410 [39]. *Z. v A Government department, The Board of management of a community school* para 53 In this case the Court held that “[t]he Court has consistently held that indirect discrimination on grounds of sex arises where a national measure, albeit formulated in neutral terms, puts considerably more workers of one sex at a disadvantage than the other.» The Court also cited the following cases in support; *Hellen Gerster v Freistaat Bayern* para 30; *Waltraud Brachner v Pensionsversicherungsanstalt* para 56; *Nadežda Riežņiece v Zemkopības ministrija and Lauku atbalsta dienests* para 39.

¹³⁰ *Lourdes Cachaldora Fernández v Instituto Nacional de la Seguridad Social (INSS) and Tesorería General de la Seguridad Social (TGSS)* 2015 ECLI:EU:C:2015:215 [28]; *Waltraud Brachner v Pensionsversicherungsanstalt* (n 135) para 56 and the case law cited; *Isabel Elbal Moreno v Instituto Nacional de la Seguridad Social (INSS), Tesorería General de la Seguridad Social (TGSS)* 2012 EU:C:2012:746 [29]. *Lourdes Cachaldora Fernández v Instituto Nacional de la Seguridad Social (INSS) and Tesorería General de la Seguridad Social (TGSS)* para 28; *Waltraud Brachner v Pensionsversicherungsanstalt* (n 135) para 56 and the case law cited; *Isabel Elbal Moreno v Instituto Nacional de la Seguridad Social (INSS), Tesorería General de la Seguridad Social (TGSS)* para 29.

¹³¹ *Bilka - Kaufhaus GmbH v Karin Weber von Hartz* 1986 E.C.R. I–204; *Debra Allonby v Accrington & Rossendale College, Education Lecturing Services, trading as Protocol Professional and Secretary of State for Education and Employment* 2004 E.C.R. I–00873; *Ingrid Rinner-Kühn v FWW Spezial-Gebäudereinigung GmbH & Co KG* 1989 E.C.R. I-02743. *Bilka - Kaufhaus GmbH v Karin Weber von Hartz*; *Debra Allonby v Accrington & Rossendale College, Education Lecturing Services, trading as Protocol Professional and Secretary of State for Education and Employment*; *Ingrid Rinner-Kühn v FWW Spezial-Gebäudereinigung GmbH & Co. KG.*

of individuals of one sex as compared with individuals of the other sex,”¹³² and similar ambiguous terms.¹³³ Attempting to translate these terms into quantified thresholds, or to match them with the results of fairness metrics and other bias tests (see: Section 0), is difficult if not impossible.

Recognising this, Equality Bodies should not seek to find a “one size fits all” set of fairness and bias tests to be used for all high-risk AI systems. Equality is a highly contextual concept, and measurements of bias, fairness, and discrimination in AI need to be similarly flexible. Rather, Equality Bodies should seek to ensure, through their rights of access and testing (see: Section 4), that high-risk AI providers are (1) searching for performance gaps affecting a wide range of broad, narrow, and intersectional groups (see: Section 4.1.2.1), and (2) using a range of fairness metrics that address different performance characteristics (see: Section 4.1.2.2). Equality bodies should see their role as ensuring high-risk AI providers are held accountable for producing consistent, high-quality, and wide-ranging evidence about performance disparities in their systems, rather than seeking to give bias, fairness, or discrimination the same substantive meaning in all cases.¹³⁴

5.2. Using statistical evidence in legal cases

As these examples suggest, Equality Bodies will need to contend with the judiciary’s historical reliance on intuition in measuring discrimination when considering whether to bring a legal case on behalf of a complainant. Statistical evidence created using technical tools which reveals a gap in performance or outcomes between demographic groups can underpin the sorts of ambiguous thresholds used historically. It can also help reverse the burden of proof from the claimant to the accused party to show that the disparity in question was not illegal.

While direct discrimination cases require claimants to show they were personally treated unfavourably due to a protected attribute, indirect discrimination can be shown through statistical evidence in cases where protected attributes are not explicitly used in the contested rule or model.¹³⁵ Statistical evidence can be crucial to demonstrating correlations between protected attributes and the contested rule or model (e.g., establishing a particular set of attributes used by a model as a proxy for a protected attribute). Statistical evidence

¹³² *Violeta Villar Láiz v Instituto Nacional de la Seguridad Social (INSS) and Tesorería General de la Seguridad Social* 2019 ECLI:EU:C:2019:382 [38]; *Lourdes Cachaldora Fernández v Instituto Nacional de la Seguridad Social (INSS) and Tesorería General de la Seguridad Social (TGSS)* (n 136) para 28 as well as the cited case law. *Violeta Villar Láiz v Instituto Nacional de la Seguridad Social (INSS) and Tesorería General de la Seguridad Social para 38*; *Lourdes Cachaldora Fernández v Instituto Nacional de la Seguridad Social (INSS) and Tesorería General de la Seguridad Social (TGSS)* (n 136) para 28 as well as the cited case law.

¹³³ For a full list of terms used in place of quantified thresholds, see: Wachter, Mittelstadt and Russell, ‘Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI’ (n 32).

¹³⁴ For an overview of using fairness testing for consistent assessment rather than interpretation of equality, see: *ibid.*

¹³⁵ Makkonen (n 95) 31. *ibid.*

can also support direct discrimination claims when proving an unlawful pattern, such as a company's refusal to hire people of a certain ethnicity despite their significant presence in the population.¹³⁶ Thankfully, legal systems generally accept statistical evidence for questions of discrimination. However, it should be noted that the admission of particular statistical evidence cannot be guaranteed in all cases,¹³⁷ seen for example in prior claims involving unequal pay by gender, age-related redundancy, and racial segregation.¹³⁸

5.3. Algorithmic discrimination is unintuitive and remote

AI complicates establishing *prima facie* discrimination as claimants must experience or anticipate inequality, which becomes more subtle and intangible with automated systems.¹³⁹ Unlike traditional discrimination, automated discrimination is harder to detect and prove, as victims may not realize they have been disadvantaged.¹⁴⁰ The ability to compare experiences, such as job promotions or pricing, is reduced in an algorithmic context, making it difficult for individuals to recognize unfair treatment.¹⁴¹

Although discrimination may not be directly felt, discriminatory practices can persist, and obtaining evidence becomes challenging, especially when system controllers restrict access to protect intellectual property or avoid lawsuits.¹⁴² Explicit and implicit biases embedded in the data used to train and run AI models may not evoke an intuitive feeling of inequality, inhibiting

¹³⁶ Makkonen (n 95) 32. *ibid.*

¹³⁷ Makkonen (n 95) 30. *ibid.*

¹³⁸ Lilla Farkas and Declain O'Dempsey, 'How to Present a Discrimination Claim: Handbook on Seeking Remedies under the EU Non-Discrimination Directives' (Publ Off of the Europ Union 2011) 49. *ibid.*

¹³⁹ Brent Mittelstadt and others, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3 *Big Data & Society* <<http://bds.sagepub.com/lookup/doi/10.1177/2053951716679679>> accessed 15 December 2016; Sandra Wachter, 'Affinity Profiling and Discrimination by Association in Online Behavioural Advertising' (2020) 35 *Berkeley Technology Law Journal* 42–43, 45–46 <<https://papers.ssrn.com/abstract=3388639>> accessed 9 February 2020; Tal Zarsky, 'The Trouble with Algorithmic Decisions An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making' (2016) 41 *Science, Technology & Human Values* 118. Mittelstadt and others; Wachter, 'Affinity Profiling and Discrimination by Association in Online Behavioural Advertising' 42–43, 45–46; Zarsky.

¹⁴⁰ Wachter, 'Affinity Profiling and Discrimination by Association in Online Behavioural Advertising' (n 145); Jenna Burrell, 'How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms' [2016] *Big Data & Society*. Wachter, 'Affinity Profiling and Discrimination by Association in Online Behavioural Advertising' (n 145).

¹⁴¹ Wachter, 'Affinity Profiling and Discrimination by Association in Online Behavioural Advertising' (n 145); Brent Mittelstadt, 'From Individual to Group Privacy in Big Data Analytics' (2017) 30 *Philosophy & Technology* 475. Wachter, 'Affinity Profiling and Discrimination by Association in Online Behavioural Advertising' (n 145).

¹⁴² Burrell (n 146); Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR' (2018) 3 *Harvard Journal of Law & Technology* 841; Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 *International Data Privacy Law* 76; Jeremy B Merrill Ariana Tobin, 'Facebook Is Letting Job Advertisers Target Only Men' (*ProPublica*, 18 September 2018) <<https://www.propublica.org/article/facebook-is-letting-job-advertisers-target-only-men>> accessed 24 March 2019; *ProPublica* Data Store, 'COMPAS Recidivism Risk Score Data and Analysis' (*ProPublica* Data Store, 2 May 2016) <<https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis>> accessed 7 May 2019. Burrell (n 146); Wachter, Mittelstadt and Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR'; Wachter, Mittelstadt and Floridi; Ariana Tobin; Store.

individuals and groups from approaching Equality Bodies with potential cases in the first place.¹⁴³ They likewise challenge traditional notions of discrimination as they may not follow familiar human patterns or biases.¹⁴⁴ AI systems process large datasets and find unexpected correlations to classify cases or make predictions which are not guaranteed to be interpretable or understandable by humans,¹⁴⁵ and yet may be proxies for protected characteristics.¹⁴⁶

The unique nature of algorithmic bias and discrimination indicate a need for Equality Bodies to be creative in their search for potential illegal discrimination. While Equality Bodies are limited by the sectoral nature of EU non-discrimination law to pursuing cases of discrimination based on legally protected attributes in specific protected sectors, they should nonetheless be expansive in their search for proxy attributes due to the unintuitive way in which AI can discriminate.

5.4. Discrimination against new, unprotected groups

While they may create hidden proxies for protected characteristics, posing risks to legally protected groups, AI also raises broader challenges for the foundations of non-discrimination law. Groups defined by characteristics which are both unintuitive or incomprehensible to humans, and likewise not covered by existing law, can nonetheless experience severe disparity from uses of AI. Imagine, for example, if a particular user segment defined by (1) the web browser they use and (2) historical search behaviour was to experience significant price discrimination that would be illegal if experienced by a protected demographic group. This gap points towards a need to assess whether current non-discrimination law and protected attributes are sufficiently broad in scope to effectively measure and mitigate AI-driven inequality.¹⁴⁷

¹⁴³ On how biased data leads to biased outcomes see Alexandra Chouldechova, 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments' (2017) 5 *Big Data* 153; see also Jerry Kang and others, 'Implicit Bias in the Courtroom' (2011) 59 *UCLA L. rev.* 1124; Marion Oswald and Alexander Babuta, 'Data Analytics and Algorithmic Bias in Policing'. On how biased data leads to biased outcomes see Chouldechova; see also Kang and others; Oswald and Babuta.

¹⁴⁴ Sandra Wachter and BD Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' (2019) 2019 *Columbia Business Law Review* 494; Timo Makkonen, 'Equal in Law, Unequal in Fact: Racial and Ethnic Discrimination and the Legal Response Thereto in Europe' (Univ 2010) 57, 64; Jennifer Cobbe and Jatinder Singh, 'Regulating Recommending: Motivations, Considerations, and Principles' (2019) forthcoming *European Journal of Law and Technology* <<https://papers.ssrn.com/abstract=3371830>> accessed 28 February 2020. Wachter and Mittelstadt; Makkonen 57, 64; Cobbe and Singh.

¹⁴⁵ Luciano Floridi, 'The Search for Small Patterns in Big Data' (2012) 2012 *The Philosophers' Magazine* 17. *ibid.*

¹⁴⁶ Datta and others (n 109); Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 *California Law Review*; Brent Mittelstadt and Luciano Floridi, 'The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts' (2016) 22 *Science and Engineering Ethics* 303. Datta and others (n 109); Barocas and Selbst; Mittelstadt and Floridi.

¹⁴⁷ Sandra Wachter, 'The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law' (2022) 97 *Tul. L. Rev.* 149. *ibid.*

5.5. Gaps in data needed to measure bias and fairness

Bias and fairness tests need data about affected parties to work. This can come in the form either of (1) access to “ground truth” data about the distribution of protected attributes in the affected population, or (2) equivalent inferred data about protected attributes. Access to such data is often limited for sound historical reasons or due to existing data gaps affecting the population.

At the same time, researchers, policy bodies, and even the European Court of Justice recognise the need for data about protected attributes to measure algorithmic discrimination.¹⁴⁸ Recent years have seen broad calls for wider collection of protected attributes data to help detect and mitigate algorithmic bias and discrimination. The European Union, European Committee of Social Rights, EU High Level Group on Non-discrimination, Equality, and Diversity, and the United Nations Special Representative on Extreme Poverty and Human Rights, among others, have explicitly called on organisations to collect equality data, which includes sensitive data, to support equality law cases.¹⁴⁹ Equality bodies seeking to measure algorithmic bias and discrimination will need to address these gaps by accessing such data collected (or inferred) by others to effectively use the tools provided by technical standards. Two Directives on Standards for Equality Bodies adopted in May 2024 likewise create obligations for EU Member States to facilitate the collection and use of equality data by Equality Bodies and could be leveraged to promote the availability of equality data.¹⁵⁰

Equality bodies may face an uphill battle in arguing for the existence of illegal discrimination. Obtaining information about the predictions received by a critical mass of people affected by a particular AI system can be very difficult, especially in cases where this information is shared by an AI provider who has not collected the demographic data necessary to compare performance between groups and identify possible discrimination.

¹⁴⁸ *Nadežda Riežniece v Zemkopības ministrija and Lauku atbalsta dienests (n 135); Asociația Accept v Consiliul Național pentru Combaterea Discriminării*, 2013 E.C.R. I–275.

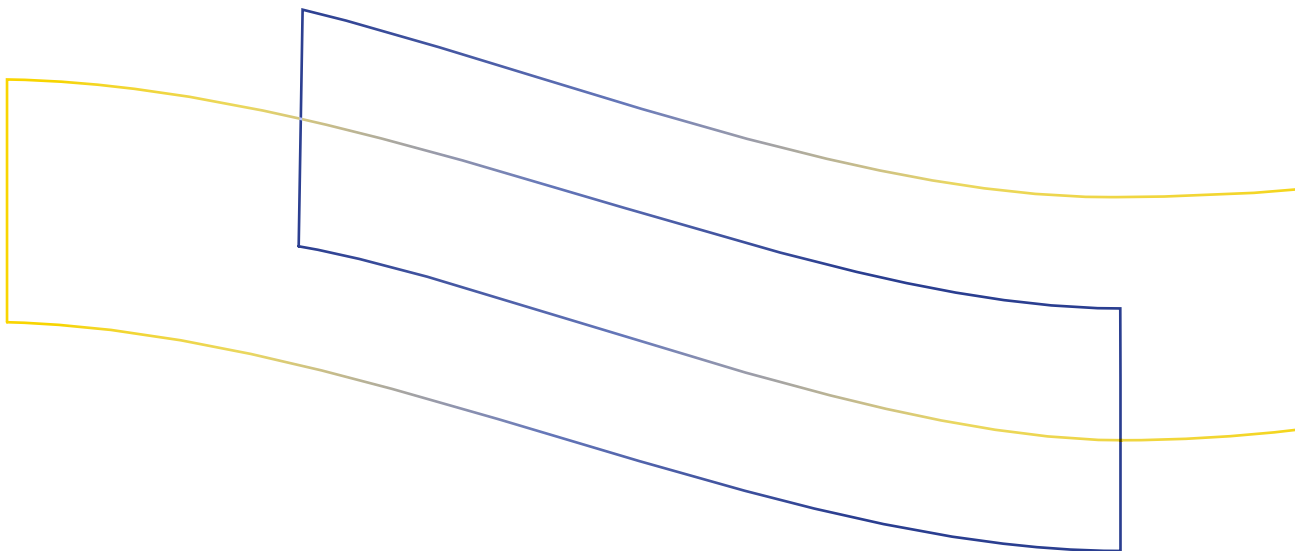
¹⁴⁹ Lilla Farkas and others, ‘The Meaning of Racial or Ethnic Origin in EU Law: Between Stereotypes and Identities.’ (2017) 118 <<http://bookshop.europa.eu/uri?target=EUB:NOTICE:DS0116914:EN:HTML>> accessed 9 February 2020. *ibid.*

¹⁵⁰ DIRECTIVE (EU) 2024/1500 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 May 2024 on standards for equality bodies in the field of equal treatment and equal opportunities between women and men in matters of employment and occupation, and amending Directives 2006/54/EC and 2010/41/EU 2024 (2024/1500); Council Directive (EU) 2024/1499 of 7 May 2024 on standards for equality bodies in the field of equal treatment between persons irrespective of their racial or ethnic origin, equal treatment in matters of employment and occupation between persons irrespective of their religion or belief, disability, age or sexual orientation, equal treatment between women and men in matters of social security and in the access to and supply of goods and services, and amending Directives 2000/43/EC and 2004/113/EC 2024 (2024/1499).

5.6. Aligning fairness measures with legal foundations

Equality bodies should be aware of how the definitions of fairness and bias used by AI providers align with core concepts of non-discrimination law. The majority of existing research on fairness and bias in AI is grounded in American anti-discrimination and equality law, and particularly influenced by their concepts of “disparate treatment” and “disparate impact.”¹⁵¹ It is important to note this gap, as these concepts do not translate cleanly or directly to European legal notions of non-discrimination. Statistical measures of bias and fairness do not capture the contextual nature of non-discrimination law in the EU.¹⁵²

As algorithmic discrimination is frequently indirect, deployers in sectors and regions that have substantive equality duties may therefore want to avoid enforcing fairness on their systems using “bias preserving” metrics that do not support substantive equality (see: Section 4.1.2.2). Equality bodies should be aware of difference between bias preserving metrics that are useful for measuring direct discrimination, and bias transforming metrics that are useful for measuring indirect discrimination. They should likewise question any usage of bias preserving fairness metrics by providers or deployers of high-risk AI systems to enforce fairness in AI products and services deployed on the EU market.



¹⁵¹ Barocas and Selbst (n 152); Pauline T Kim, ‘Data-Driven Discrimination at Work’ (2016) 58 Wm. & Mary L. Rev. 857; Crystal Yang and Will Dobbie, ‘Equal Protection Under Algorithms: A New Statistical and Legal Framework’ [2019] Available at SSRN 3462379; Zach Harned and Hanna Wallach, ‘Stretching Human Laws to Apply to Machines: The Dangers of a ‘Colorblind’ Computer’ [2019] Florida State University Law Review, Forthcoming; Thomas Nachbar, ‘Algorithmic Fairness, Algorithmic Discrimination’ [2020] Virginia Public Law and Legal Theory Research Paper.

¹⁵² Wachter, Mittelstadt and Russell, ‘Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law’ (n 86).



6. Conclusion and key recommendations

How to use the Artificial Intelligence Act to investigate AI bias and discrimination: A guide for Equality Bodies



Equality Bodies are granted new, important powers by the AIA to aid in their investigation of bias and discrimination in AI systems. This report has explained the basis and scope of these new powers in the AIA, how they connect with bias-related requirements in the AIA and harmonised technical standards, how to use the powers effectively, and highlighted open challenges. To conclude, this section summarises best practices and key lessons for Equality Bodies to use their AIA powers and harmonised standards effectively in practice.

1. Use documentation access and testing rights to investigate AI bias and discrimination

The AIA grants Equality Bodies and other NFRAs significant powers to obtain documentation about high-risk AI systems and use this documentation to investigate biases and potential discrimination. When Equality Bodies nominated under Article 77 are notified by a MSA that a risk of discrimination exists, they can request further testing and collaborate with MSAs to evaluate problematic systems. Equality Bodies should make full use of these new and powerful rights to protect against AI discrimination. In particular, they should request results of bias tests and fairness mitigations carried out under a provider's risk management system (see: Section 4.5). This may prove to be invaluable information to establish *prima facie* discrimination and reverse the burden of proof in particular uses of high-risk AI systems.

2. Collaboratively audit how fairness is measured and enforced by AI providers

Using fairness metrics and enforcement methods on AI systems requires technical expertise and system access that Equality Bodies may not have in practice (see: Section 4.1). Nonetheless, Equality Bodies will receive statistical results and information about how these techniques have been used to measure and reduce bias. Equality Bodies could ideally develop the necessary in-house expertise to understand and use this information in their investigations, or otherwise partner with MSAs, external researchers and other third-parties to interpret statistical evidence obtained through their right to access documentation. In particular, Equality Bodies should examine the costs and potential harms of using debiasing and fairness enforcement methods such as loss of accuracy (see: Section 4.1.2.3) and ensure any harms resulting from fairness enforcement, such as levelling down, do not disproportionately harm marginalised groups.

3. Collaborate closely with market surveillance authorities

The rights to documentation and testing can be very robust enforcement mechanisms to protect fundamental rights against AI harms when linked to Article 79 (see: Section 4.1). This article calls on MSAs to conduct post-market monitoring of high-risk AI systems, recommend specific actions to bring non-compliant systems into compliance with the AIA, and call for the withdrawal or recall of the system from the EU market in cases where the deployers fail to correct the system. This is the strongest and clearest path available to Equality Bodies to govern equality in AI systems under the AIA as this Article also creates an obligation for MSAs to fully cooperate with Equality Bodies listed as Article 77 authorities. By collaborating with MSAs, Equality Bodies can identify harmful AI systems and call for compliance or withdrawal from the EU market.¹⁵³ Close collaboration with MSAs should thus be a key priority for Equality Bodies under the AIA.

4. Use standards but make your own assessment about whether AI is discriminatory

Harmonised standards are unlikely to define specific thresholds or performance gaps between groups of people that would constitute illegal discrimination. Rather, decisions about which biases are intentional or unwanted, and the point at which AI becomes discriminatory, will be made on a system-by-system or case-by-case basis. Equality Bodies should be aware that standards will set expectations about how bias can be measured and mitigated in AI systems but will not determine whether these biases are illegal or unethical.

¹⁵³ AIA Article 79(5).

5. Check for ethics washing by AI providers

The gap in deciding whether an AI system is problematically biased or discriminatory will largely be filled by AI providers by default. Providers will have significant power to decide which biases or performance gaps are acceptable or intentional through the technical documentation, voluntary FRIAs, and risk management systems they implement. Specifically, how they define the “intended purpose” and “expected level of accuracy” for specific groups of people will inform the types of risks monitored by deployers and MSAs. Equality Bodies should critically assess how providers define the system’s intended use and expected levels of accuracy for different groups to close this significant enforcement loophole and ensure they are not “ethics washing” their systems in practice (see: Section 4.1.1).

6. Question how groups are defined because it can hide bias and discrimination

AI providers will decide how to measure performance gaps in their systems, including which groups of people they are comparing. This is an important decision in the context of non-discrimination law where defining appropriate “disadvantaged” and “comparator” groups is a key battleground in court cases. When reviewing the results of performance testing obtained through their right to access documentation, Equality Bodies should pay close attention to how groups are defined in statistical performance measures.

Defining groups in specific ways, especially intersectional groups where a “divide and conquer” approach can hide *prima facie* discrimination,¹⁵⁴ is a tempting way to make high-risk AI systems appear fairer or less biased than their real-world impact would suggest. Discrimination affecting broad, narrow, or intersectional groups can disappear if the size or scope of the groups being compared changes; imagine, for example, discrimination experienced by Black female applicants which disappears when focusing on Black or female applicants only.¹⁵⁵ Equality Bodies should ensure providers adequately address intersectional groups in their testing, as focusing solely on single-characteristic groups may conceal deeper disparities. In cases where providers (1) fail to report the results of multiple fairness metrics reporting on different performance gaps (e.g., precision, recall, error rates) or (2) narrowly define impacted groups, Equality Bodies should consider making a request to the relevant national MSA to carry out further specific tests with additional fairness metrics or groups.

¹⁵⁴ Wachter, Mittelstadt and Russell, ‘Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI’ (n 32).

¹⁵⁵ *ibid.* *ibid.*

7. Ensure fairness enforcement methods are aligned with non-discrimination law

Not all fairness enforcement methods are created equally; many have been developed with US anti-discrimination law in mind and are inappropriate or illegal to use in the EU.¹⁵⁶ Equality Bodies should investigate whether the fairness enforcement methods used by AI providers and reported in their risk management or technical documentation align with substantive equality duties in EU non-discrimination law (see: Section 5.6). In particular, Equality Bodies should be vigilant for systems that achieve fairness through “levelling down” which can indicate the system is creating avoidable harms to fundamental rights, health and safety in the name of fairness, for example by increasing the rate of misdiagnosis in medical applications.

This report has set out a roadmap for Equality Bodies to protect equality against harm from AI. Enforcement of the AIA will be a multi-faceted and difficult endeavour. Best practices will inevitably change as enforcement of the AIA matures and harmonised technical standards are published. For now, by following these recommendations, Equality Bodies can effectively leverage their new AIA powers and harmonised technical standards to investigate and address discrimination in AI systems. Equality Bodies are best placed to ensure AI systems deployed in the EU comply fully with equality law and protect marginalised groups against harm to their fundamental rights.



¹⁵⁶ Wachter, Mittelstadt and Russell, 'Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law' (n 86).

Equinet Member Equality Bodies

ALBANIA

Commissioner for the Protection from Discrimination
www.kmd.al

AUSTRIA

Austrian Disability Ombudsperson
www.behindertenanwalt.gv.at

AUSTRIA

Ombud for Equal Treatment
www.gleichbehandlungsanwaltschaft.gv.at

BELGIUM

Institute for the Equality of Women and Men
www.igvm-iefh.belgium.be

BELGIUM

Unia (Interfederal Centre for Equal Opportunities)
www.unia.be

BOSNIA AND HERZEGOVINA

Institution of Human Rights Ombudsman of Bosnia and Herzegovina
www.ombudsmen.gov.ba

BULGARIA

Commission for Protection against Discrimination
www.kzd-nondiscrimination.com

CROATIA

Office of the Ombudsman
www.ombudsman.hr

CROATIA

Ombudsperson for Gender Equality
www.prs.hr

CROATIA

Ombudswoman for Persons with Disabilities
www.posi.hr

CYPRUS

Commissioner for Administration and Human Rights (Ombudsman)
www.ombudsman.gov.cy

CZECH REPUBLIC

Public Defender of Rights
www.ochrance.cz

DENMARK

Danish Institute for Human Rights
www.humanrights.dk

ESTONIA

Gender Equality and Equal Treatment Commissioner
www.volinik.ee

FINLAND

Non-Discrimination Ombudsman
www.syrjinta.fi

FINLAND

Ombudsman for Equality
www.tasa-arvo.fi

FRANCE

Defender of Rights
www.defenseurdesdroits.fr

GEORGIA

Public Defender of Georgia (Ombudsman)
www.ombudsman.ge

GERMANY

Federal Anti-Discrimination Agency
www.antidiskriminierungsstelle.de

GREECE

Greek Ombudsman
www.synigoros.gr

HUNGARY

Office of the Commissioner for Fundamental Rights
www.ajbh.hu

IRELAND

Irish Human Rights and Equality Commission
www.ihrec.ie

ITALY

National Office against Racial Discrimination
www.unar.it

KOSOVO*

Ombudsperson Institution
<https://oik-rks.org>

LATVIA

Office of the Ombudsman
www.tiesibsargs.lv

LITHUANIA

Office of the Equal Opportunities Ombudsperson
www.lygybe.lt

LUXEMBURG

Centre for Equal Treatment
www.cet.lu

MALTA

Commission for the Rights of Persons with Disability
www.crpdp.org.mt

MALTA

National Commission for the Promotion of Equality
ncpe.gov.mt

MOLDOVA

Equality Council
www.egalitate.md

MONTENEGRO

Protector of Human Rights and Freedoms (Ombudsman)
www.ombudsman.co.me

NETHERLANDS

Netherlands Institute for Human Rights
www.mensenrechten.nl

NORTH MACEDONIA

Commission for Prevention and Protection against Discrimination
www.kszd.mk

NORWAY

Equality and Anti-Discrimination Ombud
www.ldo.no

POLAND

Commissioner for Human Rights of the republic of Poland
bip.brpo.gov.pl

PORTUGAL

Commission for Citizenship and Gender Equality
www.cig.gov.pt

PORTUGAL

Commission for Equality in Labour and Employment
cite.gov.pt/web/pt

ROMANIA

National Council for Combating Discrimination
www.cncd.ro

SERBIA

Commissioner for Protection of Equality
www.ravnopravnost.gov.rs

SLOVAKIA

Slovak National Centre for Human Rights
www.snsnlp.sk

SLOVENIA

Advocate of the Principle of Equality
www.zagovornik.si

SPAIN

Council for the Elimination of Ethnic or Racial Discrimination
igualdadynodiscriminacion.igualdad.gob.es

SPAIN

Institute of Women
www.inmujeres.gob.es

SWEDEN

Equality Ombudsman
www.do.se

UKRAINE

Ukrainian Parliament Commissioner for Human Rights
www.ombudsman.gov.ua

UNITED KINGDOM - GREAT BRITAIN

Equality and Human Rights Commission
www.equalityhumanrights.com

UNITED KINGDOM - NORTHERN IRELAND

Equality Commission for Northern Ireland
www.equalityni.org

** This designation is without prejudice to positions on status, and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo declaration of independence.*



ISBN 978-92-95112-87-2 | © Equinet 2024



Co-funded by
the European Union

www.equineteurope.org