Stanford – Vienna
Transatlantic Technology Law Forum

A joint initiative of
Stanford Law School and the University of Vienna School of Law

European Union Law Working Papers

No. 140

Algorithmic Bias and Discrimination in High-Risk AI Systems under the EU Artificial Intelligence Act

Lea Deman

2026

European Union Law Working Papers

Editors: Siegfried Fina and Roland Vogl

About the European Union Law Working Papers

The European Union Law Working Paper Series presents research on the law and policy of the European Union. The objective of the European Union Law Working Paper Series is to share "works in progress". The authors of the papers are solely responsible for the content of their contributions and may use the citation standards of their home country. The working papers can be found at http://ttlf.stanford.edu.

The European Union Law Working Paper Series is a joint initiative of Stanford Law School and the University of Vienna School of Law's LLM Program in European and International Business Law.

If you should have any questions regarding the European Union Law Working Paper Series, please contact Professor Dr. Siegfried Fina, Jean Monnet Professor of European Union Law, or Dr. Roland Vogl, Executive Director of the Stanford Program in Law, Science and Technology, at:

Stanford-Vienna Transatlantic Technology Law Forum
http://ttlf.stanford.edu

Stanford Law School
Crown Quadrangle
559 Nathan Abbott Way
Stanford, CA 94305-8610

University of Vienna School of Law
Department of Business Law
Schottenbastei 10-16
1010 Vienna, Austria

About the Author

Lea Deman obtained her Bachelor of Laws (LL.B.) in International Legal Studies from the University of Vienna, Austria, in the summer of 2026 and will begin a Master of Laws (LL.M.) in the same field in the fall of 2026. In spring 2024, she spent an exchange semester in Madrid, Spain, where she completed international law courses in Spanish and English. Since May 2025, she has been working at a law firm in Vienna specializing in data protection law. She is a co-author of two chapters in a forthcoming practitioner's guide on the legal framework for AI projects, one of which focuses on the AI Act.

General Note about the Content

Suggested Citation

This European Union Law Working Paper should be cited as:
Lea Deman, Algorithmic Bias and Discrimination in High-Risk AI Systems under the EU Artificial Intelligence Act, Stanford-Vienna European Union Law Working Paper No. 140, http://ttlf.stanford.edu.

Copyright

Abstract

The growing use of artificial intelligence (AI) in sensitive areas such as law enforcement, education, justice administration, and migration has raised concerns about algorithmic bias and discrimination. The EU adopted the Artificial Intelligence Act (AI Act), the first legal instrument to comprehensively regulate AI. The AI Act subjects high-risk AI systems to a distinct range of obligations under a risk-based framework.

This thesis analyzes the extent to which the AI Act effectively addresses algorithmic bias and discrimination in high-risk AI systems. First, it discusses the concepts of AI, algorithms, algorithmic bias, and algorithmic discrimination. Then, it provides a descriptive and critical analysis of the provisions most central to bias, namely Art. 9, 10, 14, 15, and 86. Finally, it assesses the Act's overall approach to discrimination.

This thesis argues that the Act is best understood as a product-safety regulation designed to address bias, a concept rooted in computer science, through the technical obligations set out in the Act. However, the Act is ill-suited to regulate and prevent algorithmic discrimination as a legal, normative concept. Recurring ambiguities in the wording as well as a regulatory gap concerning emergent bias further leave its effectiveness undetermined. Therefore, the Act plays a supporting role in seeking to minimize the risk of algorithmic discrimination by regulating algorithmic bias, whereas the legal assessment of discrimination remains with the existing EU non-discrimination framework.

Contents

1. Introduction

> ”You’re a single mother with three children aged eight to 11. You hit rock-bottom financially and think what now? My children and I sometimes had to go to bed without food.“[1]

This account is not an isolated tragedy but reflects the reality of at least 35.000 Dutch parents in what is now known as the Dutch childcare benefits scandal, one of the most prominent examples of algorithmic discrimination.[2] In 2013, the Dutch tax authorities introduced a self-learning decision-making algorithm which was meant to detect errors, irregularities and fraud in child benefits applications. However, the system relied on applicants’ nationalities as risk factors, resulting in higher risk scores for non-Dutch citizens.[3] Tens of thousands of families were wrongfully accused of committing fraud and forced to pay back large sums of money; this led to poverty, mental health issues, divorces and even loss of custody of children, while the system’s opacity left them unable to understand the decisions and exercise their legal rights.[4]

The Dutch childcare benefits scandal is one of many instances of algorithmic discrimination, i.e. when machine-learning algorithms make decisions that discriminate against people. Such discrimination often originates in algorithmic bias, characterized as technical errors in computer systems. With the increasing use of Artificial Intelligence (AI) across many domains, concerns in regard to algorithmic bias and discrimination become more and more relevant.[5] This is especially true for sensitive,

[1] 'Dutch Rutte Government Resigns over Child Welfare Fraud Scandal' *BBC News* (15 January 2021) https://www.bbc.co.uk/news/world-europe-55674146 accessed 31 July 2026.
[2] Josien Arts and Marguerite van den Berg, 'What the Dutch Benefits Scandal and Policy's Focus on "Fraud" Can Teach Us About the Endurance of Empire' (2025) 45(1) Critical Social Policy 177.
[3] Amnesty International, 'Xenophobic Machines: Discrimination Through Unregulated Use of Algorithms in the Dutch Childcare Benefits Scandal' (Index No EUR 35/4686/2021, 25 October 2021) 16.
[4] Arts and van de Berg (n 1) 177.; Amnesty International, Xenophobic Machines (n 2) 6, 13.
[5] Keketso Kgomosotho, 'Analysing the EU Ai Act’s Treatment of Algorithmic Discrimination' (2025) 9 Univ Vienna Law Rev. 76, 78.

high-risk areas such as law enforcement, education, administration of justice or migration. As a response to this threat, the EU adopted the Artificial Intelligence Act (AI Act), thereby creating the first legal instrument to comprehensively regulate AI.[6] It introduced a risk-based approach, sorting AI into four distinct categories depending on the risk they pose: Unacceptable risk, high risk, limited risk and minimal risk.[7]

This thesis analyses the extent to which the AI Act effectively addresses algorithmic bias and discrimination in high-risk AI systems. Chapter 2 sets out the methodological approach underlying this thesis. Chapter 3 discusses the notions of AI, algorithms and, most importantly, algorithmic bias and discrimination. Chapter 4 analyses the origins of algorithmic bias and Chapter 5 provides a brief introduction to the AI Act framework. The focus of this thesis lies in Chapter 6 and 7: Chapter 6 provides a descriptive and critical analysis of the main provisions regulating algorithmic bias and Chapter 7 contains a critical assessment of the AI Act's regulatory approach to algorithmic discrimination. Finally, Chapter 8 concludes this thesis with an answer to the research question as well as a brief outlook on future legal developments.

2. Methodology

This thesis employs a doctrinal legal-analytical analysis of the extent to which the EU AI Act effectively addresses both algorithmic bias and discrimination in high-risk AI systems. To answer the central research question, three sub-questions will be answered first: what algorithmic bias and discrimination are as well as their relationship to each other in Chapter 3, how and where algorithmic bias emerges throughout the AI lifecycle in Chapter 4, how the relevant provisions regulate algorithmic bias in

[6] Kgomosotho (n 5) 83.
[7] Martin Ebers, 'Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU's AI Act' (2025) 16 European Journal of Risk Regulation 684.

Chapter 6 and to what extent this regulation can effectively reduce algorithmic discrimination in Chapter 7.

The doctrinal legal-analytical analysis in this thesis involves the systematic identification, interpretation and critical evaluation of relevant legal sources, both primary and secondary. The interpretation of the most relevant primary source, the AI Act, uses statutory interpretation methods, primarily the literal meaning of a text, its systematic context within the AI Act as well as the EU legal framework, the intention of the legislator and its legislative history as well as a teleological interpretation of the Act's objectives. These methods are especially relevant for the analysis of Art. 9 and 10.

To assess the provision's effectiveness, it is necessary to clarify the benchmark against which effectiveness will be measured. First, the objectives the Act stated in its Art. 1, in particular the protection of fundamental rights; and second, the ability and capacity of the individual provisions to capture the sources of bias identified in Chapter 4.

This thesis takes on an interdisciplinary approach, since algorithmic bias is a concept rooted in computer science. A meaningful legal analysis of the Act's provisions therefore requires an understanding of the technical mechanisms surrounding algorithmic bias. While non-legal sources provide important context for a better assessment of the individual provisions and the entire AI Act, the focus of this thesis remains primarily legal. The materials used for the analysis encompass EU primary and secondary law (including relevant recitals), guidelines, proposals and other soft law instruments, CJEU case law, academic commentary as well as a wide array of legal and non-legal journal articles.

Finally, the scope of this thesis is deliberately limited to (1) only high-risk AI systems and to (2) the provisions with a direct relevance to algorithmic bias and discrimination,

namely Art. 9, 10, 14, 15 and 86. Other provisions, such as Art. 11 or 13, will be analysed in conjunction with the aforementioned articles due to their connected relationships.

3. Algorithmic Bias and Discrimination

To understand what algorithmic bias and discrimination are, it is important to define artificial intelligence (AI) first – which is not an easy task, since there is no universally accepted definition. For the purposes of this paper, the definition provided by the European Commission's High-Level Expert Group will be used, according to which AI systems refers to systems that "display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals"[8].

This definition, however, describes AI only at a functional level; a meaningful analysis of algorithmic bias and discrimination also requires a clear understanding of the underlying computational mechanism at hand: the algorithm. While the terms algorithmic bias and algorithmic discrimination are often used interchangeably in general discussion, there are significant distinctions primarily rooted in legal implications as well as disciplinary origins. Dourish defines an algorithm as an "abstract, formalized description of a computational procedure"[9]. Algorithms can be classified by level of automation, i.e. fully automated algorithms functioning without any human intervention (e.g. e-mail spam filters) and partly automated algorithms serving

[8] High-Level Expert Group on Artificial Intelligence, 'A Definition of AI: Main Capabilities and Disciplines' (European Commission, 18 December 2018) https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf accessed 10 July 2026.

[9] Paul Dourish, 'Algorithms and their others: Algorithmic culture in context' (2016) 3 (2) Big Data & Society <https://journals.sagepub.com/doi/epub/10.1177/2053951716665128> accessed 3 July 2026.

an assisting function in human decision making (e.g. credit assessments)[10]. The main distinction is that between rule-based algorithms and machine-learning algorithms. The former functions on a simple "if this, then that" logic, guaranteeing highly predictable outcomes at maximum efficiency.[11] For instance, in tax law: *if* taxable income falls within a certain bracket, *then* the corresponding statutory tax rate is applied. By contrast, machine-learning algorithms are capable of autonomously detecting and analysing data while improving their performance over time without relying on predefined human instructions.[12] Lehr and Ohm define machine-learning as an "automated process of discovering correlations (sometimes alternatively referred to as relationships or patterns) between variables in a dataset, often to make predictions or estimates of some outcome"[13]. Of the two, machine learning is regarded as a subset or a scientific discipline in the field of AI.[14]

Against this background, the analysis can now turn to algorithmic bias and, subsequently, algorithmic discrimination. The European Commission defines algorithmic bias as "systematic and repeatable errors in a computer system that create unfair outcomes, such as favouring one arbitrary group of users over others"[15]. A notable distinction lies in their differing disciplinary origins, since the notion of algorithmic bias is rooted in computer science and refers to statistical imbalances at

[10] Frederik Zuiderveen Borgesius, *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making* (Council of Europe, Directorate General of Democracy 2018) 8 https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 accessed 3 July 2026.
[11] Janneke Gerards and Raphaële Xenidis, *Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Non-Discrimination Law* (European Commission, Directorate-General for Justice and Consumers, Publications Office 2021) 35 https://data.europa.eu/doi/10.2838/544956 accessed 3 July 2026.
[12] Gerards and Xenidis (n 11) 33.
[13] David Lehr and Paul Ohm, 'Playing with the Data: What Legal Scholars Should Learn About Machine Learning' (2017) 51 UC Davis L Rev 653, 671.
[14] Zuiderveen Borgesius (n 10) 9.; Kgomosotho (n 5) 76.
[15] Commission, 'Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts' SWD (2021) 84 final https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021SC0084 accessed 5 July 2026.

the input level. Measures aimed at identifying and reducing these disparities are primarily technical and quantitative in nature.[16] This includes mathematical fairness metrics such as statistical parity, which requires members of different groups to receive positive predictions at the same rate, as well as equalized odds, which ensures that predictions, specifically the True Positive Rate and the False Positive Rate, are the same across all demographics. Other relevant fairness metrics include predictive parity and calibration, which assess whether a system's accuracy and risk scores remain consistent across different groups.[17]

Conversely, algorithmic discrimination is a normative concept and describes the legal outcome produced by algorithmic bias. Discrimination can be categorized in many ways (e.g. structural, organizational and interpersonal)[18] and it focuses on preventing unequal treatment where such treatment lacks an objective and reasonable justification and is based on protected grounds such as gender, race or sexual orientation.[19] Under EU law a key distinction is that between direct and indirect discrimination. Direct discrimination occurs when "one person is treated less favourably than another is, has been or would be treated in a comparable situation" because of protected characteristics.[20] In the context of algorithmic discrimination, however, this concept quickly reaches its limits because of a phenomenon known as algorithmic proxy discrimination: Most AI systems do not use protected characteristics as input variables but rather rely on neutral data points that strongly correlate with

[16] Kgomosotho (n 5) 85.
[17] Pratyush Garg, John Villasenor and Virginia Foggo, 'Fairness Metrics: A Comparative Analysis' in *2020 IEEE International Conference on Big Data (Big Data)* (IEEE 2020) 3662 https://doi.org/10.1109/BigData50022.2020.9378025 accessed 3 July 2026.
[18] Solon Barocas, Moritz Hardt and Arvind Narayanan, *Fairness and Machine Learning: Limitations and Opportunities* (MIT Press 2023) 209.
[19] Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights) art. 14.; Burden v United Kingdom (2008) 47 EHRR 38 [60].
[20] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin [2000] OJ L180/22, art 2(2)(a).

those protected attributes (e.g. postal codes as a proxy for ethnic background or past career gaps as a proxy for gender, as demonstrated by the Austrian AMS algorithm[21]).[22] As a result, it becomes difficult to establish that a protected attribute was a decisive factor in the algorithmic outcome. The challenge of proxy discrimination can be addressed more adequately through the notion of indirect discrimination:[23] Indirect discrimination occurs when "an apparently neutral provision, criterion or practice would put [individuals belonging to a protected group] at a particular disadvantage compared with other persons, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary".[24] While this framework works to cover proxy discrimination more adequately, its formulation is highly open-ended and its "objective justification" can override principles of non-discrimination, thereby creating legal uncertainty.[25]

The AI Act's focus lies in the identification and mitigation of algorithmic bias at a technical level, primarily through the requirements listed in Art. 10 as well as other applicable provisions. These provisions are the focus of this paper and will be analysed in Chapter 6.

4. Sources of Algorithmic Bias

Algorithmic bias is not the result of a single error, but rather an interplay of many, oftentimes overlapping factors throughout the lifecycle of an AI system.[26] While the

[21] Doris Allhutter and others, *Der AMS-Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS)* (ITA-Projektbericht Nr 2020-02, Institut für Technikfolgen-Abschätzung, Österreichische Akademie der Wissenschaften 2020) https://epub.oeaw.ac.at/ita/ita-projektberichte/2020-02.pdf accessed 3 July 2026.
[22] Gerards and Xenidis (n 11) 44.
[23] Gerards and Xenidis (n 11) 71.
[24] Council Directive 2000/43/EC (n 1), art 2(2)(b).
[25] Zuiderveen Borgesius (n 10) 20.
[26] Kristof Meding, 'It's Complicated: The Relationship of Algorithmic Fairness and Non-Discrimination Regulations in the EU AI Act' (2025) https://arxiv.org/abs/2501.12962 accessed 6 July 2026.

basis of all sources of bias is ultimately human judgement, the following analysis distinguishes them by stage of algorithmic lifecycle at which they are introduced; as the human factor is present throughout all stages, this aspect will be addressed and analysed within each respective category.

The OECD categorizes the AI lifecycle through seven cyclical stages:[27]

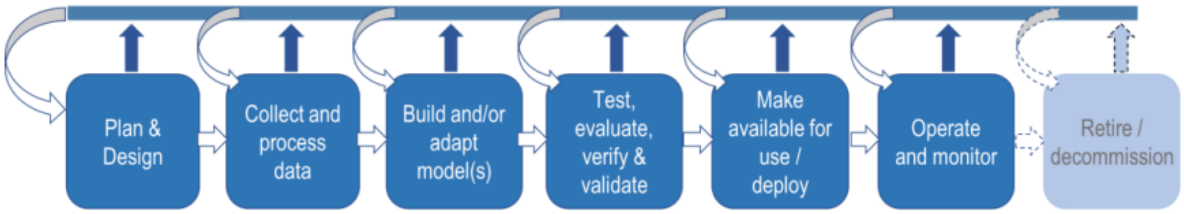

Figure 1: The OECD Seven-Stage Model

As seen in the figure above, a series of seven stages can be identified: (1) planning and designing, (2) collection and processing of data, (3) building and/or adapting models, (4) testing, evaluation, verification and validation of the model, (5) deployment, (6) operation and monitoring and (7) retiring and decommissioning.

While the OECD model provides a comprehensive account of the AI lifecycle, the seven stages are often simplified into a three-stage model comprising the planning, development and use stages. This framework creates a useful structure for analysing the points at with algorithmic bias may arise.

Although the following discussion is organized according to the lifecycle phases listed above, the sources of bias can be grouped into three categories: Design-related sources, data-related sources and contextual sources. Although distinct, these categories are not mutually exclusive and may emerge across different stages of the AI lifecycle.

[27] OECD, *Report on the Implementation of the OECD Recommendation on Artificial Intelligence* (Meeting of the Council at Ministerial Level, 2–3 May 2024, C/MIN(2024)17, 2024) https://one.oecd.org/document/C/MIN(2024)17/en/pdf accessed 8 July 2026.

4.1. Planning Stage

4.1.1. Target Variables

The design and configuration of an algorithm constitute potential sources of bias since an algorithm is shaped by a range of technical and methodological choices made during its development. One of the earliest design choices concerns the definition of the algorithm's objective or outcome that a model is designed to predict: the target variable. To create a target variable an algorithm can understand, a developer must first translate an abstract real-world concept into a measurable computational objective.[28] Barocas and Selbst illustrate this point with the example of a "good employee": There is no universally accepted definition of a "good employee", it is therefore necessary to "translate" this concept through clear indicators such as sales performance or employee tenure.[29] Once the target variable has been determined, developers assign class labels, which are categories the outcome is divided into, such as good vs. bad employee. A crucial point at which bias may be introduced is the initial selection and definition of the target variables. It involves subjective judgements about which characteristics should be considered relevant (e.g. what constitutes a good employee), which in turn determines what the algorithm learns to optimise, thereby introducing bias. For instance, as Barocas and Selbst point out, defining a "good employee" primarily in terms of predicted tenure instead of productivity may systematically disadvantage groups with higher historical turnover rates.[30] Consequently, bias can emerge not because of an inherent error in its design, but rather because of the normative, subjective choices embedded in the objective the algorithm is designed to achieve.

[28] Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 Calif L Rev 671, 678.
[29] Barocas and Selbst (n 28) 679.
[30] Barocas and Selbst (n 28) 680.

4.1.2. Feature Selection

A further potential source of bias within the design stage lies in a process known as "feature selection", whereby developers decide which specific attributes to include in the algorithm's analysis.[31] The objective is to "find the relevant features (individually or in cooperation" and discard redundant and irrelevant features in order to preserve the information contained in the whole set of input variables with respect to the target class"[32]. While the definition of the target variable established what an algorithm is designed to predict, feature selection determines the information and characteristics on which these predictions are based on. One point at which bias may be introduced is when developers attempt to reduce complex human realities into a dataset, which will naturally be an inaccurate oversimplification and generalisation. If important information about a protected group is missing from the selected features, the algorithm may make systematically less accurate predictions about that group – resulting in bias.[33] For instance, a company might rely on university prestige as an indicator to predict employee quality; this, however, does not account for the fact that not every graduate from a prestigious university is highly capable and not every graduate from a less prestigious university is less capable.[34]

4.1.3. The Proxy Problem

Even if an algorithm is designed with unbiased objectives in mind and relevant features, it may still produce biased outcomes if certain variables function as proxies for protected attributes. Barocas and Selbst identify the core problem to be "redundant encodings", whereby "data or certain values for that piece of data are highly correlated

[31] Barocas and Selbst (n 28) 688.
[32] Gustavo Sosa-Cabrera and others, 'Feature Selection: A Perspective on Inter-attribute Cooperation' (2023) 17 Int. J. Data Sci. Anal. 139, 140.
[33] Barocas and Selbst (n 28) 688.
[34] Barocas and Selbst (n 28) 689.

with membership in specific protected classes".[35] Even by removing explicit references to protected characteristics, non-discrimination cannot be guaranteed, as increasingly accurate algorithms are able to identify alternative variables that reveal the same underlying information.[36] While this can happen without the developer's intent to discriminate[37], it can also be intentional through a process known as "masking", whereby prejudicial views are deliberately embedded into datasets and hidden by neutral proxies.[38]

4.2. Development Stage

The quality and accuracy of an algorithmic outcome largely depend on the quality of the input data, as reflected by the computer-science idiom "garbage in, garbage out".[39] Training data forms the basis upon which algorithms "learn" to make predictions and analyse data.[40]

One important source of bias arises during the "construction" of the training data itself, particularly in the process of assigning labels to data points. Bias may emerge during the subjective, manual process of categorisation: Algorithms treat assigned labels as correct and internalize and reproduce them in the future. If these labels reflect subjective and/or biased human judgements and prejudice, machine-learning models will learn and systematize that discrimination and apply it to new cases under the assumption that it represents the objective truth.[41] This is illustrated by a system

[35] Barocas and Selbst (n 28) 691-692.
[36] Ibid.
[37] Gerards and Xenidis (n 11) 44.
[38] Barocas and Selbst (n 28) 692.
[39] Sandra G Mayson, 'Bias In, Bias Out' (2019) 128 Yale LJ 2218,2224.
[40] Barocas and Selbst (n 28) 680.
[41] Barocas and Selbst (n 28) 681.

trained on biased admissions data at St. George's Hospital, which learned to systematically disadvantage women and racial minorities.[42]

Beyond biased labelling, distortions may arise from the representativeness of the data itself. Data frequently reflects historical patterns of discrimination; even if an algorithm is properly designed and functions flawlessly, if it is fed with incorrect, biased or unrepresentative data, this will be reflected in the algorithmic outcome. This becomes especially relevant when data does not include or account for members of protected classes (e.g. facial recognition software that underperforms on Black women in comparison to White women or Black men[43]). This is further reinforced by gaps in data which occur when certain groups are systematically underrepresented in datasets due to poverty or unequal access to digital technologies, thereby creating "dark zones" in which certain communities are systemically overlooked.[44] For instance, a smartphone application called Street Bump, which was used to collect road condition data in Boston, resulted in the underreporting of potholes and other road problems in poorer communities, where smartphone ownership was lower.[45]

Conversely, bias may also arise if certain groups are overrepresented within a dataset, whereby members of protected groups are subject to higher levels of observation (e.g. at their place of work[46]), resulting in a dataset that records their behaviour and potential misconduct more extensively in comparison to members of other groups. Algorithms trained on such data may identify this overrepresentation as evidence of a genuine

[42] Stella Lowry and Gordon Macpherson, 'A Blot on the Profession' (1988) 296 Brit. Med. J. 657.
[43] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) 81 PMLR 1.
[44] Barocas and Selbst (n 28) 684.
[45] Barocas and Selbst (n 28) 685.
[46] Barocas and Selbst (n 28) 687.

correlation between group membership and misconduct and reproduce those patters in future decisions.[47]

Overall, these examples demonstrate that data-related choices during the development of an algorithm represent a critical source of algorithmic bias by reflecting and reproducing existing societal inequalities.

4.3. Use Stage

4.3.1. Automation bias

Lastly, bias may also arise from the algorithm's use in a real-world context and the way it evolves over time, particularly from the way humans interact with its outputs. One of the most prominent examples of this is automation bias, whereby individuals place excessive trust in algorithmic outputs and blindly follow an algorithm's suggestions because they are perceived as inherently objective and accurate.[48] Automation bias can occur in two distinct forms: Errors of omission and errors of commission[49]. The former arises when decision-makers fail to take the necessary measures due to the lack of warning from the algorithm, despite various other indicators suggesting a different course of action. Conversely, errors of commission occur when decision-makers follow the algorithm's recommendations without critically analysing it, even though other, more reliable factors indicate that the algorithmic output is incorrect.[50]

At its core, automation bias is a cognitive phenomenon: Humans will often over-rely on automated systems as a way to reduce cognitive strain instead of independently

[47] Barocas and Selbst (n 28) 687.
[48] Gerards and Xenidis (n 11) 42.
[49] Linda J Skitka, Kathleen L Mosier and Mark Burdick, 'Does Automation Bias Decision-Making?' (1999) 51 Intl J Human-Computer Studies 991, 993.
[50] Kate Goddard, Abdul Roudsari and Jeremy C Wyatt, 'Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators' (2012) 19 J Am Med Inform Assoc 121.

analysing the algorithm's outputs.[51] This is also exacerbated by the perception of AI as a "partner" or even authority figure, resulting in a diffusion of responsibility.[52] Other core drivers of automation bias include complex tasks, task experience, heavy workload as well as time pressure.[53] Particularly in professional settings, algorithmic suggestions may serve as an anchor, whereby an initial point of reference influences subsequent judgements.[54] This significantly impacts real-world scenarios such as criminal sentencing or medical diagnostics.[55] For instance, in a study involving 30 physicians the presentation of incorrect automated diagnoses significantly reduced diagnostic accuracy, from 84.86% to 41.66% in fellows and, even more drastically, from 86.38-27.43% in non-fellows.[56]

4.3.2. Emergent bias

Another source of algorithmic bias emerging during the use stage of an algorithm is known as emergent bias. Emergent bias develops through the interaction between multiple AI systems, which, as the models collaborate and learn from each other, may result in the unintentional reinforcement of certain patters, leading to biased algorithmic outcomes that cannot be explained solely by looking at one AI system in isolation.[57] This phenomenon demonstrates that algorithmic bias can arise from the dynamics of the machine learning model itself, which poses new challenges to regulatory frameworks attempting to mitigate bias, such as the EU AI Act.[58]

---

[51] Giuseppe Romeo and Daniela Conti, 'Exploring Automation Bias in Human–AI Collaboration: A Review and Implications for Explainable AI' (2026) 41 AI & Soc 259.
[52] Skitka, Mosier and Burdick (n 49), 992.
[53] Goddard, Roudsari and Wyatt (n 50), 124.
[54] Falk Lieder, Tom L Griffiths, Quentin JM Huys and Noah D Goodman, 'The Anchoring Bias Reflects Rational Use of Cognitive Resources' (2018) 25 Psychon Bull Rev 322.
[55] Gerards and Xenidis (n 11) 42.
[56] Romeo and Conti (n 51) 265.
[57] Maeve Madigan and others, 'Emergent Bias and Fairness in Multi-Agent Decision Systems' (2025) https://arxiv.org/abs/2512.16433 accessed 11 July 2026.
[58] Ibid

4.3.3. Feedback loops

A further manifestation of algorithmic bias are feedback loops, which occur when an algorithm's outputs influence the very data used to retrain or evaluate the model, thereby creating an "echo chamber" where biased predictions become self-reinforcing over time.[59]

Algorithms are designed to learn from user and environmental feedback, e.g. a search engine relies on click-through rates and likes/dislikes to refine future outputs. While this mechanism is intended to improve accuracy, they may conflate correlation with validation, meaning that a link that is clicked most often may simply be the one ranked first, rather than the one that is objectively most relevant. Because clicks are used to inform future rankings, this dynamic can worsen and increase over time, leading to a systematic favouritism of items which perform well early on.[60]

This becomes problematic once user feedback reflects and thereby reinforces bias and prejudice. This issue is highlighted by predictive policing systems, whereby an algorithm that identifies certain neighbourhoods as high-risk for crime, increased police deployment to the affected areas results in more recorded stops and arrests – not necessarily because criminal activity is higher in those areas compared to others, but because there is more enforcement to record and witness potential misconduct and crime.[61] The new arrest data is then fed back into the system as training data, appearing to validate the algorithm's original prediction, which leads to the deployment of even more police in the affected regions.[62] Lum and Isaac's analysis of the PredPol

[59] Frederik Zuiderveen Borgesius, Philipp Hacker, Nina Baranowska and Alessandro Fabris, 'Non-Discrimination Law in Europe: A Primer for Non-Lawyers' (7 April 2024) https://ssrn.com/abstract=4786956 accessed 11 July 2026.
[60] Barocas, Hardt and Narayanan (n 18) 13.
[61] Barocas, Hardt and Narayanan (n 18) 14.
[62] Ibid

algorithm in Oakland demonstrates this effect: They found that Black residents were targeted for drug-crime related policing at approximately twice the rate of White residents despite comparable rates of drug use.[63] These discrepancies intensified over time through repeated feedback loops, despite the absence of race as an explicit input variable, signifying once more the increasing relevance of proxy discrimination.[64]

4.4. Synthesis

This analysis demonstrates that algorithmic bias is not the result of one single error but rather arises cumulatively across the entire lifecycle of an algorithm. At the planning stage, bias is introduced through subjective choices in defining target variables and selecting the relevant features. At the development stage, bias becomes embedded in the data itself via biased labelling or unrepresentative data, which may even reproduce historical patterns of discrimination. Finally, at the use stage, bias may emerge through the continuous interaction of humans with the algorithmic outputs as well as through feedback loops amplifying earlier distortions.

It is important to recall, however, that algorithmic bias, as discussed before, is widely regarded as a computer science term, whereas algorithmic discrimination is a legal, normative concept. Whether algorithmic bias gives rise to algorithmic discrimination is a separate question that lies outside the scope of technical detection of bias alone. This distinction is central to the analysis that follows in Chapter 7.

[63] Kristian Lum and William Isaac, 'To Predict and Serve?' (2016) 13(5) Significance 14.
[64] Ibid

5. The EU Artificial Intelligence Act

5.1. Legislative Context and Objectives

The AI Act, formally Regulation (EU) 2024/1689, is the first comprehensive, horizontal legal framework governing artificial intelligence at the regional level.[65] Article 1 of the Act defines three principal objectives: (1) unifying the internal market, (2) promoting human-centric and trustworthy AI, including a high-level protection of fundamental rights, health, safety and the environment as well as (3) fostering innovation.[66]

The Act's primary legal basis lies in Art. 114 Treaty on the Functioning of the European Union (TFEU), which, for the purpose of internal market unification, confers competence on the EU to adopt necessary measures and provisions.[67] The other objectives, most notably the protection of fundamental rights, do not in itself confer legislative competence upon the EU.[68] Consequently, since one of the central aims of the Act remains the protection of fundamental rights, the measures adopted by the Act must remain tethered to the objective of internal market integration[69] (rather than human rights protection as such) and must be pursued through mechanisms used in market regulation: EU product safety law and the New Legislative Framework (NLF)[70].[71]

---

[65] Ebers (n 7) 684.

[66] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 1.

[67] Patrick van Eecke and Bartholomäus Regenhardt, 'Article 1' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), *The EU Artificial Intelligence (AI) Act: A Commentary* (Kluwer Law International 2024) 9.

[68] Charter of Fundamental Rights of the European Union [2012] OJ C326/391, art 51(2).

[69] van Eecke and Regenhardt (n 67) 9.

[70] The New Legislative Framework is an EU legislative framework introduced in 2008 to harmonise product safety law by establishing common principles for conformity assessment, market surveillance and accreditation. (European Commission, 'New Legislative Framework' (Internal Market, Industry, Entrepreneurship and SMEs) https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en accessed 11 July 2026.)

[71] Almada & Radu, 'The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy' (2024) 25 German Law Journal 646, 648.

A related question is the Act's definition of discrimination. It is not directly defined in the Act, but it rather relies on the existing EU non-discrimination framework.[72] Recital 45 clarifies that the Regulation does not affect the existing Union law prohibiting discrimination.[73]

5.2. Scope

The material scope of the Act encompasses AI systems, as defined in Art 3 (1), and General Purpose AI Models (GPAI models), as defined in Art 3 (63). The definition of AI systems contains five criteria: (1) machine-based, (2) varying levels of autonomy, (3) adaptiveness after deployment, (4) inference from input and (5) impact on environments.[74] GPAI models, on the other hand, display "significant generality" and are "capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and can be integrated into a variety of downstream systems or applications".[75] GPAI models are subject to their own, largely self-contained framework set out in Chapter V of the Act and are therefore not further analysed in this thesis.

The personal scope distinguishes between the provider, who develops an AI system (or a GPAI model) or instructs others to do so and places it on the market under their own name,[76] and the deployer, who uses an AI system (or GPAI model) under its own authority in a professional context.[77] This distinction is relevant since the bulk of ex ante compliance requirements, such as the data-governance requirements of Art. 10,

[72] Kgomosotho (n 5) 88.
[73] AI Act (n 66) recital 45; Kgomosotho (n 5) 88.
[74] AI Act (n 66) art 3 (1).
[75] AI Act (n 66) art 3 (63).
[76] AI Act (n 66) art 3 (3).
[77] AI Act (n 66) art 3 (4).

fall on providers, while deployers bear responsibilities during the use of the AI system (e.g. technical and operational measures, human oversight, monitoring data quality).[78]

Temporally, the AI Act entered into force on August 1st 2024 and applies, as a rule, from August 2nd 2026, with certain chapters taking effect earlier: Chapters I and II from February 2nd 2025[79]; Chapter III Section 4, Chapter V, Chapter VII, Chapter XII and Art. 78 from August 2nd 2025.[80] Following the European Parliament's adoption of the Digital Omnibus on AI on June 16th 2026 several of the deadlines shifted: Art. 50 to December 2nd 2026, the requirements for Annex III high-risk systems to December 2nd 2027 and those for Annex I high-risk systems to August 2nd 2028.[81]

5.3. The Risk-Based Approach

At the core of the AI Act lies a risk-based approach, whereby the intensity of regulatory obligations depends on the potential impact of an AI system on health, safety, fundamental rights and other protected interests.[82] The rationale behind the risk-based framework is to reconcile two objectives, namely fostering innovation and benefitting from the opportunities offered by AI systems as well as ensuring the protection of fundamental rights.[83] Four tiers can be distinguished: unacceptable risk, high-risk, limited risk and minimal risk.

At the top sits the category of prohibited AI practices, which are listed in Chapter II Art. 5 of the Act. The prohibition of these practices is based on the recognition that, despite the many beneficial applications of AI, certain uses may be manipulative and

[78] AI Act (n 66) art 26.; Katerina Yordanova, 'Article 10' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), *The EU Artificial Intelligence (AI) Act: A Commentary* (Kluwer Law International 2024) 263.
[79] AI Act (n 66) art 113 (a).
[80] AI Act (n 66) art 113 (b).
[81] Commission, 'Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI)' COM(2025) 836 final, recital 22.
[82] Ebers (n 7) 684.
[83] Ebers (n 7) 685.

exploitative and therefore incompatible with Union values such as human dignity, freedom and equality.[84] Art 5 sets up a non-exhaustive list[85] of applications that are banned outright: subliminal techniques (a), exploitative systems targeting vulnerable groups (b), social scoring (c), predictions of criminal conduct (d), systems that create or expand facial recognition databases (e), systems that infer emotions (f), biometric categorisation systems (g) and real-time biometric categorization systems (h).[86] To guarantee effective application of Art. 5, the European Commission published guidelines on prohibited AI practices.[87]

At the opposite end lie minimal risk AI systems, which face no specific obligations beyond the AI literacy duty in Art. 4,[88] though a separate category of systems with limited risk carries transparency obligations irrespective of a high-risk classification.[89]

High-risk AI systems constitute the primary focus of this thesis and are therefore analysed in more detail. Stand-alone AI systems qualify as high-risk, when, considering both the severity and likelihood of potential harm, their intended purpose creates a high risk to health, safety or fundamental rights.[90] The AI Act distinguishes two distinct approaches in this category: A product-related system under Art. 6 (1) in conjunction with Annex I is a high-risk AI systems if it is a product or a safety component of a product already subject to third-party conformity assessment under Annex I harmonization legislation.[91] The Act differentiates between products governed by the New Legislative Framework, to which the Act applies directly (e.g. machinery,

[84] AI Act (n 66) recital 28.
[85] AI Act (n 66) art 5 (8); AI Act (n 58) recital 45.
[86] AI Act (n 66) art 5.
[87] Commission, 'Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act)' C(2025) 5052 final, 29 July 2025.
[88] Paul Voigt and Nils Hullen, *Handbuch KI-Verordnung: FAQ zum EU AI Act* [Handbook on the AI Regulation: FAQ on the EU AI Act] (Springer 2024) 30.
[89] Voigt and Hullen (n 88) 8.
[90] AI Act (n 66) recital 52.
[91] AI Act (n 66) art 6 (1).

toy, medical devices), and products regulated outside the NLF, which are subject only to the high-risk requirements through the relevant sector-specific legislation (civil aviation, marine equipment, agricultural vehicles).[92] The second, and for the purpose of this thesis, more significant category are context-related AI systems: under Art. 6 (2) in conjunction with Annex III, an AI-system is considered to be high-risk if it falls within one of the eight listed categories, regardless of any link to a product regulated under Art 6 (1).[93] The categories are as follows:[94]

1. **Biometrics**, in so far as their use is permitted under relevant Union or national law
2. Safety components in **critical infrastructure**
3. **Educational and vocational training**
4. **Employment**, workers' management and access to self-employment
5. Access to and enjoyment of **essential private services** and **essential public services** and benefits
6. **Law enforcement**, in so far as their use is permitted under relevant Union or national law
7. **Migration, asylum and border control management**, in so far as their use is permitted under relevant Union or national law
8. **Administration of justice** and **democratic processes**

Nevertheless, the use of AI in critical Annex III areas does not automatically entail classification as high-risk, as Art. 6 (3) carves out AI systems that do not pose significant risks to health, safety or fundamental rights.[95] This is the case if an AI

[92] Voigt and Hullen (n 88) 50.
[93] AI Act (n 66) art 6 (2).; Voigt and Hullen (n 88) 54.
[94] AI Act (n 66) annex III
[95] AI Act (n 66) art 6 (3).

system merely performs narrow procedural tasks, improves the result of a previously completed human activity, detects decision-making patterns without replacing human review or performs a preparatory task.[96]

Providers of high-risk AI systems are subject to a large range of obligations spanning the entire system lifecycle, including (but not limited to) risk management (Art. 9), data governance (Art. 10), technical documentation (Art. 11), record-keeping (Art. 12), transparency and provision of information to deployers (Art. 13), human oversight (Art. 14), and accuracy, robustness and cybersecurity requirements (Art. 15).[97] Deployers, in turn, bear complementary obligations that are mainly regulated in Art. 26 and Art. 27, such as taking appropriate technical and organisational measures and assigning human oversight.[98] Within this broad compliance framework, only certain provisions, notably Art. 9, 10, 14, 15 and 86, are directly relevant to algorithmic bias and will therefore serve as the basis for the analysis in Chapter 6.

6. Regulation of Algorithmic Bias under the AI Act

Before analysing the Act's provisions in detail, it is important to recall what the Act regulates. As stated briefly in previous chapters, the Act does not define or prohibit discrimination as such, it rather regulates bias, a computer-science term.[99] This link is explicitly stated in Art. 10 (2) (f), which requires providers to examine data for "possible biases that are likely to…lead to discrimination under Union law".[100]

[96] AI Act (n 66) recital 53; Voigt and Hullen (n 88) 68-69.
[97] Voigt and Hullen (n 88) 88.
[98] Tim Wybitul, *Arbeitsbuch AI Act: Gebrauchsanleitung für Unternehmen zur Umsetzung der KI-Verordnung* [Workbook AI Act: Instructions for Companies on Implementing the AI Regulation] (Fachmedien Recht und Wirtschaft 2025) 47.
[99] Kgomosotho (n 5) 84.
[100] Ibid; AI Act (n 66) art 10 (2) (f).

The following sections will analyse how this strategy is implemented through three sets of provisions: risk management system (Art. 9), data and data governance obligations (Art. 10), human oversight and feedback loop measures (Art. 14 and 15 (4)). These provisions aim to address bias at a technical level. The last provision to be analysed, Art. 86, operates differently: Rather than addressing bias ex ante, it applies after a decision has been taken by granting individuals a right to an explanation. It is included in this chapter because it constitutes a part of the Act's broader approach at addressing bias and discrimination. To avoid repetition, the provisions will be examined both descriptively and critically in the same. However, whether this form of regulation can reduce discrimination generally throughout the entire AI Act will be discussed separately in Chapter 7.

6.1. Risk Management System (Art. 9)

6.1.1. Scope and Function

High-risk AI systems constitute the primary regulatory focus of the Act. As a means of risk reduction Art. 8-15 contain a range of different requirements for these systems. Art. 9 introduces a risk management system. Recital 46 identifies this provision as the key mechanism that such systems "do not pose unacceptable risks to important Union public interests",[101] while Art. 8 (1) emphasizes that compliance with Art. 9, as well as the other provisions of Chapter III Section 2, is a precondition for the system's admissibility.[102] In other words, Art. 9 does not only prescribe an internal management process, but by requiring providers to reduce a system's risk to an acceptable residual

[101] AI Act (n 66) recital 46.
[102] AI Act (n 66) art 8 (1).; David M Schneeberger, Walter Hötzendorfer and Christof Tschohl, 'Article 9' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), *The EU Artificial Intelligence (AI) Act: A Commentary* (Kluwer Law International 2024) 241.

level, it determines whether high-risk systems may be lawfully placed on the market at all.[103]

The obligation to establish, implement and maintain a risk management system falls on the provider[104], which is part of the broader quality management requirement laid out by Art. 17 (g).[105] Risk itself is defined in Art. 3 (2) as "the combination of the probability of an occurrence of harm and the severity of that harm", whereas the terms "risk management", "established", "implemented" and "maintained" are not defined in the Act.[106]

6.1.2. The Risk Management Process

Art. 9 (3) limits the scope of the entire provision to risks that may be reasonably mitigated or eliminated through the development, design or provision of adequate technical information.[107] Risks outside of this category, which includes those that cannot be mitigated at all, are excluded not only from the obligation to implement risk-control measures but also from the identification and analysis duties as well.[108] This consequence is difficult to reconcile with the preventive function of Art. 9 as well as the entire AI Act.[109]

Art. 9 (2) describes risk management as a "continuous and iterative process" throughout the entire lifecycle of an AI system, consisting of four steps:

1. Risk Identification and Analysis (a): This first step is restricted to risks the system poses to health, safety or fundamental rights (only!) if the system is used

[103] Ibid.
[104] AI Act (n 66) art 16 (c).
[105] AI Act (n 66) art 17 (g); Schneeberger, Hötzendorfer and Tschohl (n 102) 241.
[106] AI Act (n 66) art 3 (2).; Schneeberger, Hötzendorfer and Tschohl (n 102) 241.
[107] AI Act (n 66) art 9 (3).
[108] Schneeberger, Hötzendorfer and Tschohl (n 102) 242.
[109] Ibid.

in accordance with its intended purpose, rather than risks associated with all conceivable uses.[110] This provision is of particular importance to non-discrimination law due to its link to fundamental rights.[111]

2. Estimation and Risk Evaluation (b): The second step includes risks that may arise from both the intended use, as defined in Art. 3 (12),[112] and reasonably foreseeable misuse, as defined in Art. 3 (13).[113] Therefore, this also covers uses that are reasonably predictable from ordinary human behaviour, leaving out risks falling outside the provider's control.[114] This gives rise to an interpretative difficulty, since risk identification is confined to risks from the system's intended purpose, but risk evaluation extends to reasonably foreseeable misuse – yet subparagraph (d) only refers to risks under (a), creating a loophole whereby an entire category of risks is identified and assessed but not required to be mitigated.[115] However, this inconsistency *may* be the result of an editorial oversight in the final text.[116]
3. Evaluation of Other Risks (c): The third step connects the Art. 9 risk management system to the post-market monitoring obligation set out under Art. 72, requiring providers to collect and analyse data after the system has been placed on the market, thereby ensuring that risk management does not end after deployment.[117]

[110] Schneeberger, Hötzendorfer and Tschohl (n 102) 245.
[111] Meding (n 26) 7.
[112] AI Act (n 66) art 3 (12): 'intended purpose' means the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation.
[113] AI Act (n 66) art 3 (13): 'reasonably foreseeable misuse' means the use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behaviour or interaction with other systems, including other AI systems;
[114] Schneeberger, Hötzendorfer and Tschohl (n 102) 246.
[115] Ibid.
[116] Ibid.
[117] AI Act (n 66) art 72 (2).; Schneeberger, Hötzendorfer and Tschohl (n 102) 247.

4. Adoption of Risk Management Measures (d): The fourth and final step follows a predefined hierarchy of measures and must be read in conjunction with Art. 9 (4) and (5). Art 9 (5) requires that (1) risks should be eliminated or reduced "as far as technically feasible" (note here the change from the original phrasing "as far as possible", implying less strict standards)[118] through the design and development of the system, (2) the implementation of mitigation and control measures addressing risks which cannot be eliminated under phase one and (3) that information and, where appropriate, training be provided to deployers pursuant to Art. 13.[119] Thus, the Act follows an "ex ante before ex post" strategy with design-and-development-stage measures taking precedence over deployer-facing measures.[120] The aim lies in reducing each individual hazard as well as the residual risk of the overall system.[121] These stages must be repeated until the risks are watered down to a level deemed acceptable.[122] Art. 9 (4) further requires careful consideration of interactions between risk management measures and other obligations set out in Chapter III as a way to resolve any tension resulting from their interactions.[123]

6.1.3. The Testing Procedure

The risk management process is supported by a testing obligation set out in Art. 9 (6) to (8) to verify which risk management measures are the most appropriate, that the system works consistently with its intended purpose and complies with the other requirements in Art. 8-15.[124] These obligations acknowledge the notion that no AI

[118] Schneeberger, Hötzendorfer and Tschohl (n 102) 250.
[119] AI Act (n 66) art 9 (5).
[120] Schneeberger, Hötzendorfer and Tschohl (n 102) 249.
[121] AI Act (n 66) art 9 (5).
[122] Schneeberger, Hötzendorfer and Tschohl (n 102) 250.
[123] AI Act (n 66) art 9 (4); Schneeberger, Hötzendorfer and Tschohl (n 102) 251.
[124] AI Act (n 66) art 9 (6).

system can be assumed to be entirely free of risks prior to deployment and therefore testing obligations serve to detect and mitigate residual risks before the system enters into use.[125] Testing may include testing in real-world conditions in accordance with Art. 60 and is required whenever it is necessary to identify the most appropriate and targeted risk management measures.[126] Testing may occur at any stage of system development, but in any event before it is placed on the market.[127] Testing must be assessed against "prior defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the system".[128] These are terms that the Act itself does not explain nor provide guidance on, thereby leaving providers to select them on a case-by-case basis.[129]

6.1.4. Complementary provisions

Art. 9 (9) requires providers to give specific consideration to whether the system is likely to adversely affect persons under the age of 18 and, as appropriate, other vulnerable groups (e.g. ethnic minorities).[130] Lastly, Art. 9 (10) addresses the relationship between Art. 9 and other risk management requirements set out in Union law, insofar as the Art. 9 requirements may be integrated into or combined with the already existing regimes to avoid a duplication of compliance duties.[131]

6.1.5. Interim Conclusion

Art. 9 aims to target bias arising at the planning stage of an AI system.[132] Read together, the ambiguities present throughout Art. 9 leave providers with considerable room for interpretation, most notably through the scope-limiting effect of Art. 9 (3) as

[125] Schneeberger, Hötzendorfer and Tschohl (n 102) 252.
[126] AI Act (n 66) art 9 (7); Schneeberger, Hötzendorfer and Tschohl (n 102) 252.
[127] AI Act (n 66) art 9 (8).
[128] Ibid.
[129] Schneeberger, Hötzendorfer and Tschohl (n 102) 253.
[130] AI Act (n 66) art 9 (9).
[131] AI Act (n 66) art 9 (10).; Schneeberger, Hötzendorfer and Tschohl (n 102) 254.
[132] Kgomosotho (n 5) 86.

well as the potential regulatory gap through the interaction between Art. 9 (a), (b) and (d). Since Art. 9 ultimately determines whether an AI system is deemed acceptable to be put on the market, these uncertainties and gaps may impact the level of protection Art. 9 can provide.

6.2. Data and Data Governance (Art. 10)

Art. 10 (1) requires that high-risk systems developed using data-driven techniques, such as machine learning, be built on training, validation and testing data sets which meet the quality standards set out in paragraphs 2 to 5, whenever such data sets are used.[133] This provision constitutes the Act's key mechanism for addressing bias at the data/development level as discussed in Chapter 4.2.

Art. 10 applies to a wide and non-exhaustive range of data-driven techniques, including supervised, unsupervised, semi-supervised, reinforcement, transfer and self-supervised learning.[134] Paragraph 6 extends application of paragraph 3 to high-risk systems not using model training at all, such as rule-based systems.[135]

The Act defines training, validation and testing data in Article 3 (29) – (32), with testing data specifically defined as data used confirm expected performance before a system is placed on the market.[136] This is not reconcilable with the wording in paragraph 1 that quality criteria need to be met whenever such data sets are used, as well as with Art. 11 and 15's obligations spanning the entire system lifecycle. If this provision is to be read literally, this tension reflects a broader limitation of Art. 10, which is structurally oriented towards preventing bias before deployment while offering limited mechanisms for addressing bias emerging thereafter, such as those discussed in Chapter 4.3.

[133] AI Act (n 66) art 10 (1).
[134] Yordanova (n 78) 267.
[135] Yordanova (n 78) 268.
[136] AI Act (n 66) art 3 (32).

6.2.1. Data Governance and Management Criteria

Art. 10 (2) requires data sets to be subject to data governance and management practices appropriate to the system's intended purpose, set out across subparagraphs a to h. These can be categorised by data lifecycle: data creation (relevant design choices (a), data collection processes and origin (b)), data processing and storage (data-preparation processes (c) and assessment of availability, quantity and suitability (e)) and data usage (formulation of assumptions (d) and identification of data gaps (h)).[137] Since these provisions do not impose any direct obligations concerning non-discrimination, they will not be discussed further.[138] Paragraphs (f) and (g) require the examination of data sets for possible biases likely to affect health, safety, fundamental rights or lead to discrimination prohibited under Union law (with particular attention to situations where data outputs influence future data inputs – feedback loops), as well as the adoption of appropriate measures to detect, prevent and mitigate any biases identified in the previous stage. Bias testing is subject to the documentation requirements laid down in Art. 11, which is essential for ensuring accountability.[139] Paragraphs (f) and (g) do not fit in the conventional data lifecycle but rather constitute an additional stage "bias management and ethical compliance".[140] The term "bias" is not explicitly defined in the Act, despite the varying definitions across different disciplines, such as deviation from rational judgement in psychology or systematic errors in statistics.[141] Because different disciplines will call for different mitigation measures, this terminological uncertainty hinders provider's ability to determine which measures they must take under (g) and (g).[142] Art. 10 (2) addresses biases that "are

[137] Yordanova (n 78) 269.
[138] Meding (n 26) 8.
[139] Ibid; Zuiderveen Borgesius and others (n 59) 8.
[140] Yordanova (n 78) 269.
[141] Yordanova (n 78) 272.
[142] Yordanova (n 78) 273.

likely" to affect fundamental rights or cause discrimination, leaving unclear whether certainty or probability of harm is required, unlike Art. 9 (2) (a), thereby creating regulatory uncertainty and divergence within the Act's own risk vocabulary.[143]

6.2.2. Quality Requirements for Data Sets

Art. 10 (3) requires data sets to be relevant, sufficiently representative, to the best extent possible free of errors and complete in view of the intended purpose.[144] Art 10 (4) requires that data sets take into account the characteristics or elements particular to the geographical, contextual, behavioural or functional setting in which the system is intended to be used as well as regular updates and feedback evaluation to reflect ongoing developments.[145] When compared to general data quality frameworks commonly used in literature, which typically distinguish between relevance, completeness, accuracy, consistency, validity, reliability and timeliness, the requirements set out in Art. 10 (3) primarily *only* address relevance, completeness and accuracy.[146] This omission does not address the potential negative impact of data entry errors, system glitches or even intentional falsification.[147] Additionally, precise explanations of what is considered to be "sufficiently representative" and "to the best extent possible" are not found in the Act, creating legal uncertainty for providers.[148]

6.2.3. Processing of Special Categories of Personal Data for Bias Correction

Art. 10 (5) permits providers of high-risk systems to process special categories of personal data, as defined in Art. 9 (1) GDPR[149] (e.g. data revealing racial origin,

[143] Kgomosotho (n 5) 93.
[144] AI Act (n 66) art 10 (3).
[145] AI Act (n 66) art 10 (4); Yordanova (n 78) 274.
[146] Ibid.
[147] Ibid.
[148] Yordanova (n 78) 275.
[149] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, art 9.

political opinions), where strictly necessary for the purposes of bias detection and correction, subject to appropriate safeguards for fundamental rights and freedoms.[150] Considering that Recital 41 clarifies that the AI Act does not constitute a legal basis for such processing, providers must identify a separate legal basis under Art. 6 (1) GDPR and an exception under Art. 9 (2) GDPR before relying on Art. 10 (5).[151]

The rationale underlying Art. 10 (5) is that biases affecting certain groups cannot be detected and corrected using only non-sensitive data, with existing research already proving that processing special categories of personal data can be instrumental in mitigating bias.[152]

Art. 10 (5) (e) requires special categories of personal data to be deleted once bias has been corrected or the retention period has ended. But because bias monitoring is regarded to be a continuous process, as established in previous chapters, it is difficult and probably inaccurate to determine the precise moment at which bias has been corrected.[153] Additionally, Art 10 (5) leaves room for exploitation, since providers may process sensitive data under the cover of bias correction.[154]

6.2.4. Further Limitations and Interim Conclusion

Art. 10 primarily aims to target bias emerging at the development phase, with the exception of paragraph 2 (f) addressing feedback loops, which emerge in the use stage. A more fundamental limitation arises from the fact that even if all the Art. 10 obligations are met, no dataset can ever be truly unbiased, because even if the requirements of Art. 10 eliminate biased labelling or unrepresentative, biased data one

[150] AI Act (n 66) art 10 (5).
[151] Yordanova (n 78) 277-279.
[152] Kgomosotho (n 5) 98.
[153] Yordanova (n 78) 280.
[154] Kgomosotho (n 5) 102.

fundamental problem remains: The Proxy Problem.[155] As discussed in earlier chapters, access restriction to protected characteristics does not eliminate bias, since the system can identify alternative variables which reveal the same information. This supports the argument that however clean data may be, "all data is dirty" and the requirements set out in Art. 10 therefore fall short of completely erasing bias in data.[156] Additionally, bias is introduced throughout the entire system lifecycle, as established in Chapter 4: via collection, labelling, feature selection and after deployment through feedback loops, emergent bias or automation bias. This directly impacts how much Art. 10 with its narrow focus on input data can realistically achieve.[157] This, alongside the various ambiguities throughout the provision discussed above, suggests that Art. 10 constitutes a set of technical obligations that play a meaningful role in addressing bias, but in itself is not sufficient.[158]

6.3. Human Oversight (Art. 14) and Accuracy, Robustness and Cybersecurity (Art. 15)

6.3.1. Human Oversight

Art. 14 requires high-risk systems to be designed and developed in such way that they can be effectively overseen by natural persons during their use, which reflects the Act's anthropocentric rationale.[159] The aim, as stated in Art. 14 (2), is to prevent or minimise risks to health, safety or fundamental rights that may arise when a system is used in accordance with its intended purpose or foreseeable misuse, mirroring the scope of Art. 9 (2) (b).[160] Most relevant to this thesis' focus on bias, Art. 14 (4) (b) offers a definition of automation bias (as discussed in Chapter 4.3.1) as "the possible tendency

[155] Kgomosotho (n 5) 96.
[156] Kgomosotho (n 5) 94.
[157] Kgomosotho (n 5) 95.
[158] Kgomosotho (n 5) 96.
[159] AI Act (n 66) art 14 (1).; Argyri Panezi, 'Article 14' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), *The EU Artificial Intelligence (AI) Act: A Commentary* (Kluwer Law International 2024) 358.
[160] AI Act (n 66) art 14 (2).

of automatically relying or over-relying on the output produced by a high-risk AI system".[161] It requires that deployers enable oversight measures "as appropriate and proportionate to the circumstances",[162] thereby acknowledging the human tendency towards over-reliance on AI system's outputs.[163]

For oversight to be effective, Art. 14 imposes further safeguards: Firstly, systems must be designed with appropriate human-machine interfacing tools to enable interaction between the human and the system.[164] However, rather than mitigating automation bias, human-machine interfaces may also amplify them if not properly designed.[165] Secondly, the provider must identify and build appropriate oversight measures into the system before it is placed on the market[166] and the person overseeing the system must be able to understand its capacities and limitations as well as detect anomalies[167] and disregard, override, reverse[168] or even intervene or interrupt the system output.[169] A further limitation of human oversight is that, while it is attractive to regulators because it is "easily implementable and observable", research shows that humans generally perform poorly as supervisors because of automation bias, alert fatigue or even because they introduce their own biases.[170]

6.3.2. Accuracy, Robustness and Cybersecurity

Art. 15 requires that high-risk systems be designed in such a way that they achieve an appropriate level of accuracy, robustness and cybersecurity throughout their

[161] AI Act (n 66) art 14 (4) (b).
[162] AI Act (n 66) art 14 (4).
[163] Panezi (n 159) 366.
[164] AI Act (n 66) art 14 (1).
[165] Panezi (n 159) 367.
[166] AI Act (n 66) art 14 (3).
[167] AI Act (n 66) art 14 (4) (a).
[168] AI Act (n 66) art 14 (d).
[169] AI Act (n 66) art 14 (e).
[170] Riikka Koulu, 'Proceduralizing Control and Discretion: Human Oversight in Artificial Intelligence Policy' (2020) 27(6) Maastricht J Eur & Comp L 720.; Kgomosotho (n 5) 95.

lifecycle.[171] Accuracy is understood as a measure of how accurately the system predicts the value of data sets, while robustness refers to the system's capacity to continue functioning normally even under "unlikely, anomalous or adversarial conditions".[172] Lastly, cybersecurity means the system's capabilities to withstand attempts to compromise it, such as cyber-attacks aimed at extracting personal data.[173] Because of the ambiguous and unclear wording used throughout the provision (such as "appropriate level" and "as resilient as possible"), which sets only general objectives without defining parameters and because no relevant case law yet clarify what even constitutes an "appropriate level", Art. 15 has been criticized as too open-ended to be effectively implemented.[174]

Art. 15 (4) addresses feedback loops as discussed in Chapter 4.3.3., therefore an aspect of bias distinct from the input-focused obligations found in Art. 10, in that it requires that "high-risk systems which continue to learn after being placed on the market be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures".[175] Thus, unlike Art. 10 (2), concerned with examining a system's input for bias, Art. 15 (4) regulates the system's output so as to avoid it becoming an

[171] AI Act (n 66) art 15 (1).
[172] Jorge Constantino, 'Article 15' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), *The EU Artificial Intelligence (AI) Act: A Commentary* (Kluwer Law International 2024) 378.
[173] Constantino (n 172) 379.
[174] Lucas G Uberti-Bona Marin and others, 'Is Your AI Model Accurate Enough? The Difficult Choices Behind Rigorous AI Development and the EU AI Act' (arXiv, 28 April 2026) https://arxiv.org/abs/2604.03254 accessed 22 July 2026.; Roberta Tamponi, Carina Prunkl, Thomas Bäck and Anna V Kononova, 'Lost in Vagueness: Towards Context-Sensitive Standards for Robustness Assessment under the EU AI Act' (arXiv, 19 November 2025) https://arxiv.org/abs/2511.15620 accessed 22 July 2026.
[175] AI Act (n 66) art 15 (4).

"echo chamber" for bias.[176] The vagueness critique discussed above also extends to Art. 15 (4), due to the ambiguous wording of "as far as possible".[177]

6.3.3. Interim Conclusion

Taken together, Art. 14 and 15 extend the Act's approach to bias beyond risk management and data governance to address bias that may emerge during the use stage of an AI system. Due to the potential of possibly re-introducing bias into a model (Art. 14) as well as interpretative and practical uncertainty in wording (Art. 15), the effectiveness of both provisions remains uncertain.

6.4. Right to an Explanation (Art. 86)

Art. 86 confers on any person subject to a decision taken by the deployer based on the output of a high-risk AI system the right to obtain a "clear and meaningful explanation" of the role of the AI system in the decision-making procedure and the main elements of the decision, where that decision produces legal effects or similarly affects them (such as impacts on health, safety or fundamental rights).[178] In contrast to the provisions discussed above, which target bias at the planning, development and use stages, Art. 86 applies after a decision has been made by providing affected individuals with the tools to identify potential discriminatory outcomes and seek redress.

6.4.1. Personal and Material Scope

As already discussed in Chapter 4 bias can arise throughout the life cycle of an AI system with different players responsible for different stages, yet the obligation to provide a meaningful explanation rests exclusively on the deployer; providers, importers or distributors face no obligations in that matter. This allocation of

[176] Zuiderveen Borgesius and others (n 59) 8.; Meding (n 26) 8-9.
[177] Tamponi, Prunkl, Bäck and Kononova (n 165).
[178] AI Act (n 66) art 86 (1).

responsibility is peculiar, since deployers typically only supply the input data used for a decision, whereas providers are responsible for the system's underlying algorithm, relevant design choices, model assumptions etc. – a deployer may therefore lack the insight to provide a meaningful explanation.[179] The Act offers an explanation for this allocation in Recital 93, indicating that deployers, as the ones actually using the systems, are "best placed to understand how the high-risk AI system will be used concretely and can therefore identify potential significant risks that were not foreseen in the development phase".[180] In order to supply deployers with the information needed to provide an explanation, the transparency obligations under Art. 13 play a crucial role; they require high-risk systems to be designed to ensure a level of transparency that enables deployers to interpret and use the system's output.[181] However, whether the Art. 13 is in practice sufficient depends largely on the type of system deployed; an issue that will be addressed below.[182]

Art. 86 is restricted further to decisions based on the output of a high-risk system listed in Annex III, meaning that other systems capable of affecting fundamental rights (such as content moderation systems) fall entirely outside of the article's scope. Those systems might still produce discriminatory outcomes. This restriction is striking, considering Art. 86 underlying purpose of enabling individuals to make informed decisions about pursuing legal remedies.

6.4.2. Conditions for Application

As stated above, Art. 86 only applies when a decision is based on the output of an AI system and produces legal effects or "similarly significantly affects that person in a way

[179] Maja Brkan and Hana Palčič Vilfan, 'Article 86' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), *The EU Artificial Intelligence (AI) Act: A Commentary* (Kluwer Law International 2024) 1198.
[180] AI Act (n 66) recital 93; Brkan and Palčič Vilfan (n 179) 1199.
[181] Ibid.
[182] Ibid.

that they consider to have an adverse impact on health, safety or fundamental rights".[183] However, there exists ambiguities in both elements: First, it is not clear whether "based on the output" is a two-step process, whereby humans rely on the outcome of an AI system when making their own decisions, or whether it also extends to decisions made entirely by AI systems without any human intervention whatsoever.[184] Second, the notion of "similarly significant effects" also carries uncertainties, namely how closely the significant effect must be equivalent to the legal effect, why it is restricted to the domains of health, safety and fundamental rights, whether a similarly significant impact might not already amount to a legal effect and whether the wording "in a way that they consider" imports a subjective perception rather than relying on objective parameters.[185]

6.4.3. Defining a Clear and Meaningful Explanation

The Act requires the explanation offered under Art. 86 to be clear and meaningful, covering both the role of an AI system in the decision-making procedure as well as the main elements of the decision taken.[186] The latter is not defined in the Act, but covers, inter alia, the algorithm types, their underlying logic, input and training data as well as personal data used, the weight accorded to different data points and the relationships formed between categories of data.[187] The Act does not provide guidance on the level of detail required for the explanation, which may in part due to the fact that the more complex an AI system becomes, the more detailed and therefore less understandable

[183] AI Act (n 66) art 86 (1).
[184] Brkan and Palčič Vilfan (n 179) 1203.
[185] Brkan and Palčič Vilfan (n 179) 1204-1205.
[186] AI Act (n 66) art 86 (1).
[187] Brkan and Palčič Vilfan (n 179) 1207.

an explanation is.[188] This, however, may result in "inconsistent application of the right across different jurisdictions, AI systems or contexts".[189]

As already hinted at above, a significant limitation arises from the type of AI system that is deployed. A deployer is much more likely to be able to identify and communicate to individuals certain data sets if the underlying AI system is comparatively simple and interpretable. However, when a system relies on complex neural networks and deep learning models its internal decision-making processes may not be comprehensible to humans (even to the designers and developers), thereby raising the question if such "black-box systems" can even comply with the transparency obligation under Art. 13 and the right to an explanation under Art. 86.[190] Within the context of bias and discrimination, unexplainable black-box systems mean that bias – and as a result discriminatory outcomes – may go undetected, leaving individuals without the possibility of receiving an explanation, and, by extension, without the information necessary to effectively exercise their rights.[191] The development of explainable AI (XAI) and interpretable AI is on the rise as way of addressing this opacity issue. The former aims to provide understandable explanations, and the latter focuses on the design of the AI system as way of making it more understandable to humans.[192] However, XAI has already been criticized as generating outputs that are inconsistent across similar inputs and being sensitive to minor input perturbations, thereby increasing the risk of external manipulation. Despite the risk black-box systems pose,

[188] Kgomosotho (n 5) 107.
[189] Ibid.
[190] Brkan and Palčič Vilfan (n 179) 1207.
[191] Brkan and Palčič Vilfan (n 179) 1208.
[192] Cecilia Panigutti and others, 'The Role of Explainable AI in the Context of the AI Act' in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)* (ACM 2023) 1139.

the Act does not prohibit them, with the singular reference to opacity found in Recital 72.[193]

6.4.4. Interim Conclusion

An analysis of Art. 86 reveals that its practical value in addressing bias is contested by the allocation of responsibilities to deployers rather than providers (or both), its narrow material scope, ambiguities in wording and, most notably, the type of system deployed. The broader relationship between the Act's reliance on explainability as a tool of enforcement and the technical limits in complex AI system are discussed in Chapter 7 within the context of the black-box problem.

7. Synthesis and Critical Assessment

7.1. The Act as a Regulation of Bias, not Discrimination

An overall analysis of the provisions discussed in Chapter 6 reveals that the Act lowers the risk of bias substantially by establishing a sound technical framework. However, one limitation all but one provision (Art. 9, 10, 15 and 86) have in common are ambiguities in wording (see Interim Conclusions to Chapter 6), that leaves room for interpretation. Whether this shared ambiguity ultimately diminishes the Act's effectiveness as a whole cannot be determined on the current evidence and is therefore not resolved here, due to the lack of relevant case, seeing as the requirements for high-risk AI systems are not yet in force. Additionally, following from the analysis in Chapter 4 in conjunction with Chapter 6, all but one form of bias find coverage under the Act's provisions: Emergent Bias. The two other forms of bias arising under the Use Stage (feedback loops and automation bias) are discussed in Art. 10 (2), Art. 14 (5) and Art. 14 (4), yet no provision addresses emergent bias. To

[193] Ibid.

recall, this form of bias arises from the interaction between multiple AI systems. Yet the Act's articles provide obligations and requirements for individual systems. Every other source of bias identified in Chapter 4 is addressed, at least to some extent, in the Act. Whether and to what extent this omission results in discrimination cannot be determined within the scope of this thesis and is a question for future legal research. As for the Act's approach to discrimination, the following can be said:

As already mentioned in Chapter 5, Art. 114 TFEU functions as the legal basis for the AI Act, which enables the adoption of measures designed to secure the establishment and functioning of the internal market. As a consequence, the measures adopted under the Act must be connected to the goal of market integration at least on some level, even though (one of) the stated aims of the Act remains the protection of fundamental rights, as stated in Art. 1 of the Act.[194] This tension between the Act's stated aim of fundamental rights protection, the market-integration logic as the underlying legal basis and the product safety instruments used to pursue this aim has attracted criticism, with some scholars questioning whether Art. 114 TFEU can serve as a sufficient legal basis for a regulation whose actual aim appears to be fundamental rights protection.[195] More central to the point of this thesis is, however, the consequence of Art. 114 as a legal basis, since it commits the Act to the instruments native to market regulation, namely the EU product safety legislation as well as the New Legislative Framework, which work best when targeting technical properties of products.[196] The Act therefore works well to target quantifiable issues; however, more normative, value-driven issues that cannot be captured through computational representation are not adequately addressed. This has a direct effect on how bias and discrimination are targeted in the

[194] van Eecke and Regenhardt (n 67) 9.
[195] van Eecke and Regenhardt (n 67) 10.
[196] van Eecke and Regenhardt (n 67) 9.

Act: Algorithmic bias, as a computer science term, can be regulated much more adequately through product safety mechanisms, since the measures used to combat it are technical and quantitative in nature (as discussed in Chapter 3). Algorithmic discrimination, a legal assessment of an outcome against a prohibited ground, however, is a normative concept that cannot be accurately quantified and, by extension, is not the object of the AI Act.[197] The measures defined in the Act are of a technical nature, and, as Kgomosotho puts it "a purely technical approach cannot fully capture or govern these complex socio-technical interactions that lead to discriminatory outcomes in the real world".[198]

This is illustrated by the provisions analysed in Chapter 6:

Art. 9 introduces a risk management system aimed at identifying and mitigating risks through design, development or provision of adequate technical information. The Act follows a concept of risk rooted in probability and statistics, meaning that a risk can always be calculated using the "combination of the probability of an occurrence of harm and the severity of that harm".[199] It is, however, questionable whether this concept of risk can address more fundamental and non-quantifiable phenomena such as discrimination, emphasizing the argument above.[200]

Art. 10 implements quality requirements for training, validation and testing data sets and therefore governs bias at the input-level but cannot adequately address the proxy problem and therefore falls short of eliminating discrimination that can result from biased datasets.

[197] Kgomosotho (n 5) 86.
[198] Kgomosotho (n 5) 87.
[199] AI Act (n 66) art 3 (2).
[200] van Eecke and Regenhardt (n 67) 16.

Art. 14 and 15 introduce safeguards of human oversight as well as accuracy and robustness, but these measures only address the technical aspects of how an AI system should be designed and developed and not if a system output amounts to discrimination.

Even Art. 86, a provision that (in contrast to the others) is the only one focusing on the system output, is limited to ensuring that a decision taken on the basis of the output is understandable and does not establish whether or not the decision was discriminatory.

When read together, it becomes clear that the Act's provisions, most notably Art. 9, 10, 14, 15 and 86, address the technical phenomenon of bias, yet don't address the normative outcome: discrimination. The Act functions as a technical framework focused on bias mitigation with the aim of ultimately lowering the likelihood of discrimination, but due to its internal logic it cannot achieve a definitive prevention of discrimination as such.[201] Even if a system fully complies with the requirements set out in the Act, it may still result in discriminatory outcomes, because none of the provisions are designed to determine or fully prevent discrimination.[202] The Act thus serves only a supporting role in reducing the likelihood of algorithmic discrimination through the technical obligations of Art. 9, 10, 14, 15 and 86; yet the final legal assessment whether discrimination has taken places remains governed by the existing EU non-discrimination framework.[203] However, based on the foregoing analysis, this thesis argues that a genuine tension remains: A regulation that states a "high-level protection of fundamental rights" as one of its core aims, while relying on a framework which falls

[201] Kgomosotho (n 5) 87.
[202] Kgomosotho (n 5) 114.
[203] Kgomosotho (n 5) 87.

short of preventing the harm it seeks to address, gives rise to the question of whether the objective has even been met at all.[204]

7.2. The Burden of Proof

If the Act is not designed to prevent discrimination, this now raises the question of what happens after discrimination has taken place; more specifically, whether affected individuals can even prove that discrimination has occurred and exercise their rights effectively.

Supporting the argument of the Act as a product-safety regulation compared to fundamental rights document, affected individuals need to turn to remedies laid out in Union and national law in order to succeed with a claim of discrimination;[205] first, lodging an internal complaint with the respondent and second, in case of failure, initiating formal legal proceedings.[206] The legal process follows a sharing of the burden of proof, meaning that the claimant does not need to prove a case of definitive discrimination at the outset, but rather needs to "establish facts from which it may be presumed that there has been direct or indirect discrimination" – a prima facie case.[207] For direct discrimination the claimant must prove that they are members of a protected class, suffered unfavourable treatment and that a causal link connects that treatment and the protected characteristic.[208] In cases of indirect discrimination, the claimant must establish that a certain neutral measure or criterion puts them at a disadvantage and that this disadvantage is likely to affect persons sharing the same specific protected characteristic.[209] Only after this has been established, the burden of proof

[204] AI Act (n 66) art 1 (2) (a).
[205] AI Act (n 66) recital 170.
[206] Kgomosotho (n 5) 109.
[207] Julie Ringelheim, 'The Burden of Proof in Antidiscrimination Proceedings: A Focus on Belgium, France and Ireland' [2019] 2 European Equality Law Review 49, 51.
[208] Ringelheim (n 207) 52.
[209] Ringelheim (n 207) 54.

moves to the respondent, who can now prove that this unfavourable treatment did not take place or, in cases of indirect discrimination, that it is objectively justified by a legitimate aim.[210] Within the context of AI decision-making, the standards of indirect discrimination are applied (its application not without faults: see Chapter 3).[211] Thus, in order to establish a prima facie case of discrimination claimants need to meet an initial threshold and in order to do that, they require information. This is where Art. 86 of the Act, as discussed in Chapter 6.4, plays a crucial role, since it provides claimants with the right to a clear and meaningful explanation. However, its effectiveness is limited by three major problems that arise in its practical application:[212]

First, the black-box problem discussed in Chapter 6.4: Where an AI system's internal decision-making processes are not comprehensible, affected individuals are unable to pinpoint the exact source of bias, the criteria the system relied on, and the weight assigned to those factors.[213] More importantly, this leaves individuals unable to challenge discriminatory outcomes, since they cannot meet the initial threshold necessary to establish a prima facie case.[214] Art. 86, which is examined in detail in the chapter above, is meant to address this issue of opacity, but due to the inherent nature of the black-box phenomenon as well as the provision's own limitations, such as the lack of guidance on the level of detail required, the allocation of responsibilities and the scope restriction to Annex III high-risk systems, it cannot be relied upon to effectively address the opacity and burden of redress problem. Additionally, XAI as a technical approach is unable to "resolve fundamentally normative issues rooted in value conflicts" (for further limitations of XAI see Chapter 6.4).[215]

[210] Ringelheim (n 207) 55-56.
[211] Zuiderveen Borgesius (n 10) 19.
[212] Kgomosotho (n 5) 111.
[213] Ibid.
[214] Ibid.
[215] Ibid.

Second, establishing a case of algorithmic discrimination typically requires statistical and technical evidence, such as an AI system's source code, data inputs and other technical information, which is generally regarded as confidential.[216] While Art. 77 does grant national public authorities the power to "request and access any documentation created or maintained under this Regulation", this right lies with public authorities and not with the individual person.[217] Similarly, Art. 86 is unlikely to encompass such detailed technical documentation, leaving the affected person with no means to access the information necessary to build a prima facie case.[218]

Finally, even if a claimant were to successfully obtain all the necessary information, its analysis demands a high level of statistical and technical expertise as well as significant time and resources that individual claimants are unlikely to have. This ultimately translates into high costs for technical support on top of the already expensive legal representation, which may result in the exclusion of affected individuals with insufficient financial means.[219]

Given the substantial burden on the practical availability of redress, the EU proposed a way reducing these difficulties through the AI Liability Directive[220] (AILD), which introduced, inter alia, disclosure obligations, a rebuttable presumption of non-compliance if a defendant fails to comply with an order by a court to disclose or preserve evidence as well as a rebuttable presumption of a causal link.[221] However, this proposition is not without faults, given that the presumption is excluded when a defendant can prove sufficient evidence on the part of the defendant, thereby

[216] Kgomosotho (n 5) 112.
[217] AI Act (n 66) art 77.
[218] Kgomosotho (n 5) 112.
[219] Ibid.
[220] Commission, 'Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (Artificial Intelligence Liability Directive)' COM(2022) 496 final.
[221] Commission, AI Liability Directive Proposal (n 220) art 3-4.

effectively only applying if proving a causal link would otherwise be excessively difficult.[222] Additionally, the scope of the AILD only extends to high-risk AI systems, thus leaving the same regulatory gap for non-high-risk systems identified in Chapter 6.4.1.[223] On February 11th 2025, the Commission disclosed its intent to withdraw the AILD in the 2025 Work Programme due to “no foreseeable agreement”.[224]

Affected individuals will consequently continue to face the burdens of proof discussed above in a discrimination claim, without this, albeit limited, alleviation, thereby leaving one of Act’s central shortcomings unresolved: the opacity of AI decision-making continues to be a hurdle in establishing a prima facie case of discrimination.

8. Conclusion

The AI Act is the first legal instrument to comprehensively regulate AI, a necessary first step in light of its growing societal impact and increasing use in decision-making processes across high-risk areas, which poses substantial challenges in regard to algorithmic bias and discrimination. This thesis explores the relationship between algorithmic bias and discrimination and to what extent the AI Act regulates both. It argues that the Act is designed as a product-safety regulation that addresses the technical phenomenon of bias by introducing technical obligations, namely Art. 9, 10, 14, 15 and 86, though recurrent ambiguities in the provision’s wording as well as a regulatory gap regarding emergent bias leave the potential impact on their effectiveness undetermined. The analysis further demonstrates that a technical framework built around preventing bias, a term rooted in computer science, is not adept at regulating the normative, legal concept of discrimination; neither upstream, at the

[222] Sandra Wachter, 'Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond' (2024) 26 Yale JL & Tech 671, 710.
[223] Ibid.
[224] Commission, 'Commission Work Programme 2025: Moving Forward Together: A Bolder, Simpler, Faster Union' COM(2025) 45 final, annex IV, 11 February 2025.

level of prevention and design, nor downstream, at the level of redress, due to the black-box nature of AI systems impacting the establishment of a prima facie case of discrimination. The Act rather pursues its stated aim of fundamental rights protection through product safety instruments, however, given the two major limitations identified, it is contestable whether it achieves that aim. The relationship between the Act's product safety mechanisms and its Art. 1 objective reveals that the Act is best understood as serving a supporting role by reducing algorithmic bias through technical obligations as a way of attempting to ultimately reduce the likelihood of algorithmic discrimination, while the legal assessment of discrimination remains within the existing non-discrimination framework.

The answers to the unresolved questions identified in this thesis will depend on future legal developments: The withdrawal of the AI Liability Directive leaves the opacity and burden of proof problem unaddressed; however, the Commission reserves the right to assess this issue in the future with another proposal or an altogether different approach. Future developments will need to be monitored. Lastly, the potential impact of the aforementioned ambiguities in wording and regulatory gap for emergent bias on the effectiveness of the individual provisions as well as the entire Act awaits relevant case law and practice once high-risk requirements apply in December 2027.

9. Bibliography

Primary Sources

EU Legislation

Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin [2000] OJ L180/22

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L 2024/1689

Treaties and Conventions

Charter of Fundamental Rights of the European Union [2012] OJ C326/391

Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights)

Case Law

Burden v United Kingdom (2008) 47 EHRR 38

Secondary Sources

Allhutter D and others, Der AMS-Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS) (ITA-Projektbericht Nr 2020-02, Institut für Technikfolgen-Abschätzung, Österreichische Akademie der

Wissenschaften 2020) https://epub.oeaw.ac.at/ita/ita-projektberichte/2020-02.pdf accessed 3 July 2026

Almada and Radu, 'The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy' (2024) 25 German Law Journal 646

Amnesty International, 'Xenophobic Machines: Discrimination Through Unregulated Use of Algorithms in the Dutch Childcare Benefits Scandal' (Index No EUR 35/4686/2021, 25 October 2021)

Arts J and van den Berg M, 'What the Dutch Benefits Scandal and Policy's Focus on "Fraud" Can Teach Us About the Endurance of Empire' (2025) 45(1) Critical Social Policy 177

Barocas S, Hardt M and Narayanan A, Fairness and Machine Learning: Limitations and Opportunities (MIT Press 2023)

—— and Selbst AD, 'Big Data's Disparate Impact' (2016) 104 Calif L Rev 671

Brkan M and Palčič Vilfan H, 'Article 86' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), The EU Artificial Intelligence (AI) Act: A Commentary (Kluwer Law International 2024)

Buolamwini J and Gebru T, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) 81 PMLR 1

Commission, 'Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts' SWD (2021) 84 final https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021SC0084 accessed 5 July 2026

—— 'Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (Artificial Intelligence Liability Directive)' COM(2022) 496 final

—— 'Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act)' C(2025) 5052 final

—— 'Commission Work Programme 2025: Moving Forward Together: A Bolder, Simpler, Faster Union' COM(2025) 45 final

—— 'Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI)' COM(2025) 836 final

—— 'New Legislative Framework' (Internal Market, Industry, Entrepreneurship and SMEs) https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en accessed 11 July 2026

Constantino J, 'Article 15' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), The EU Artificial Intelligence (AI) Act: A Commentary (Kluwer Law International 2024)

'Dutch Rutte Government Resigns over Child Welfare Fraud Scandal' BBC News (15 January 2021) https://www.bbc.co.uk/news/world-europe-55674146 accessed 31 July 2026

Dourish P, 'Algorithms and their Others: Algorithmic Culture in Context' (2016) 3(2) Big Data & Society https://journals.sagepub.com/doi/epub/10.1177/2053951716665128 accessed 3 July 2026

Ebers M, 'Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU's AI Act' (2025) 16 European Journal of Risk Regulation 684

Garg P, Villasenor J and Foggo V, 'Fairness Metrics: A Comparative Analysis' in 2020 IEEE International Conference on Big Data (Big Data) (IEEE 2020) https://doi.org/10.1109/BigData50022.2020.9378025 accessed 3 July 2026

Gerards J and Xenidis R, Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Non-Discrimination Law (European Commission, Directorate-General for Justice and Consumers, Publications Office 2021) https://data.europa.eu/doi/10.2838/544956 accessed 3 July 2026

Goddard K, Roudsari A and Wyatt JC, 'Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators' (2012) 19 J Am Med Inform Assoc 121

High-Level Expert Group on Artificial Intelligence, 'A Definition of AI: Main Capabilities and Disciplines' (European Commission, 18 December 2018) https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf accessed 10 July 2026

Kgomosotho K, 'Analysing the EU AI Act's Treatment of Algorithmic Discrimination' (2025) 9 Univ Vienna Law Rev 76

Koulu R, 'Proceduralizing Control and Discretion: Human Oversight in Artificial Intelligence Policy' (2020) 27(6) Maastricht J Eur & Comp L 720

Lehr D and Ohm P, 'Playing with the Data: What Legal Scholars Should Learn About Machine Learning' (2017) 51 UC Davis L Rev 653

Lieder F and others, 'The Anchoring Bias Reflects Rational Use of Cognitive Resources' (2018) 25 Psychon Bull Rev 322

Lowry S and Macpherson G, 'A Blot on the Profession' (1988) 296 Brit Med J 657

Lum K and Isaac W, 'To Predict and Serve?' (2016) 13(5) Significance 14

Madigan M and others, 'Emergent Bias and Fairness in Multi-Agent Decision Systems' (2025) https://arxiv.org/abs/2512.16433 accessed 11 July 2026

Mayson SG, 'Bias In, Bias Out' (2019) 128 Yale LJ 2218

Meding K, 'It's Complicated: The Relationship of Algorithmic Fairness and Non-Discrimination Regulations in the EU AI Act' (2025) https://arxiv.org/abs/2501.12962 accessed 6 July 2026

OECD, Report on the Implementation of the OECD Recommendation on Artificial Intelligence (Meeting of the Council at Ministerial Level, 2–3 May 2024, C/MIN(2024)17, 2024) https://one.oecd.org/document/C/MIN(2024)17/en/pdf accessed 8 July 2026

Panezi A, 'Article 14' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), The EU Artificial Intelligence (AI) Act: A Commentary (Kluwer Law International 2024)

Panigutti C and others, 'The Role of Explainable AI in the Context of the AI Act' in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23) (ACM 2023)

Ringelheim J, 'The Burden of Proof in Antidiscrimination Proceedings: A Focus on Belgium, France and Ireland' [2019] 2 European Equality Law Review 49

Romeo G and Conti D, 'Exploring Automation Bias in Human–AI Collaboration: A Review and Implications for Explainable AI' (2026) 41 AI & Soc 259

Schneeberger DM, Hötzendorfer W and Tschohl C, 'Article 9' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), The EU Artificial Intelligence (AI) Act: A Commentary (Kluwer Law International 2024)

Skitka LJ, Mosier KL and Burdick M, 'Does Automation Bias Decision-Making?' (1999) 51 Intl J Human-Computer Studies 991

Sosa-Cabrera G and others, 'Feature Selection: A Perspective on Inter-attribute Cooperation' (2023) 17 Int J Data Sci Anal 139

Tamponi R and others, 'Lost in Vagueness: Towards Context-Sensitive Standards for Robustness Assessment under the EU AI Act' (arXiv, 19 November 2025) https://arxiv.org/abs/2511.15620 accessed 22 July 2026

Uberti-Bona Marin LG and others, 'Is Your AI Model Accurate Enough? The Difficult Choices Behind Rigorous AI Development and the EU AI Act' (arXiv, 28 April 2026) https://arxiv.org/abs/2604.03254 accessed 22 July 2026

van Eecke P and Regenhardt B, 'Article 1' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), The EU Artificial Intelligence (AI) Act: A Commentary (Kluwer Law International 2024)

Voigt P and Hullen N, Handbuch KI-Verordnung: FAQ zum EU AI Act (Springer 2024)

Wachter S, 'Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond' (2024) 26 Yale JL & Tech 671

Wybitul T, Arbeitsbuch AI Act: Gebrauchsanleitung für Unternehmen zur Umsetzung der KI-Verordnung (Fachmedien Recht und Wirtschaft 2025)

Yordanova K, 'Article 10' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), The EU Artificial Intelligence (AI) Act: A Commentary (Kluwer Law International 2024)

Zuiderveen Borgesius F, Discrimination, Artificial Intelligence, and Algorithmic Decision-Making (Council of Europe, Directorate General of Democracy 2018)

https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 accessed 3 July 2026

—— and others, 'Non-Discrimination Law in Europe: A Primer for Non-Lawyers' (2024) https://ssrn.com/abstract=4786956 accessed 11 July 2026