Analysing the EU Ai Act's Treatment of Algorithmic Discrimination

Keketso Kgomosotho[*]

Contents

I. Introduction

The use of Artificial Intelligence ("Ai")[1] systems to make and support decisions is increasing throughout various sectors of society. This is because machine learning ("ML"), a subset of Ai, facilitates a level of data processing, analysis, and decision-making precision that promises to revolutionise industries by maximising efficiency,

[*] Keketso Kgomosotho is a doctoral researcher and *Ars Iuris* fellow from South Africa. His current research focus is on the intersection between Machine Learning operational logic, international legal governance and consciousness (qualia) in the context of Ai decision making. He is also an Attorney of the High Court of South Africa.

[1] I employ a lowercase 'i' in the term 'artificial intelligence' throughout this paper to signify that I do not regard these systems as 'intelligent'. This typographical choice to use a lowercase 'i' in 'intelligence' underscores this perspective.

University of Vienna Law Review, Vol. 9 No 1 (2025), pp. 76-123, https://doi.org/10.25365/vlr-2025-9-1-76.

cutting costs, and accelerating production.[2] For instance, Ai algorithms are increasingly used to make and support decisions with a legal or similarly significant impact for individuals in criminal justice,[3] employment,[4] finance,[5] surveillance,[6] law enforcement,[7] education,[8] in autonomous lethal weapons,[9] cyber-attacks and cyber warfare,[10] and the distribution of public services.[11] While on occasion this new level of accuracy and efficiency in decision-making can lead to significant improvements, it is accompanied by significant trade-offs, unprecedented risks and challenges of a novel and unique nature.[12] Notably, this "accuracy is not distributed equally among

[2] AI Index Steering Committee, 'The AI Index 2023 Annual Report' (Stanford University, 4 April 2023 <https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf> accessed 4 May 2025. See also: Zhang, Kirsty and others, 'The AI Index 2024 Annual Report' (AI Index Steering Committee, May 2024) <https://arxiv.org/abs/2405.19522> accessed 4 May 2025.

[3] Dass, Kumar and others, 'Detecting Racial Inequalities in Criminal Justice: Towards an Equitable Deep Learning Approach for Generating and Interpreting Racial Categories Using Mugshots' (2023) *AI&Soc* 897.

[4] Pan and others, 'The Adoption of Artificial Intelligence in Employee Recruitment: The Influence of Contextual Factors' (2022) *IJHRM* 1125; Köchling and Wehner, 'Discriminated by an Algorithm: A Systematic Review of Discrimination and Fairness by Algorithmic Decision-Making in the Context of HR Recruitment and HR Development' (2020) *Bus Res* 795, wherein on the basis of a systematic review of 36 journal articles from 2014 to 2020, the authors review discrimination by algorithmic decision-making in the human resource management context.

[5] Cao, 'AI in Finance: Challenges, Techniques, and Opportunities' (2022) *ACM CS* 64:1.

[6] Kosta, 'Algorithmic state surveillance: Challenging the notion of agency in human rights' (2022) *IJRG*, 16 (224).

[7] Joh, 'The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing' (2016) *10 HL&PR* 15; Citron and Pasquale, 'The Scored Society: Due Process for Automated Predictions' (2014) *WLR*1 (89).

[8] Crompton and Diane Burke, 'Artificial Intelligence in Higher Education: The State of the Field' (2023) 20 *IJETHE* 22; Pham and Sampson, 'The Development of Artificial Intelligence in Education: A Review in Context' (2022) *JCAL* 38 (1408).

[9] Surber, 'Artificial Intelligence: Autonomous Technology (AT), Lethal Autonomous Weapons Systems (LAWS) and Peace Time Threats' Artificial Intelligence, (2018) ICT for Peace <https://ict4peace.org/wp-content/uploads/2018/02/2018_RSurber_AI-AT-LAWS-Peace-Time-Threats_final.pdf> accessed 12 May 2025.

[10] Yamin and others, 'Weaponized AI for Cyber Attacks' (2021) *JISA* 57.

[11] Marwala, '*Closing the Gap: The Fourth Industrial Revolution in Africa'* (2020) *PMSA* 12; Adams, and others, '*Human rights and the fourth industrial revolution in South Africa'* (HSRC, 2021); Cozgarea and others, 'Artificial Intelligence Applications in the Financial Sector,' (2008) *TAEEE*, 12 (57).

[12] Solove, 'Artificial Intelligence and Privacy' (2024) *FLRev* 1 (77).

different demographics because of system bias."[13] As such, one of the more pressing concerns in the context of governing Ai technology is the potential for Ai to perpetuate and amplify existing patterns of discrimination.

For instance, in February 2025, the Netherlands Institute for Human Rights, also known as the Dutch College for Human Rights, concluded that Meta's Facebook algorithm unlawfully discriminated against women by disproportionately showing job advertisements aligned with historic gender stereotypes. The case, brought by NGOs Stichting Clara Wichmann and Global Witness, revealed that Meta's algorithm displayed secretary roles to 85–97% female users and mechanic roles to 96% male users. In its findings, the Dutch College applied Directive 2000/78/EC on Equal Treatment in Employment rather than newer digital regulations like the Digital Services Act ("DSA") or Ai Act. This case is similar to the 2022 case of *Real Women in Trucking v Meta Platforms, Inc.* which involves allegations of systemic gender and age discrimination in job advertisement distribution on Facebook, this time in California. The charge, filed with the USA's Equal Employment Opportunity Commission by Real Women in Trucking, an NGO advocating for women in the trucking industry, similarly contended that Meta's algorithms used to optimise advertising decisions disproportionately steered job advertisements away from women and older individuals, in violation of federal anti-discrimination laws, including Title VII of the Civil Rights Act of 1964, and the Age Discrimination in Employment Act of 1967. The central allegation here is that even when advertisers requested their job ads be shown to all genders, Facebook's ad delivery algorithm independently determined that trucking job advertisements were more relevant to men than women, resulting in discriminatory distribution of ads. Advertisements for traditionally male-dominated jobs like truck driving or mechanics were shown overwhelmingly to men – up to 99% – while ads for traditionally female-dominated roles like administrative assistants were shown predominantly to women. At the time of writing, the case is in the investigatory phase with the EEOC. At first glance, each of the algorithms in question seemed to have complied with non-discrimination law prohibitions by avoiding sensitive data points like sex, gender or age, which are clearly prohibited data points on the basis of which to make decisions. Instead, the algorithmic relies on ostensibly neutral data points which act here as proxies or indirect indicators of said sensitive data points, to determine relevance and distribution patterns. This mechanism is referred to as the Proxy Problem.

[13] Barocas and Selbst, 'Big Data's Disparate Impact' (2016) *CLR* 104 (671-732).

In response to growing concerns around Ai's legal governance more broadly, the EU adopted its landmark Artificial Intelligence Act ("Ai Act" or "the Act") in 2024, becoming the first regional legal system to establish a legal framework for the governance of Ai.[14] Among other things, the Act identifies Ai system bias that is likely to lead to discrimination as a significant risk to fundamental rights in the context of Ai systems. Accordingly, it attends to bias through a multi-faceted approach comprising multiple and mutually reinforcing measures throughout the Act, aimed at minimising the risk of bias, either directly or indirectly (I refer to these measure as "movements").

Following a systematic review of the Ai Act to identify provisions relevant to non-discrimination, this paper unpacks the Ai Act's approach to algorithmic discrimination, concentrated at Article 10 of the Act. The paper is structured around four regulatory movements: (1) the invoking of existing EU non-discrimination frameworks, (2) the emphasis on bias and data quality, and the establishment of data quality criteria, (3) leaving open an exception to GDPR's prohibition of processing special categories of personal data, and finally, (4) actor-specific obligations aimed at making the use and deployment of high risk Ai systems transparent and explainable. By critically examining these movements, this paper assesses the Act's effectiveness in addressing the complex and evolving challenges of algorithmic discrimination, in an attempt to locate the Act within the broader non-discrimination law framework that precedes it. It finds that when it comes to addressing algorithmic discrimination, the Act adopts a strategy deeply rooted in technical requirements and processes, particularly for high-risk Ai systems. While this technical approach is necessary for tackling the unique ways Ai can perpetuate or create unfairness, it inherently limits the Act to a supporting function within the broader architecture of EU non-discrimination law.

II. Understanding Algorithmic Discrimination: The Proxy Problem

To effectively situate the EU Ai Act within the broader non-discrimination landscape, and to effectively assess its contribution therein, we must first understand the problem to which these legal frameworks attend. The proxy problem occurs where seemingly neutral features (or data points) are used as indirect stand-ins or proxy for prohibited grounds, such as ZIP codes, educational institutions, or linguistic patterns, etc. These

[14] Regulation 2024/1689/EU of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts, June 2024, OJ L 252/1.

ostensibly neutral data points serve as statistical proxies or indirect indicators of prohibited grounds, thereby enabling indirect discrimination without explicit use of data on any prohibited grounds. This gives the impression that the algorithm has complied with non-discrimination prohibitions, while in practice, it achieves the very outcome the law sought to avoid. For example, instead of making a prediction based on "gender" due to legal restrictions in the form of prohibited grounds, the algorithms will instead rely on a proxy or indirect indicator of gender, like the applicant's purchasing patterns, membership of a particular group, their name, communication style, or occupation from which to infer their gender. Prohibited grounds (also referred to as protected characteristics or protected attributes in some jurisdictions) are the (non-exhaustive) sensetive personal characteristics or group memberships that non-discrimination law explicitly protects from discriminatory treatment and outcomes.[15] These legally recognised categories vary across jurisdictions and legal instruments, but typically include personal or group characteristics such as race, ethnicity, national origin, sex, gender, religion, disability status, age, and sexual orientation.

The nature of predictive ML algorithms is to find connections (correlation) between input data and target variables, regardless of those connections' normative or legal character (causation).[16] To the algorithms, it matters only that there is a predictive correlation, for example between gender and future income. The potential explanations (causal link) for the relationship between these connections (e.g., patriarchy, hetero-sexism, gender pay gap or history of exclusion) doesn't matter at all.[17] ML algorithms simply do not possess the cognitive infrastructure to understand any other these qualitative, contextual reasons or causalities.[18] Sensitive (prohibited) grounds, or personal characteristics have a highly predictive or probative value in predicting humans' future behaviour, because past and ongoing human

[15] For a discussion on the non-exhaustive prohibited grounds, see Shelton, 'Prohibited Discrimination in International Law' in Aristotle Constantinides and Nikos Zaikos (eds), *The Diversity of International Law: Essays in Honour of Professor Kalliopi K. Koufa* (Martinus Nijhoff Publishers 2009) 261-292; Fredman, Discrimination Law (Oxford University Press 2011) 139.

[16] EU Artificial Intelligence Act; Anupam Datta and others, 'Proxy Non-Discrimination in Data-Driven Systems' (2017) *arXiv.org> accessed 29 May 2025*; Johnson, 'Algorithmic Bias: On the Implicit Biases of Social Technology' *Synthese* 198 (10).

[17] Johnson, (2021) Synthese 198 (10).

[18] Wachter *et al.*, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) CL&SR 105567; Nishant *et al*, 'The Formal Rationality of Artificial Intelligence-based Algorithms' (2024) JIT 39 (20); Selbst *et al.*, 'Fairness and Abstraction in Sociotechnical Systems' (2019) CFAT' 59.

discrimination has created unequal starting points for different individuals and groups, while different social structures have cemented this unequal status quo. Therefore, restricting access to highly predictive data points makes algorithms statistically less accurate in predicting future human behaviour or outcomes.[19] To compensate for this and maintain statistical accuracy, algorithms rely instead on a "proxy" or an indirect indicator of that prohibited ground or restricted sensitive data – thereby promoting the very outcomes the law seeks to avoid. In the discourse, this is called the "Proxy problem."[20]

Moreover, ML algorithms have no cognitive capacity to *understand* meaning of ethics, principles or norms in the same way conscious human minds do. As notes by Deck *et al.*, "legal concepts relying on flexible ex-post standards and human intuition are in tension with the mathematical need for precision and ex-ante standardisation."[21] Unlike human decision-makers, ML decision systesm do not depart from any theory or hypothesis about what types of characteristics may prove useful for predicting the target variable. Rather, it uses "brute force" [22] to learn from scratch which attributes or behaviours predict the outcome of interest. As a formal computational data-processing system, ML algorithms can understand hard, individual technical, empirical, mathematical parameters, to the exclusion of substantive principles and norms, such as equality and non-discrimination.[23] While the prohibited grounds prohibition can target proxy discrimination by human actors (understood as indirect discrimination), it fails in the context of proxy discrimination by Ai data-analysing algorithms. I submit that the approach finds more success with human decision makers, because unlike Ai, human decision makers are conscious entities, with cognitive qualities that allow a capacity to "know" and understand meaning of the prohibitions, to have direct experience and phenomenal awareness of the normative and ethical reasons and causalities between different connections,

[19] Johnson, (2021) Synthese 198 (10), discussing the trade off with accuracy in algorithmic decisions and predictions.

[20] Johnson, (2021) Synthese 198 (10).

[21] Deck and others, 'Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness' (2024) <*arXiv.org*> *accessed 21 March 2025.*

[22] Coglianese and Lehr, 'Transparency and Algorithmic Governance' (2019) 71 *ALR* 1 (15), noting that '[t]he algorithm itself tries many possible combinations of variables, figuring out how to put them together to optimize the objective function.'

[23] Nishant *et al*, 'The Formal Rationality of Artificial Intelligence-based Algorithms' (2024) JIT 39 (20); Xie, 'An explanation of the relationship between artificial intelligence and human beings from the perspective of consciousness' (2021) *CoS* 4(3) (124).

and by extension, the legal and ethical nature of those connections, too.[24] I submit that the law implicitly replies on this quality on the part of the human subject or decision maker – a capacity to "understand" the underlying normative dimensions, reasons, parameters and spirit of the prohibition – in ways that exceed empirical, mathematical, formal fairness. This way, a human decision maker will understand that a prohibition on gender will necessarily include all other indirect indicators of gender, along with the meta-ethics, and contextually-embedded meanings thereof, on a case-by-case basis.

Moreover, as Ai algorithms continue to evolve, they become more sophisticated and complex, and become increasingly adept at identifying and relying on an even broader array of proxies or indirect indicators of prohibited grounds. These proxies grow increasingly more subtle, less obvious and less intuitive to the human mind, often appearing disconnected.[25] For instance, the presence or absence of certain mobile apps, the frequency of software updates on devices, series or music preference, keyboard typing speed, frequency of contact with customer service can all be used to indirectly indicate individual characteristics that are otherwise restricted, such as a person's age, income levels (class), cultural background or risk aversion.

Proxy discrimination is "inherent and ubiquitous" in both algorithmic and human decision-making. In inductive reasoning, we (humans) often rely on some characteristics to "stand in" for other deeper characteristics to which we often don't have access.[26] The same is true for algorithmic inductive reasoning. Take, for example, object recognition algorithms; they are never in contact with the actual objects they aim to identify. Here, the algorithm relies rather on images, where some collection of pixel values will be a proxy for some other feature, such as texture or shape. These in turn act as a proxy for the target attribute, such as a cat, a missile, or a pedestrian.[27] In fact, it is common to train ML algorithms on proxy characteristics that are easier to measure than the characteristics or attributes we want the system to predict. As such, discrimination that results from algorithmic use of proxy characteristics cannot be avoided or mitigated using overt filtering techniques such as

[24] Xie, (2021), 4(3) 1(34).

[25] Datta and others, (2017) *ArXiv*; Prince, and Schwarcz, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data' (2020) 105 *ILR* 1257; Patty and Penn, 'Algorithmic Fairness and Statistical Discrimination' (2022) *PC* 1289 (1); Johnson, (2021) Synthese 198 (11).

[26] For instance, human minds often infer emotions from facial expressions, use money as a proxy for value, or use job titles as proxy for expertise; See Kemp and Tenenbaum, 'Structured Statistical Models of Inductive Reasoning' (2009) *PR* 116 (20); Johnson, (2021) Synthese 198 (10).

[27] Johnson, (2021) Synthese 198 (10).

those of the "prohibited grounds approach" employed by existing EU and national non-discrimination framework. As Goodman and Flaxman put it, "it is widely acknowledged that simply removing certain variables from a model does not ensure predictions that are, in effect, uncorrelated to those variables."[28]

The EU Ai Act is, in part, a response to the problem of bias and discrimination produced by (high-risk) ML algorithms, or more specifically, algorithmic proxy discrimination. This is revealed variously in the Act and its recitals. However, and as is discussed in the next section, discrimination remains a peripheral objective in the Act, receiving limited, often indirect attention from the legislator. The result, regrettably, is that this fundamental challenge to non-discrimination in the form "the Proxy Problem" remains under-attended to under the EU Ai Act.

III. Unpacking Ai Act's Treatment of Algorithmic Discrimination

In its fundamental nature, the Ai Act is an enhanced product safety legislation, following a mixed regulatory approach combining product safety and fundamental rights. It adopts a product safety legislative approach to managing risk – the risk-based approach,[29] wherein risk is defined as "the combination of the probability of an occurrence of harm and the severity of that harm."[30] This approach shapes all obligations and requirements in the Act, bending them towards a fundamentally risk-based orientation, focusing on risk produced by (high-risk) Ai systems. Typically, product safety legislation targets risks to health and safety; however, given Ai's impact on society, the Ai Act includes an additional risk dimension; it attends, in addition, to risks to fundamental rights – becoming an enhanced product legislation.[31]

As made clear repeatedly in the Act, one of its many objectives is the prevention of discrimination produced by Ai systems, a phenomenon otherwise known as algorithmic discrimination. The preamble of the Act consistently emphasises the principle of non-discrimination and acknowledges the associated risks in the context of Ai. However, discrimination remains a peripheral objective in the Act, receiving

[28] Goodman and Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'' (2017) AiMaga 38.

[29] Pouget and Zuhdi, 'AI and Product Safety Standards Under the EU AI Act' (2024) *CE* (11).

[30] This definition aligns with the definition of risk in other NLF legislation, e.g. with Safety Risk Management per ISO Guide 51.

[31] Martens, 'The European Union AI Act: Premature or Precocious Regulation?' (*Bruegel*, 23 May 2024) <https://www.bruegel.org/analysis/european-union-ai-act-premature-or-precocious-regulation> accessed 30 June 2024.

only limited, indirect attention from the Act. This approach, read with the Act's focus on (technical) bias rather than discrimination, positions the Act in a supporting, rather than primary role when it comes to the governance of algorithmic discrimination, refraining from duplicating existing Union non-discrimination frameworks. The core legal definitions of prohibited discrimination, the grounds on which it is prohibited, the burden of proof in legal proceedings, and the remedies available to victims of discrimination are all contained within existing Union non-discrimination directives and related jurisprudence. These frameworks provide the comprehensive legal infrastructure for challenging and redressing discrimination, including that facilitated by Ai. The Ai Act does not replicate this infrastructure, nor does it establish a parallel system for adjudicating discrimination claims. These established (traditional) frameworks remain the primary legal avenues for individuals to challenge discriminatory treatment, including that which is caused by ML decision systems. Instead, the Act explicitly invokes "discrimination prohibited under Union law" at Article 10(2)(f), signalling that the legal standard for what constitutes discrimination remains firmly within the existing body of EU non-discrimination directives. The interaction and relation between the Act and existing non-discrimination frameworks begins to clarify.

The Act's contribution is to provide specific rules tailored to the unique technical characteristics of Ai systems that are likely to give rise to discriminatory risks at the output stage. To that end, the Act has identified a number of specific risks, including risks of "biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law," as stated in Article 10(2)(f).[32] In the latter section, the Act creates a link between the different concepts of "bias" in computer science and "discrimination" in law. It regulates at the level of bias to support efforts to prevent discrimination.

Kristof Meding's analysis offers us a useful framework for understanding the Act's attempt at regulating bias to support the achievement of non-discrimination in Ai systems.[33] Meding argues that there is a difference between the concepts of "bias," and "algorithmic fairness" as understood in computer science, and "non-discrimination" within the domain of law. Terms like "bias" and "fairness" in the computer science context do not fully align with their legal counterparts. Bias in a dataset (input level), for instance, might be a statistical imbalance or disparity, but it

[32] See Article 10(2)(f) of the Ai Act.

[33] Meding, 'It's complicated. The relationship of algorithmic fairness and non-discrimination regulations in the EU AI Act' (2025) <*arXiv.org* 2501.12962v2> accessed 12 June 2025.

translates into legal discrimination only if it leads to unjustified differential treatment on the basis of a prohibited ground (output). While both domains are broadly moving towards the same goal of attending to unfair outcomes, their approaches diverge. Algorithmic fairness, bias mitigation and debiasing techniques are rooted in computer science. They primarily focus on the formal, quantitative technical aspects of detecting, preventing, and mitigating statistical imbalances within data sets used to train algorithms, relying on statistical and mathematical metrics, formulas and rules.[34] Much of these efforts are concentrates at the input level, concerned with what goes into the system.

Non-discrimination, in contrast, is premised in normative, value-driven principles. It is concerned with protecting individuals and groups from objectively unjustifiable differential treatment based on selected protected grounds such as race, ethnicity, gender, religion, disability, age, or sexual orientation. The test for discrimination under non-discrimination frameworks involves analysing the outcome of a practice or decision. The jurisprudence of the ECtHR sets the test as follows: "a difference in the treatment of persons in relevantly similar situations… is discriminatory if it has no objective and reasonable justification; in other words, if it does not pursue a legitimate aim or if there is not a reasonable relationship of proportionality between the means employed and the aim sought to be realised."[35] In *Belgian Linguistics*, the ECtHR established that an objective and reasonable justification is established where the measure in question has a legitimate aim and there is "a reasonable relationship of proportionality between the means employed and the aim sought to be realised."[36]

The Act's primary strategy, as reflected in Article 10 and other related provisions concerning data quality and risk management, leans heavily on a technical paradigm of bias detection and bias mitigation. In practice, this occurs through formal, mathematical fairness metrics, debiasing techniques, etc., which employ quantitative methods and metrics to assess and correct imbalances in datasets and algorithmic outputs. The focus is on ensuring that training, validation, and testing data are of high

[34] Barocas and others, '*Fairness and Machine Learning: Limitations and Opportunities'* (2023) *MITP*, Lamba and Ghani, 'Empirical Observation of Negligible Fairness–Accuracy Trade-Offs in Machine Learning for Public Policy' (2021) *NMI* (10) 896; Selbst *et al.*, 'Fairness and Abstraction in Sociotechnical Systems' (2019) *CFAT* 51.

[35] *Burden v The United Kingdom* App no 13378/05 (ECtHR [GC], 29 April 2008), para. 60; decisions of the ECtHR can be accessed via https://hudoc.echr.coe.int/eng with their case number or party names; *Guberina v Croatia* App no 23682/13 (ECtHR, 22 March 2016), para. 69; *D.H. and Others v the Czech Republic* App no 57325/00 (ECtHR [GC], ECHR 2007-IV), para. 175.

[36] *Belgian Linguistics v Belgium* App nos 1474/62, 1677/62, 1691/62, 1769/63, 1994/63, 2126/64 (ECtHR, 23 July 1968).

quality, relevant, and sufficiently representative, and that measures are in place to identify and address statistical biases within these datasets that are likely to impact non-discrimination prohibited under Union law.

This approach is categorically not sufficient on its own to guarantee non-discrimination. Bias arises not just from biased data, but also from the design of the algorithm itself, the choice of objectives, the context of deployment, and the interaction of the Ai system with existing societal inequalities. Accordingly, the Act identifies 3 sources of this risk of discrimination: bias in the data used to train the Ai system;[37] the design of the algorithm itself;[38] and the way the Ai system is deployed and used, including the context.[39] However, and to the point of this paper, bias arises from a more fundamental location: ML's confinement to formal operational logic and techniques, which cannot simulate components of non-discrimination and fairness. Non-discrimination and equality are not merely a matter of statistical parity or the absence of bias in data, a position shared by Meding and a broad range of cross disciplinary experts, spanning law, technology, data ethics, statistics, information systems, sociology, and computer science .[40] Instead, it is a substantive, normative concept rooted in dignity equality principles and context.[41] Non-discrimination involves normative judgments about fairness, proportionality, and the legitimacy of differential treatment in specific contexts, which cannot be reduced solely to statistical properties of the input data. This more fundamental limitation explains and accounts for algorithms' reliance on proxies, and remains unresolved even with hygienic,

[37] This is evidenced by a heightened focus on data quality, (see Article 10 of the EU AI Act generally). See also Recital 67, which highlights that biases can be inherent in underlying data sets and emphasise the importance of high-quality data sets for training Ai systems, particularly to avoid perpetuating and amplifying existing discrimination.

[38] See for example Article 9 of the EU Ai Act which requires a risk management system for high-risk Ai systems, which includes identifying and mitigating risks related to the Ai system's design. While Article 15 requires that high-risk Ai systems be designed and developed in a way that ensures their accuracy, robustness, and cybersecurity.

[39] To that end, Article 10(4) requires that data sets take into account, the characteristics or elements that are particular to the specific geographical, contextual, behavioural or functional setting within which the high-risk Ai system is intended to be used. Accordingly, Article 14 requires appropriate human oversight of high-risk Ai systems. Further, Recital 13 introduces the concept of 'reasonably foreseeable misuse,' in recognition of the fact that the way an Ai system is deployed and used can also lead to discriminatory outcomes. Recital 86 also emphasises the need for deployers to understand the context of use and identify potential risks not foreseen in the development phase.

[40] Wachter *et al.*, (2021) CL&SR (21);; Nishant *et al*, (2024) JIT 39 (20); Lindebaum *et al.*, 'Insights From 'The Machine Stops' to Better Understand Rational Assumptions in Algorithmic Decision Making and Its Implications for Organisations' (2020) *AMR* 45 (247); Selbst *et al.*, (2019) CFAT' 59.

[41] Wachter *et al.*, (2021) CL&SR (21).

representative data, fairness metrics or debiasing techniques. The latter, while useful for identifying potential biases in data sets, cannot capture the full spectrum of discrimination, or its principle of substantive quality, as understood in law juridically or jurisprudentially. Different, often contradictory, fairness metrics exist. However, research is already finding that they are insufficient to meet the requirements of non-discrimination law – for the perhaps simple reasons that fairness or equality cannot be simulated through formal techniques like fairness metrics in ML.[42]

Bias in an Ai system's data or design is a technical phenomenon that can cause discrimination, while, discrimination itself is a legal concept defined by its impact and lack of legal justification. The Act's technical measures target the cause (bias), but the legal frameworks for non-discrimination govern the effect or discriminatory outcomes. A purely technical approach, even one that can process sensitive data for bias detection, cannot fully capture or govern these complex socio-technical interactions that lead to discriminatory outcomes in the real world. The normative landscape of equality, non-discrimination and fairness in democratic societies is inherently substantive, qualitative, contextual, causal and meaning-laden. This becomes clearer when considered in light of the specific proxy mechanism employed by ML operational logic.

Thus, in essence, the Ai Act contributes as a layer of Ai-specific regulation that imposes upstream obligations on providers of high-risk Ai systems to reduce the *likelihood* of discrimination occurring. However, the ultimate legal assessment of whether discrimination has taken place, the grounds on which it is prohibited, and the available remedies are still governed by the foundational Union non-discrimination frameworks. To show it, the Act itself does not require Ai system to be unbiased or discrimination to be eliminated. By refraining from creating a new, comprehensive legal framework for non-discrimination in the Ai context, the Act structurally avoids duplicating the extensive body of existing Union non-discrimination law. Rather, it provides tools and obligations[43] that are necessary to identify and address these Ai-specific technical risks contributing to the problem of discrimination.[44]

[42] Wachter *et al.*, (2021) CL&SR; Nishant *et al.* (2024) JIT 39 (21);

[43] Like those in Article 10, including the ability to process sensitive data under strict conditions via Article 10(5)

[44] These are design requirements for high-risk Ai systems; actor specific obligations; including obligations promoting transparency and explainability. This interpretation is shared by Lukas Arnold,

Algorithmic fairness and bias mitigation provide valuable, even necessary tools, however, they cannot be a direct substitute for or equivalent to the normative requirements of legal non-discrimination frameworks. In the context of algorithmic discrimination, the Act cannot make any claim of solving the problem. Its provisions and approach are limited to a supporting role, supporting existing non-discrimination law frameworks. In the sections that follow, I move the analysis to a more granular level, unpacking the Ai Act's treatment of algorithmic discrimination into four key movements and offering a critical analysis of each, in turn, to demonstrate their supporting, rather than primary function in the governance of algorithmic discrimination.

A. Movement 1: Invoking Existing EU Non-Discrimination Law

As part of its strategy of achieving consistency with and avoiding duplication within the EU regulatory landscape,[45] the Act invokes pre-existing EU frameworks that prohibit non-discrimination. At Recital 45[46] to the Preamble, it explicitly provides that it does not affect the application of existing Union law prohibiting discrimination, thereby confirming that existing non-discrimination protections remain in place in the context of Ai.[47] For instance, at Article 10(2)(f), the Act creates an obligation for

'How the European Union's AI Act Provides Insufficient Protection Against Police Discrimination' (2024) *UPCLS* (1).

[45] At para 10 of the Preamble to the Act, it makes clear that it 'does not seek to affect the application of existing Union law governing the processing of personal data. See also para 45, 48, 67, 70 of Preamble.

[46] At para 45 of the Preamble to the Act makes clear that 'Practices that are prohibited by Union law, including ... non- discrimination law, ... should not be affected by this Regulation. While paragraphs 29 (based on guiding principles of the 2019 Ethics guidelines for trustworthy AI developed by the independent AI HLEG) understands 'Diversity, non-discrimination and fairness' as meaning 'AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law. *The opening line of para 67 of the Preamble underscores the importance of High-quality data in ensuring that* 'high-risk AI system performs as intended and safely and it does not become *a* source of discrimination prohibited by Union law.' It goes on to confirms that '*The data sets* should also have the appropriate statistical properties... *with specific attention to the mitigation of possible biases in the data sets, that are likely to ... lead to discrimination prohibited under Union law.'*

[47] See also Recital 7, that the AI Act aims to complement existing Union law, including legislation on fundamental rights and non-discrimination.

providers of high risk Ai systems to examine high risk Ai systems for "possible biases that are likely to ... lead to discrimination prohibited under Union law."[48]

Those prohibitions, under existing law, can be found for instance in various national laws on non-discrimination, EU directives and regulations, as well as in regional European human rights instruments. They include Article 14 of the European Convention of Human Rights,[49] Article 21 of the Charter of Fundamental Rights of the European Union,[50] Article 2 of Directive 2000/43/EC on racial or ethnic origin;[51] Article 1 of Directive 2000/78/EC equal treatment in employment and occupation;[52] Article 4 to Directive 2006/54/EC;[53] Article 4 to Directive 2004/113/EC;[54] and Article

[48] Finally, and to the extent that this reliance on existing union non-discrimination law is unclear, the Act at Article 96(1)(e) creates an obligation for the Commission to develop guidelines on its practical application, including '*detailed information on the relationship of this Regulation with ... relevant Union law.*'

[49] Council of Europe, European Convention on Human Rights, as amended by Protocols Nos. 11, 14 and 15, ETS No. 005, 4 November 1950, <https://www.refworld.org/legal/agreements/coe/1950/en/18688> accessed 29 September 2025, which provides that '[t]he enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.'

[50] Charter of Fundamental Rights of the European Union [2012] OJ C 326/391, which provides that '[a]ny discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.'

[51] 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 L 180/22, which provides that 'the principle of equal treatment shall mean that there shall be no direct or indirect discrimination based on racial or ethnic origin.'

[52] Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation [2000] OJ L 303/16, which 'lay[s] down a general framework for combating discrimination on the grounds of religion or belief, disability, age or sexual orientation as regards employment and occupation.'

[53] Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) [2006] OJ L 204/23, which provides that 'For the same work or for work to which equal value is attributed, direct and indirect discrimination on grounds of sex with regard to all aspects and conditions of remuneration shall be eliminated.'

[54] Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services [2004] OJ L 373/37, which provides that 'the principle of equal treatment between men and women shall mean that (a) there shall be no direct discrimination based on sex, including less favourable treatment of women for reasons of pregnancy and maternity; (b) there shall be no indirect discrimination based on sex.'

1 and 2 of Directive Proposal (COM(2008)462).[55] Each of these prohibitions set out a framework for the prohibition of discrimination on the basis of selected prohibited grounds to establish a uniform minimum level of protection within the State, EU and European continent respectively. These (non-exhaustive) lists of prohibited grounds present a "societal rejection" of those grounds as acceptable foundation for differentiation in decision-making.[56]

The EU Ai Act fully accepts and relies upon the definitions, scope, and principles of discrimination as they currently exist within the Union legal order. The technical requirements imposed by the Act, particularly on high-risk Ai systems, are designed to provide duty-holders (like Ai developers and deployers) with specific technical obligations and processes (such as robust data governance and bias mitigation) that help them comply with their pre-existing legal obligations under non-discrimination law. The Act offers Ai-specific methods and standards to reduce the risk of engaging in conduct that would already be considered discriminatory under existing laws. The Act's preventative focus on identifying and mitigating bias at the technical level is a strategy aimed at *avoiding* outcomes that would trigger a violation of existing non-discrimination law. It's about building Ai systems in a way that respects and upholds the principles already enshrined in Union law, rather than defining those principles anew. This ensures coherence within the EU legal system and reinforces that existing non-discrimination laws remain the primary and comprehensive framework for defining and addressing discriminatory conduct, regardless of whether it is facilitated by Ai. Thus, the invocation of pre-existing non-discrimination law frameworks clarifies further the structural positioning and interaction between the Ai Act, and those existing non-discrimination law frameworks.

B. Movement 2: A Focus on Bias and Data Quality

Article 10(1) of the Act establishes a heightened data quality criteria for those Ai systems classified as high risk, to the extent that they rely on data and ML techniques.[57] This quality criteria consists of 3 requirements:

[55] Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation, SEC (2008) 2180, which provides a framework for 'combating discrimination on the grounds of religion or belief, disability, age, or sexual orientation'.

[56] Gerards and Borgesius, 'Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence' (2021) CTLJ 55.

[57] To the extent that the High-Risk Ai system does not rely on data and Machine Learning techniques, paragraphs 2 to 5 apply only to the testing data sets.

1. The "[t]raining, validation and testing data sets shall be subject to data governance and management practices" that must relate to 8 aspects outlined at subparagraph 2(a) to (h).[58] Key of these aspects are items (f) and (g), which respectively provide that these data governance and management practices must concern "examination in view of possible biases that are likely to ... lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations" (feedback loops); and that they must include "appropriate measures to detect, prevent and mitigate possible biases identified according to point (f)." This is a procedural obligation, not a substantive obligation, meaning, to comply fully with this obligation, it is sufficient for a provider or deployer to demonstrate that an appropriate data governance and management practice is in place, and that it relates to the 8 aspects outlined at subparagraph 2(a) to (h).[59]

2. The training, validation and testing data sets must be "relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose." And further that the data sets must be "statistically representative" with regards to persons or groups of persons in relation to whom the high-risk Ai system is intended to be used.[60]

3. Data sets must be contextual, taking into account "the characteristics or elements that are particular to the specific geographical, contextual, behavioural or functional setting within which the high-risk AI system is intended to be used."[61]

[58] Article 10(2)(a)-(h) provides that '*[t]raining, validation and testing data sets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular: (a) the relevant design choices; (b) data collection processes and the origin of data, and in the case of personal data, the original purpose of the data collection; (c) relevant data-preparation processing operations, such as annotation, labelling, cleaning, updating, enrichment and aggregation; (d) the formulation of assumptions, in particular with respect to the information that the data are supposed to measure and represent; (e) an assessment of the availability, quantity and suitability of the data sets that are needed; (f) examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations; (g) appropriate measures to detect, prevent and mitigate possible biases identified according to point (f); (h) the identification of relevant data gaps or shortcomings that prevent compliance with this Regulation, and how those gaps and shortcomings can be addressed.' Own Emphasis.*

[59] That is to say, it is not an obligation of result, rather, its obligation of conduct.

[60] Article 10(3) of the EU Artificial Intelligence Act .

[61] Article 10(4) of the EU Artificial Intelligence Act .

The Act's focus on bias and data quality is premised on the understanding that Ai systems (re)produce discriminatory outcomes when trained with biased and inaccurate data. Garbage in, garbage out.[62] At Recital 67 the Preamble, the Act recognises that "[b]iases can for example be inherent in underlying data sets, especially when historical data is being used, or generated when the systems are implemented in real world settings." Thus, by establishing an obligation that high-risk Ai systems be developed using training, validation, and testing data-sets that "meet the quality criteria referred to in paragraphs 2 to 5",[63] the Act establishes a technical pillar supporting the legal principle of non-discrimination.

The precise nature of the "bias" addressed by Article 10(2) remains a point of contention. As Kristof Meding critically observes, the focus may be overly narrow; "It seems that the regulators had a more technical definition of bias in mind, focusing on the diversity of training data in different dimensions compared to social, ethical, or structural biases."[64] Furthermore, he highlights a crucial consequence: "This makes it very hard to determine the regulatory content" of Article 10(2), pointing to the resulting ambiguity for developers and deployers seeking to comply."[65]

Meding also examines the scope of the phrase "that are likely," arguing that it modifies all subsequent conditions in the sub-paragraph. As Meding notes, it contrasts with the definition of "risk" found elsewhere in the Act, which involves a balancing of both likelihood *and* severity of harm. By suggesting that the "likely" threshold in Article 10(2)(f) primarily concerns the probability of the adverse effects manifesting, Meding implies that this specific obligation for bias examination is triggered by a likelihood assessment, different from a comprehensive risk evaluation triggered by the additional element of severity.[66] "Compared to the risk in Article 9(2) AIA, in Article 10(2)(f) AIA, the severity of the harm is not taken into account, only the likelihood."[67]

[62] Wortham, '*Garbage in, toxic data out: a proposal for ethical artificial intelligence sustainability impact statements*' (2023) *AI&E* 3 (135–1142); Geiger and others, 'Garbage in, Garbage out' Revisited: What Do Machine Learning Application Papers Report about Human-Labelled Training Data?' (2021) *QSS* 2 (795).

[63] Article 10(2) EU Artificial Intelligence Act.

[64] Meding, *arXiv.org*, 10.

[65] Meding, *arXiv.org*, 10.

[66] Ibid. This supported by reference to translational consistency in the German translation.

[67] Ibid.

The requirement to address biases leading to outcomes "that are likely" to negatively affect fundamental rights or constitute discrimination presents an "open question" regarding the necessary threshold – certainty or probability.[68] Meding leans towards a probabilistic reading, arguing it "seems more consistent" with the Act's broader focus on *risk*, particularly when compared to Article 9(2)(a)'s explicit mention of risk to fundamental rights.[69] This interpretation finds additional support through linguistic comparison, specifically in translations like the German text of the same Article 10(2)(f), which makes clear that "it is more evident that the wording 'that are likely' applies to all conditions."[70] This suggests the intent was likely to capture probable harms, aligning with a risk-management framework. Adopting this view means Article 10(2)(f) mandates addressing not just definite, but also probable, discriminatory outcomes or rights infringements stemming from data bias.[71]To the extent that it contrasts with the use of "risk" in other articles, for instance Article 9(2)(a), Meding argues that such discrepancies create "unnecessary regulatory uncertainty," noting that it is "unclear why the legislator opted for the difference in applicability between Article 9 AIA and Article 10 AIA."[72]

What Article 10 offers in a supporting function to non-discrimination law is the provision of enforceable technical standards aimed at preventing discrimination upstream. Article 10 offers concrete, Ai-specific requirements that, if met, reduce the likelihood of an Ai system producing results that would violate those existing legal standards. However, even its supportive function is ultimately limited; non-discrimination law requires the elimination and reduction of discrimination. The measures employed in the Act cannot guarantee non-discrimination. An Ai system can comply with Article 10 and still result in discrimination. The fundamental reason why an Ai system can comply with the technical requirements of Article 10 and still result in discrimination lies in the inherent difference between technical bias detection at the development stage and the complex, context-dependent nature of legal discrimination in real-world deployment. Therefore, while Article 10 and similar technical provisions in the Act are necessary for building Ai systems that are more likely to be fair and non-discriminatory by addressing technical biases at the source, they are not a fail-safe. The possibility remains that even an Ai system

[68] Ibid.

[69] Ibid.

[70] Ibid.

[71] Ibid.

[72] Ibid.

developed in full technical compliance with Article 10 could – due to the inherent complexities of Ai and its interaction with society – lead to discriminatory outcomes.

First, this is because all data is "dirty data." As Meredith Broussard poignantly puts it, "[A]ll data is dirty. All of it."[73] There is no such thing as "clean data" or "unbiased data."[74] The phrase itself is an oxymoron. This is because all data is a product of human language, experiences, cultural artifacts, human interpretation, selection and decision-making at some stage throughout the data life cycle, whether at collection, categorisation, or presentation.[75] As Eaglin contends, data is not just found but selected and crafted with particular value judgments. To illustrate this, she points to the example of input data in ADM systems used to assess risk of recidivism, noting that the data centres on "policy questions about who should be considered a risk and how much risk society tolerates."[76]

Consider the process of data cleaning or processing for example. Once data is collected, it goes through a 'chewing' process, wherein all data is standardised to the algorithm's training requirement, to facilitate the interoperability of data within quantitative frameworks and tools. Data that deviates from the norm or exhibits unique characteristics is often filtered out or refined by human data scientists or programmers to align better with the rest. The process also removes significant portions of the original formatting and context, introducing a profound decontextualization of the data, which in turn has an impact on the interpretation and understanding of this data.[77] As Solovo contends, "[w]hen qualitative data is removed,

[73] Broussard, 'Artificial Unintelligence: How Computers Misunderstand the World' (2018) *MITP* 103. The quote continues "All data is dirty. All of it. Data is made by people going around and counting things or made by sensors that are made by people. In every seemingly orderly column of numbers, there is noise. There is mess. There is incompleteness. This is life."

[74] Delbosc, 'There Is No Such Thing as Unbiased Research – Is There Anything We Can Do about That?' (2023) TR 33 (155); Maatman, 'Unbiased Machine Learning Does Not Exist (LBBOnline' October 2018) <https://www.lbbonline.com/news/unbiased-machine-learning-does-not-exist-3> accessed 19 October 2023; Johnson, (2021) Synthese 198 (10); du Preez, 'AI and Ethics 'Unbiased Data Is an Oxymoron' (Diginomica, 31 October 2019) <https://diginomica.com/ai-and-ethics-unbiased-data-oxymoron> accessed 19 October 2023.

[75] Bower, 'The Nature of Data and Their Collection', in Bower (ed.), *Statistical Methods for Food Science: Introductory Procedures for the Food Practitioner*, 2nd edn. (Hoboken, 2013) 15.

[76] Eaglin, 'Constructing Recidivism Risk' (2017) *ELJ* 67 (59).

[77] Burk, 'Algorithmic Legal Metrics' (2020) *NDLR* 96 (1147).

leaving just quantifiable data, the nuance, texture, and uniqueness of individuals is lost…decisions are being made about people based on a distorted picture of them."[78]

Where the algorithms require labelled data, the individuals labelling the data bring their own (often implicit) biases and interpretations to this process, which will subsequently be learned by the algorithm. Further, the Act itself recognises at Article 10(2) that certain design choices or features chosen to represent the data can further introduce bias into the algorithm, for instance where important features are omitted or where irrelevant features are given undue weight.[79] To this point, Cathy O'Neil correctly notes that development of algorithmic models involves selective attention to certain data while excluding others, highlighting the inherent subjectivity involved in the development of algorithmic systems more broadly. To O'Neil, "[t]hose choices are not just about logistics, profits, and efficiency. They are fundamentally moral."[80]

Those fulfilling the human oversight role or interpreting the algorithmic output/results also bring their own pre-existing conscious and unconscious biases; or if they lack a deep understanding of the algorithm's limitations, they might draw biased conclusions.[81] Barocas and Selbst succinctly summarise the issue: "Big data claims to be neutral. It isn't."[82] As the authors point out, machine learning relies on data collected from society. Consequently, to the extent that society embodies inequality, exclusion, or other forms of discrimination, these biases will inevitably be reflected in the data.

In this way then, every data set carries with it the fingerprints of societal, cultural, and personal biases of those who played a role in its life cycle.[83] In what she calls the "input fallacy," Talia Gillis reminds us that efforts to remove or restrict access to legally protected characteristics, such as race, gender or sexual orientation from input data does not eliminate discrimination and bias in Ai outputs. Gillis correctly observes

[78] Solove, The Digital Person: Technology and Privacy in the Information Age (New York, 2004) 3.

[79] Seijo-Pardo and others, 'Biases in Feature Selection with Missing Data' (2019) *NeuroComp* 432 (97).

[80] O'Neil, '*Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown Pub, 2016) 30.

[81] Mehrabi and others, 'A Survey on Bias and Fairness in Machine Learning' (2021) *ACMCS* 54 (115); Samuel, 'Why It's so Damn Hard to Make AI Fair and Unbiased' (*Vox*, 19 April 2022) <https://www.vox.com/future-perfect/22916602/ai-bias-fairness-tradeoffs-artificial-intelligence> accessed 19 October 2023.

[82] Barocas and Selbst, (2016) *CLR* 671.

[83] Datta and others, (2017) *arXiv.org*.

that these protected characteristics can still be inferred from other available data about the individual, demonstrating that biases are intricately woven into the fabric of the data, beyond direct identifiers, i.e. the Proxy Problem.[84] The takeaway here is that the data driving ML decision systems – while seemingly neutral – are normative; they often reflect biases stemming from historical inequalities, societal norms, and the subjective choices made about which data to proceed without, which weightings to place, or variables to set. This reality undermines the idea that data can be neutral or unbiased.

The Act's primary focus on data quality undoubtedly expands the legal framework for safeguarding non-discrimination in meaningful ways. However, it is clear that is not sufficient against the challenge of algorithmic proxy discrimination. It is seductive and convenient to view the Act's as a complete solution to the problem of algorithmic discrimination. It is not.[85] Data sets that are of "high quality," "sufficiently representative," "free of errors and complete"[86] are simply not sufficient to erase biases in data sets. Neither does the best model design. The most that can be hoped for is a minimisation of system bias, not its elimination. Non-discrimination law sets a higher bar, requiring not merely the reduction of risk, but the elimination or effective mitigation of discriminatory *effects* and *treatment* experienced by individuals and groups based on prohibited grounds. Thus, working at the input level is necessary of course, but it is not sufficient.[87]

C. Movement 3: A New Exception to GDPR

Third, Article 10(5) of the Act opens a marked exception to Article 9(1) of the GDPR, which prohibits the processing of special categories of personal data. The GDPR is a technology-neutral legal framework; we see this in its expansive definition of "processing" which encompasses nearly all operations performed on personal data. As such, it can be inferred that the GDPR's provisions extend to Ai systems, to

[84] Gillis, 'The Input Fallacy' (2022) *MLR* 106 (1175).

[85] Chander, 'EU's AI Law Needs Major Changes to Prevent Discrimination and Mass Surveillance - European Digital Rights' (EDRi, 28 April 2021) <https://edri.org/our-work/eus-ai-law-needs-major-changes-to-prevent-discrimination-and-mass-surveillance/> accessed 29 May 2024.

[86] Article 10(3) of the Artificial Intelligence Act.

[87] While the emphasis in Article 10 is heavily on data inputs and governance, it's important to note that the Act for high-risk systems also includes requirements related to Risk Management System, Quality Management System, and Post-Market Monitoring. Discriminatory outcomes, particularly those impacting fundamental rights, could be monitored under these general provisions. Notably, however, the Act's provisions on output monitoring and real-world impact assessment are less detailed and prescriptive compared to the requirements for input data and development processes.

the extent that personal data is at any point processed during the Ai system's lifecycle.[88] To that end Article 9(1) of the GDPR prohibits, as a general rule, the processing of special categories of personal data. Personal data is defined as those data points that "[reveal] racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation."[89]

Article 10(5) of the Ai Act creates an exception to this prohibition, to allow providers of high-risk systems, in exceptional cases, to process special categories of personal data where it is strictly necessary for the purpose of ensuring bias detection and correction. Guadino calls this Article a "paradigm shift" in the governance of special categories of personal data, remarking optimistically that "[the] paradigm shift is an expression of a fundamental optimism that our social reality can be improved in a sustainable manner through properly regulated AI."[90] Article 10(5) provides a technical tool – the authorisation to process sensitive data – specifically for the purpose of fulfilling a requirement of bias detection and correction under Article 10(2)(f) and (g) that is explicitly linked to preventing discrimination prohibited under Union law.

The rationale of Article 10(5) is clear; you cannot effectively detect or correct biases against specific groups if you cannot analyse data related to those groups. Primarily, the Ai Act introduces this exception because system biases are not, in certain circumstances, detectable using only non-sensitive data.[91] Examining special category data can help reveal these biases by allowing the limited and controlled processing of

[88] Quezada-Tavarez, and other, 'Voicing Challenges: GDPR and AI Research' (2022) *ORE* 2(126); Demircan and Kalyna, 'Europe: The EU AI Act's Relationship with Data Protection Law: Key Takeaways' *Privacy Matters*, April 2024) <https://privacymatters.dlapiper.com/2024/04/europe-the-eu-ai-acts-relationship-with-data-protection-law-key-takeaways/> accessed 1 July 2024.

[89] See Article 9 of GDPR.

[90] Lukas Feiler and others, 'EU AI Act: Diversity and Inclusion Prevails over Data Protection' (*Lexology*, 26 June 2024) <https://www.lexology.com/library/detail.aspx?g=a978bb5a-409a-4b26-8df1-26e3244bd97f> accessed 29 June 2024. As they write 'On the one hand, ignoring personal demographic data promotes the same risk as the widely rejected idea of fairness through unawareness because legally protected attributes like race and gender usually correlate to innocuous proxy variables. If protected attributes are unavailable during model training and evaluation, these subtle correlations cannot be accounted for, nor can technical fairness metrics be tested and optimized.'

[91] Yucer and others, 'Measuring Hidden Bias within Face Recognition via Racial Phenotypes' (2021) *EEE WACV, 2022.*

special category data.[92] In theory the Ai Act aims to enable developers to uncover and address these hidden biases, making them easier to mitigate. Processing special categories of personal data can provide a more nuanced understanding of how Ai systems impact different groups. Deck *et al.*, writing on the implications of the Ai Act for non-discrimination law and algorithmic fairness, contend that the development of fair Ai systems necessitates access to sensitive demographic data to identify and mitigate biases that may correlate with, and act as proxies for, prohibited grounds such as race and gender, thereby moving beyond the limitations of "fairness through unawareness."[93] The authors contend that Article 10(5) of the Ai Act directly addresses this conflict by providing a legal basis for the exceptional processing of special categories of personal data where strictly necessary for the purpose of detecting and correcting bias in high-risk Ai systems, provided adequate safeguards are in place.[94]

There is already evidence that allowing algorithms access to personal sensitive data can help detect and mitigate system biases and discriminatory outcomes.[95] In fact, existing research supports the view that permitting algorithms to access sensitive personal data under strict and exceptional conditions can be instrumental in identifying and mitigating systemic biases which lead to discriminatory outcomes. A notable example is the Gender Shades project, which exposed significant biases in automated facial analysis algorithms and datasets by analysing their performance across different demographic groups.[96] In addition, several organisations are also developed tools and frameworks that leverage sensitive personal data – with appropriate safeguards – to detect and mitigate biases in Ai systems. IBM's Ai

[92] Artzt and Dung, 'Artificial Intelligence and Data Protection: How to Reconcile Both Areas from the European Law Perspective' (2023) VJLS 7(2):39-58; 12/11/2025 15:39:00Paterson and McDonagh, 'Data protection in an era of big data: the challenges posed by big personal data' (2019) *MU* (2).

[93] Deck *et al.*, (2023) *arXiv* (3).

[94] Ibid, as they note, 'discrimination and fairness considerations can provide a justification for data processing during the training phase of high-risk AI systems.'

[95] Marvin van Bekkum and Frederik Zuiderveen Borgesius, 'Using Sensitive Data to Prevent Discrimination by Artificial Intelligence: Does the GDPR Need a New Exception?' (2023) 48 *CL&SR* 48 (105770).

[96] Joy Buolamwini, 'Press Kit' (*MIT Media Lab*) <https://www.media.mit.edu/projects/gender-shades/press-kit/> accessed 20 May 2024.

Fairness 360 toolkit,[97] Google's What-If Tool,[98] and Microsoft's Fairlearn are prime examples of such initiatives.[99] These tools provide developers with the means to assess and improve the fairness of their Ai models by analysing their behaviour across different groups and identifying potential sources of bias.

1. Limitations

Article 10(5) provides a necessary enabler for tackling the proxy problem at the data level, but it is likely not sufficient to fully combat it on its own. Once potential proxies are identified, developers can then assess if these features are contributing to biased outcomes within the dataset or during model training, in line with the requirements of Article 10(2)(f). However, proxies are subtle, context-sensitive and therefore dynamic, involving non-linear correlations between multiple seemingly innocuous features and a protected characteristic.

A feature that acts as a strong proxy for a protected characteristic in one application context or geographical area might not do so, or might proxy a different characteristic elsewhere. For instance, certain linguistic patterns might correlate with origin in one region but not another. An Ai system trained and validated for bias using data from one specific context, relying on the ability to process sensitive data under Article 10(5) to identify proxies within that dataset, might still exhibit discriminatory behaviour when deployed in a different context where the proxy relationships are altered. The input analysis performed under Article 10 provides a snapshot valid for the *specific dataset and context of development*, not a universal guarantee against proxy-driven discrimination in all potential deployment scenarios. Identifying all such complex proxies in high-dimensional datasets can be technically challenging, even with access to sensitive data. The proxy problem is not just about the data itself, but also about *how the algorithm uses and understands* the features.

2. Safeguards

Further, this processing of special categories of personal data must be in accordance with the obligation for data governance and management practices outlined at sub

[97] IBM AI Fairness 360 is an open-source toolkit with algorithms for detecting and mitigating bias in machine learning models.

[98] A visualization tool to explore how machine learning models behave with different data inputs and help identify biases <https://pair-code.github.io/what-if-tool/> accessed 19 May 2024.

[99] Microsoft's Fairlearn is an open-source toolkit to assess and improve the fairness of AI systems: <https://fairlearn.org/> accessed 29 May 2024.

paragraph 2 of the Article, specifically points (f) and (g).[100] The Act makes this exception "subject to appropriate safeguards for the fundamental rights and freedoms of natural persons,"[101] – including requirements of necessity; technical limitations on the re-use of personal data (purpose limitation), privacy-preserving measures and cyber security; strict controls and documentation of the access of this data, confidentiality; 3rd party access, transmission and transfer prohibitions; and requirements for deleting the data once the bias has been corrected or once data retention period has expired (data minimisation).[102] This is in addition to the provisions set out in Regulation (EU) 2016/679, Directive (EU) 2016/680 and Regulation (EU) 2018/1725. Finally, the Act requires the keeping of records of processing activities, as well as "the reasons why the processing of special categories of personal data was strictly necessary to detect and correct biases, and why that objective could not be achieved by processing other data."[103] Here the Act assumes that what providers disclose will be true and correct. However, even with this, the approach raises serious concerns about risks to the privacy and data protection regime because of Ai's exponential capacity for pattern recognition and association. This offers an expanding capacity for data to be combined in a way capable of reliably

[100] While the GDPR provides the general framework for data protection, Article 10(5) acts as a *lex specialis* for the processing of special categories of personal data in the context of Ai. This means that the specific conditions and safeguards of Article 10(5) take precedence over the general provisions of the GDPR in this particular context.

[101] Article 10(5)(a)-(f).

[102] Article 10(5)(a)-(f).

[103] Article 10(5)(a)-(f): '(a) the bias detection and correction cannot be effectively fulfilled by processing other data, including synthetic or anonymised data; (b) the special categories of personal data are subject to technical limitations on the re-use of the personal data, and state of the art security and privacy-preserving measures, including pseudonymisation; (c) the special categories of personal data are subject to measures to ensure that the personal data processed are secured, protected, subject to suitable safeguards, including strict controls and documentation of the access, to avoid misuse and ensure that only authorised persons with appropriate confidentiality obligations have access to those personal data; (d) the personal data in the special categories of personal data are not to be transmitted, transferred or otherwise accessed by other parties;(e) the personal data in the special categories of personal data are deleted once the bias has been corrected or the personal data has reached the end of its retention period, whichever comes first; (f) the records of processing activities pursuant to Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680 include the reasons why the processing of special categories of personal data was strictly necessary to detect and correct biases, and why that objective could not be achieved by processing other data.'

(re)identifying, classifying and clustering individuals based on seemingly unconnected, non-personal pieces of data.[104]

Even with the ability to process sensitive data for bias detection in inputs, the gap between identifying input bias and preventing discriminatory outputs persists. An Ai system might process data containing special categories of personal data, and developers might use this to mitigate statistical biases in the training data. However, the system's behaviour upon deployment – influenced by algorithmic design and context – could still lead to discriminatory outcomes that were not fully eliminated or foreseen during the input-focused bias correction phase, or that were introduced by the phase itself. Moreover, there are compelling incentives for providers to exploit this exception. The ability to process special category data can provide a competitive edge. By analysing this data, providers can develop more accurate and nuanced Ai models, potentially leading to better performance and increased market share.[105] Processing existing special category data under the exception can be a cost-effective way to improve Ai models without investing in additional data collection and labelling efforts, which can be expensive.[106] In more exceptional circumstances, it may incentivise providers to classify their Ai systems as high-risk even if the risks associated with their systems are not significant or to intentionally engineer conditions of bias that make it seem necessary to gain access to this exception and process special category data.

Consider for example enforcement actions against the utilisation of Ai systems by EU data protection authorities, before the adoption of the Act. These enforcement actions are based on a range of issues, including a lack of transparency, lack of legal basis to process personal data or special categories of personal data, failure to fulfil data subject rights, automated decision-making abuses, and data accuracy issues. The

[104] Solove, (2024) *FLR* 1 (77); Staab, Vero, Balunović and Vechev, 'Beyond Memorization: Violating Privacy Via Inference with Large Language Models' (arXiv, 6 May 2024)<https://doi.org/10.48550/arXiv.2310.07298> accessed 4 May 2025 noting that 'It is often technically very difficult to separate personal data from non-personal data.'

[105] Moira Paterson and Maeve McDonagh, 'Data Protection in an Era of Big Data: The Challenges Posed by Big Personal Data' (2018) *MULR* 1 (44).

[106] Ai is currently very expensive to train, especially models that would be defined as high risk. See David Meyer, 'The Cost of Training AI Could Soon Become Too Much to Bear' (Fortune, May 2024) <https://fortune.com/2024/04/04/ai-training-costs-how-much-is-too-much-openai-gpt-anthropic-microsoft/> accessed 10 January 2025; Derek du Preez, 'AI Is Currently Too Expensive to Take Most of Our Jobs, Finds MIT Researchers' (Diginomica, 24 January 2024) <https://diginomica.com/ai-currently-too-expensive-take-most-our-jobs-finds-mit-researchers> accessed 28 May 2025.

most notable are Italian DPA's temporary ban on OpenAI's ChatGPT[107] and its fine against Deliveroo's Ai-enabled automated rating of rider performance,[108] and the French DPA's fine against Clearview Ai for scraping billions of photographs from the internet, including social media platforms, to create a vast database for facial recognition purposes.[109] These already demonstrate a propensity by Ai providers and deployers to exceed the lawful bounds in processing personal data and special categories of personal data, where it's in their business interest to do so.

Moreover, research illustrates that these safeguards are already proving insufficient to protect privacy of personal data in the context of algorithmic data processing and decision making.[110] As regards the first safeguard of necessity, the Act assumes here that the promoted alternative data types (synthetic or anonymised data) can adequately replicate the complexities of real-world data, which is often not the case.[111] Moreover, even though necessity is subjective and open to interpretation, the Act does not provide any guidance or criteria for making this determination, leaving room for potential misuse or abuse of this provision by providers. A more stringent and clearly defined necessity standard, coupled with robust oversight mechanisms, is required given the risks of Ai processing of personal data. Regarding technical limitations, the effectiveness of this safeguard hinges on the robustness of the technical measures implemented and the ability to enforce these limitations. The dynamic nature of Ai, with models constantly evolving and being integrated into new systems, poses a challenge in ensuring that the data remains confined to its intended

[107] For a lack of a suitable legal basis for the collection and processing of personal data for the purpose of training the algorithms underlying ChatGPT. See Natasha Lomas, 'ChatGPT Is Violating Europe's Privacy Laws, Italian DPA Tells OpenAI' (*TechCrunch*, 29 January 2024) <https://techcrunch.com/2024/01/29/chatgpt-italy-gdpr-notification/> accessed 29 May 2025.

[108] 'Italian DPA Fines Food Delivery App 2.6M Euros for GDPR Violations' (IAPP News, 2024)<https://iapp.org/news/b/italian-dpa-fines-food-delivery-app-3m-euros-for-gdpr-violations> accessed 1 July 2024.

[109] 'The French SA Fines Clearview AI EUR 20 Million' (European Data Protection Board, 2022) <https://www.edpb.europa.eu/news/national-news/2022/french-sa-fines-clearview-ai-eur-20-million_en> accessed 1 July 2024.

[110] Staab and others, 'Beyond Memorization' (arXiv, 6 May 2024); Solove, (2024) *FLR* 1 (76)

[111] Stefanie James and others, 'Synthetic Data Use: Exploring Use Cases to Optimise Data Utility' (2021) *DAI* 1 (15); Majeed and Lee, 'Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey' (2021) *IEEEA* 8512, highlighting re-identification methods used by malevolent adversaries to re-identify people uniquely from the privacy preserved data.

use.[112] Additionally, the Ai Act does not explicitly define what constitutes "technical limitations," leaving room for interpretation and potential loopholes that could be exploited.

The Act mandates the use of "state-of-the-art security and privacy-preserving measures," including pseudonymisation, to protect special categories of personal data. Studies are already demonstrating that these measures are not effective in the face of Ai capabilities.[113] Pseudonymisation refers to the process of substituting original identifiers with fictitious identifiers, commonly known as pseudonyms. While these measures aim to enhance privacy, their effectiveness can be questioned in the context of rapidly evolving Ai technologies.[114] As Boudolf notes, for instance, "researchers recently succeeded in reconstructing both pixelized and blurred faces by making use of neural networks."[115] Ai-powered neural networks can be employed to reverse-engineer anonymised data and potentially re-identify individuals. The term "state-of-the-art" is inherently fluid, as what is considered cutting-edge today may quickly become outdated. This brings into question the long-term viability of these measures in safeguarding sensitive data against emerging threats that arise with more sophisticated techniques. Pseudonymisation, while a valuable privacy-enhancing technique, does not completely anonymise data. As Ai-powered de-anonymisation techniques become more sophisticated, the risk of re-identifying individuals from pseudonymised data increases in direct proportion.

Moreover, the nature of Ai models, especially ML models that continuously learn and adapt, raise questions about the feasibility of complete data deletion. As Chourasia and Shah note, "records in a database become interdependent"[116] and the deleted data's influence remains subliminally in the remaining data due to this

[112] Mühlhoff and Ruschemeier, 'Updating Purpose Limitation for AI: A Normative Approach from Law and Philosophy' (2024) *IJLIT* (1)

[113] Paal, 'Artificial Intelligence as a Challenge for Data Protection Law: And Vice Versa' in Mueller and others (eds), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (Cambridge University Press 2022); Paterson and Maeve (2018) *MULR* 1 (44); Artzt and Dung, 'Artificial Intelligence and Data Protection: How to Reconcile Both Areas from the European Law Perspective' (2022) *VJLS*.

[114] Varanda and others, 'Log Pseudonymization: Privacy Maintenance in Practice' (2021) 63 *JISA* 103021.

[115] Boudolf, Imagery Pseudonymization: Using Ai For Privacy Enhancement (2020) *UniGent*.

[116] Chourasia and Shah, 'Forget Unlearning: Towards True Data-Deletion in Machine Learning', (2023)*PMLR* <https://proceedings.mlr.press/v202/chourasia23a.html> accessed 30 June 2024; See also Izzo and others, 'Approximate Data Deletion from Machine Learning Models' (2021) (PMLR) <https://proceedings.mlr.press/v130/izzo21a.html> accessed 30 June 2024.

interdependence. Notably, the Act does not specify the duration of the retention period, leaving it open to interpretation by Ai providers. This can lead to situations where data is retained for longer than necessary, increasing the risk of unauthorised access or misuse. In the circumstances, a more stringent approach is warranted.

Furthermore, these safeguards primarily focus on technical measures, potentially overlooking the human element in data breaches. The effectiveness of "strict access controls and documentation" hinges on the consistent and rigorous implementation of these protocols by all authorised personnel. However, human error, negligence, or even malicious intent can undermine these safeguards. The Act does not address the potential for insider threats or the need for ongoing training and awareness programs on the handling of sensitive personal data and about evolving security risks in ML models. I propose that these safeguards are insufficient to effectively respond to the foreseeable risks of Ai algorithmic processing of special categories of personal data; more specific, detailed and evolving safeguards are required.

D. Movement 4: Transparency and Explainability

While technical bias mitigation measures, such as those in Article 10, focus on prevention at the development stage, transparency and explainability provide essential mechanisms for scrutiny and challenge *after* an Ai system is deployed. Thus, transparency is relevant for algorithmic discrimination when it comes to the enforcement of obligations to eliminate bias and discrimination in Ai systems. Transparency requirements under the Ai Act, particularly for high-risk Ai systems, serve as a foundational support for non-discrimination by addressing the "black box" problem associated with complex algorithms. It is a necessary first step for individuals and oversight bodies to even *identify* that a potentially discriminatory outcome might be linked to an Ai system. Without knowing that an Ai system was involved, or having basic information about its function, it would be exceedingly difficult to even begin investigating a suspected case of algorithmic discrimination. Transparency, therefore, lifts a veil, enabling the possibility of legal contestability of algorithmic outcomes, including those that are discriminatory.

The right to an explanation, then, support individuals' ability to receive meaningful explanations for decisions made by high-risk Ai systems that affect them, often indirectly facilitating the exercise of remedies under existing non-discrimination law. The ability to understand the key factors or reasons behind an adverse decision is crucial for an individual to determine if unlawful discrimination may have occurred and to effectively challenge that decision through appropriate legal channels. Explainability reduces the opacity of algorithmic decision-making, making it possible to scrutinise the basis for differential treatment. In the absence of such transparency, it becomes impossible to evaluate or adjudicate the outcomes of an Ai system, thereby hindering the identification of discriminatory patterns against any specific group or individual. The Act acknowledges this necessity, recognising the importance of transparency for both persons affected by Ai decisions and for judicial and oversight bodies responsible for adjudicating algorithmic discrimination claims. Nevertheless, in attempting to balance transparency with the protection of intellectual property and confidentiality, the Act fails to mandate a sufficient level of transparency in Ai systems to achieve its objective of upholding non-discrimination.[117]

1. Right to an Explanation

Enforcement has consistently been a significant challenge in the context of non-discrimination law, especially in jurisdictions that primarily rely on individual litigation for the enforcement of the right.[118] Even under ordinary circumstances, persons affected by a discriminatory decision or outcome must contest with considerable difficulties in identifying, proving, and preparing a competent complaint of discrimination to a court.[119] Enforcing the non-discrimination obligation in the context of Ai systems escalates these challenges to exponential proportions. This is because affected persons often cannot detect instances of potential discrimination due to the opacity of these systems. And even where an individual suspects that unlawful discrimination may have occurred, limited access to the inner workings of the models or training data severely restricts their ability to meet the burden of proof requirements mandated by procedural law.

In an effort to address this challenge, Article 86 of the Act creates a new right to an explanation in the context of decisions taken on the basis of outputs of high-risk Ai systems. The right to an explanation for individual decision making is made necessary by the opacity and complexity of Ai systems. One of the major technical and legal challenges at the heart of Ai discourse is the inscrutable, opaque nature in which the algorithm functions. This is referred to as the "black box" problem, which obscures the intricate workings and decisional logic of complex Ai systems, making them

[117] Arnold, 'How the European Union's AI Act Provides Insufficient Protection Against Police Discrimination' (2024) *UPCLS* 1(12).

[118] Deck *et al.*, (2024) arxiv.org; See for example Laina, 'Proving an Employer's Intent: Disparate Treatment Discrimination and the Stray Remarks Doctrine after *Reeves v Sanderson Plumbing Products*' (2002) *VLR* 55 (219)*;* Fredman, *Discrimination Law*, 2nd edn. (Oxford, 2011) 45.

[119] Ponce, 'Direct and Indirect Discrimination Applied to Algorithmic Systems: Reflections to Brazil' (2022) *Computer Law and Security Review.*

"impenetrable,"[120] even as they make increasingly consequential decisions in society. Deep Learning algorithms take in millions of varied data points and correlate distinct data features to produce an output. Because this is a primarily self-directed process, it presents substantial interpretive difficulties for data scientists, programmers, and end-users.[121] As Carstens and others put it, it leads to "decreased comprehensibility of the decisions and outcomes, their underlying criteria and reasons leading to certain decisions and of the weighting between those criteria. This results in a diminished ability of those affected to detect, prove and contest adverse outcomes such as manipulations or discriminations."[122]

Moreover, algorithmic opacity is further occasioned by intellectual property protections coupled with the Act's protective approach to confidential business information. If companies can shield their Ai models and algorithms behind confidentiality and intellectual property claims, it will become difficult for affected persons, regulators and the public to oversee or scrutinise these systems as required variously under the Act. As such, at Article 86(1) the Act provides that "any affected person" who has been subject to a "decision which is taken by the deployer on the basis of the output from a high-risk Ai system listed in Annex III" has "the right to obtain from the deployer clear and meaningful explanations of the role of the Ai system in the decision-making procedure and the main elements of the decision taken."[123]

The right to an explanation is a welcomed step forward in improving transparency and accountability for high-risk Ai systems and facilitating the protection of fundamental rights. The Ai Act is more explicit than the GDPR in articulating the right to an explanation. The GDPR provides instead for a right to be informed about

[120] Artzt and Dung, (2022) *VJLS*.

[121] Pavlidis, 'Unlocking the Black Box: Analysing the EU Artificial Intelligence Act's Framework for Explainability in AI' (2024) *LIT* (3).

[122] Orwat, 'Risks of Discrimination through the Use of Algorithms: A study compiled with a grant from the Federal Anti-Discrimination Agency (*Germany Federal Anti-Discrimination Agency*, 2019) https://www.antidiskriminierungsstelle.de/EN/homepage/_documents/download_diskr_risiken_verwendung_von_algorithmen.pdf?__blob=publicationFile&v=1 accessed 21 June 2025.

[123] The list of high-risk Ai systems at Annex III is sufficiently broad to cover contexts where Ai decision systems might result in discrimination; it includes are those in administration of justice and democratic processes, migration, asylum and border control management; law enforcement; access to and enjoyment of essential private services and essential public services and benefits; employment, workers management and access to self-employment; education and vocational training; and Biometrics - to the extent that their uses are permitted under relevant Union or national law.

the logic involved in automated decision-making.[124] Specifically, Article 13(2)(f) requires controllers to provide "meaningful information about the logic involved" in automated decisions that have legal or significant effects on individuals. The latter right to "meaningful information about the logic involved" begs the question; what does it mean, and what is required to explain an algorithm's decision?[125] In the present context, the right plays a central role to accountability by empowering individuals to challenge adverse decisions made by Ai systems and seek legal remedies/redress where necessary. In this way, the right to an explanation acts as a prerequisite for effective judicial protection, because in theory it enables individuals to understand the basis of an Ai-based decision and potentially challenge it in a competent forum. However, some level of explainability in the Ai system is required for furnishing an explanation; this technical challenge is presently a "roadblock" obstructing meaningful or effective explanations.[126]

The Act is silent on the level of detail required for explanations under Article 86(1). It does not expect deployers to provide complex, detailed explanation on the functioning of the algorithm. This is in part because of a recognition that providing details in an explanation may prove challenging where the algorithm is opaque, because – as Tran Viet Dung puts it – the "logic used may not be easy to describe and might not even be understandable in the first place."[127] And the higher the level of autonomy, the more challenging it becomes to describe the decisional logic or processing activity.[128] The Ai Act mandates only that deployers provide "clear and meaningful explanations of the role of the Ai system in the decision-making procedure and the main elements of the decision taken." It leaves open the exact form or content of these explanations, leaving significant room for interpretation and potential ambiguity.[129] This in turn will risk an inconsistent application of the right across different jurisdictions, Ai systems or contexts.

[124] This right is established in Articles 13-15 and 22 of the GDPR.

[125] Goodman and Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'' (2024) *AiMag* (12).

[126] Artzt and Dung, (2022) *VJLS*.

[127] Ibid.

[128] Ibid.

[129] Kaur, 'Concerns Remain Even as the EU Reaches a Landmark Deal to Govern AI' (ProQuest, 2024)<https://www.proquest.com/docview/2900586403?parentSessionId=blMDkbwsWWQ%2BD7Gb3Oe5k1SQ5d9XcV3czoWfJyYVLSU%3D&pq-origsite=primo&accountid=14682&sourcetype=Trade%20Journals> accessed 29 June 2024.

Further, what constitutes a "clear and meaningful" explanation must ordinarily vary depending on the complexity of the Ai system, the nature of the decision, and the affected person's level of technical literacy and understanding. Similarly, there is a need for more guidance on the "main elements" of a decision. Consider for example the more specific, detailed language used at Article 13(2)(f) of the GDPR. It requires that data subjects be provided with "meaningful information about the logic involved" in automated decision-making, including "the significance and the envisaged consequences of such processing for the data subject." This GDPR provision offers more specific guidance on the content of explanations relative to the Ai Act, unfortunately. Although the Commission is scheduled to come out with more guidance on the implementation of the Act, it is unclear whether this will include details on the standard of explanations under the Act.[130]

Although writing primarily in the context of Article 13 of the GDPR, the discourse on explainability has long been proposing specific and detailed facts that may form part of a meaningful explanation to give effect to the objective behind the obligation.[131] These include the confidence level or uncertainty associated with the Ai system's output, the human oversight measures in place and how they influenced the final decision, and the potential or known risks or limitations of the Ai system that could have affected the decision. There is clearly a need for more concrete regulatory guidance on what constitutes a "clear and meaningful" explanation and the "main elements" of a decision under the Ai Act.

Lastly, Article 86 is limited in its scope of application to Ai systems that have been classified as "high risk." All other decision making or decision support systems which are not classified as high risk are not covered by this provision, leaving open a wide range of decision support systems. These systems which do not meet the threshold of "high risk" may nonetheless produce adverse legal effects or have similarly significant impacts on individuals' health, safety, or fundamental rights.

[130] See Article 73, 6 and 96 of the Act.

[131] Fink and Finck, 'Reasoned A(I)dministration: explanation requirements in EU law and the automation of public administration' (2022) *ELR* 47(376-392) <https://hdl.handle.net/1887/3439725> accessed 29 June 2025; Panigutti and others, 'The Role of Explainable AI in the Context of the AI Act' (2023) *CFAT*; Pavlidis, 'Unlocking the Black Box: Analysing the EU Artificial Intelligence Act's Framework for Explainability in AI' (2024) *LIT*.

2. Judicial Powers to Request and Access Any Documentation Under the Act

As made clear by Recital 170, the Act retains existing mechanisms for redress under Union and national law in the event of algorithmic discrimination.[132] To that end, Recital 170 of the Preamble to the Act clarifies that "Union and national law already provide effective remedies to natural and legal persons whose rights and freedoms are adversely affected by the use of AI systems," thereby further clarifying location of the Act, in relation to existing non-discrimination legal frameworks. Accordingly, the handling of algorithmic discrimination cases in the EU still follows the traditional route for non-discrimination cases: In the first instance, a complainant or person affected by a discriminatory decision has the option to pursue an internal complaint with the respondent or respondent organisation.[133]

Failing this, affected persons may lodge formal legal proceedings with the relevant national equality body or a court with jurisdiction over discrimination claims.[134] To that end, Article 77 provides "national public authorities or bodies which supervise or enforce the respect of obligations under Union law protecting fundamental rights, including the right to non-discrimination" with "the power to request and access any documentation created or maintained under this Regulation in accessible language and format when access to that documentation is necessary for effectively fulfilling their mandates within the limits of their jurisdiction." Where such documentation proves insufficient to ascertain whether an infringement of obligations under Union law protecting fundamental rights has occurred, the public authority or body may make a reasoned request to the market surveillance authority to organise testing of the high-risk Ai system through technical means.[135]

By making clear, as in Recital 170, that Union and national law already provide effective remedies for persons whose rights are adversely affected by Ai systems, and by clarifying in Recital 9 that the Act does not affect existing rights and remedies, the Act explicitly positions itself as not replacing or undermining the established legal

[132] See Recital 170 of the Preamble read with Article 77 of the Ai Act.

[133] In addition to this, individuals can also lodge complaints with national supervisory authorities, designated in each member State to oversee the Act's implementation in terms of Article 85 and Recital 170 of the EU Ai Act. These authorities are empowered to investigate, audit, and enforce corrective administrative measures; See Article 85, 79, 74, and 70 of the EU Ai Act.

[134] Recital 9 of the Preamble to the Act clarifies that the Act does not affect existing rights and remedies under Union law. This confirms that affected persons are still required to rely on existing legal frameworks and institutions, including judicial remedies, to address discrimination caused by Ai systems.

[135] See Article 77(3) of the EU Ai Act.

avenues for challenging discrimination. This confirms that the primary legal framework for defining, prosecuting, and adjudicating discrimination, including algorithmic discrimination, remains with the existing body of Union and national non-discrimination law and the competent courts and equality bodies responsible for their enforcement. The traditional route for discrimination cases, involving internal complaints and formal legal proceedings with national authorities or courts, remains the principal path for seeking redress. The Act's supporting role is particularly evident in how it enhances the effectiveness of these existing mechanisms, specifically through the provision in Article 77.

This is a welcomed measure. However, a significant challenge remains unaddressed in this context: the complainant bears the initial burden of proof to establish a *prima facie* case of discrimination.[136] This means that to succeed with an algorithmic discrimination claim, an individual or affected person must establish firstly their membership in a protected class (e.g., race, gender, disability, etc.) and, secondly, that they were subjected to treatment that is less favourable than someone in a comparable situation who does not share the same protected characteristic. Finally, and more challenging, the complaint must prove that there exists a causal connection between the unfavourable treatment and their protected characteristic.[137] The *prima facie* case must establish on a balance of probabilities that the discrimination or unfavourable treatment is *because of* a prohibited ground or protected characteristic.

While the exact elements of a *prima facie* case may vary across different jurisdictions, the requirement is generally that the complainant or person affected by a discriminatory decision must present evidence that establishes a reasonable inference of unlawful, unjustified discrimination before a court will even admit the complaint or consider its merits.[138] The initial burden of establishing a *prima facie* case can be significant in traditional cases. However, in the context of algorithmic discrimination, this burden escalates, presenting new challenges that stand as significant barriers to establishing a *prima facie* case in the context of algorithmic discrimination. This

[136] Ross, 'The Burden of Proving Discrimination' (2000) *IJDL* 4(2).

[137] Borgesius, 'Discrimination, Artificial Intelligence, and Algorithmic Decision-Making' (*Strasbourg: Council of Europe*, 2018) <https://dare.uva.nl/search?identifier=7bdabff5-c1d9-484f-81f2-e469e03e2360> accessed 30 June 2024.

[138] Ibid.

burden threatens to render remedies illusory for individuals who are not able to surmount these challenges.[139]

First is the increasing opacity and complexity of Ai Systems, particularly those using ML techniques to make or support decisions. The Act's transparency measures seem to be premised on the optimistic assumption that the visible, transparent information will be both comprehensible and truthful. As noted earlier, complex decision systems often operate as "black boxes," making it difficult and often impossible to understand the decision logic leading to a discriminatory outcome, even for experts that have created the model. It obscures or conceals the reasoning behind their decisions and the specific criteria they use or the specific factors that contribute to an outcome, or their weighing of different factors. To the point of this contribution, this opacity also makes it difficult for affected persons to identify and challenge unfair or discriminatory outcomes. Effectively, algorithmic opacity obscures the causal link between the algorithm and the discriminatory effect, making it difficult if not impossible for the complainant to pinpoint the exact source of discrimination and provide sufficient evidence of this to meet the standard of a *prima facie* case.

Explainable Ai (XAi) is a specialised branch within the technical field of artificial intelligence. It focuses on creating Ai systems whose actions and decisions can be easily understood and interpreted by humans, and systems that can otherwise help make other systems more explainable.[140] This field has been developing numerous technical bias definitions and fairness metrics, as well as practical techniques for bias detection and mitigation, with no lasting or scalable success.[141] As Deck *et al.* explain, formalisation and quantification cannot resolve fundamentally "normative issues - rooted in value conflicts."[142] While these challenges can be supported by formal technical methods, it cannot entirely address the challenge.[143]

[139] Lind, 'The Prima Facie Case of Age Discrimination in Reduction-in-Force Cases' (1995) *MLR* 94 (832); Fedorchuk, 'Prooving in Cases of Discrimination in The Field of Labour' (2020) *BTS NUK LS* 59 (1).

[140] Panigutti *et al.* 'The role of explainable AI in the context of the AI Act' (2023) FAccT '23: the 2023 ACM Conference on Fairness, Accountability, and Transparency

[141] Friedler, Scheidegger and Venkatasubramanian, 'The (Im)Possibility of Fairness' (2021) *Communications of the ACM* 136; Creel and Hellman, 'The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems' (2022) *Canadian Journal of Philosophy* 1.

[142] Deck and others, 'Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness' (2024) *arXiv.org*

[143] Ibid.

Secondly, proving algorithmic discrimination will rely on specialised types of evidence, including statistical or mathematical evidence and analysis demonstrating that the algorithmic outcomes disproportionately affect a particular group based on protected grounds such as race, gender, or age. To establish this, the complainant will likely need access to confidential information such as the algorithm's source code, decision-making processes, data inputs, expert evidence, and internal company documents like logs or other technical documentation on the Ai system's operations. This is considered proprietary information and may not be readily accessible to those alleging discrimination. While the Act does provide for an individual's right to an explanation and for judicial access to documentation, it is unlikely that the level of detail and depth of any disclosure will include proprietary or confidential information. In fact, the act employs a necessarily protective approach to confidentiality; I've written elsewhere about the interaction between the right to an explanation and the Act's sweeping and totalising confidentiality obligations.[144]

Third, the collecting and analysing of this evidentiary data will be resource-intensive and will require specialised knowledge in statistics and data science – both on the part of the applicant, and on the part of the judicial officer who must make a decision. For the complainant, this means gathering and analysing large datasets, requiring the support of data scientists or statisticians. This process will be resource intensive and time-consuming, potentially resulting in a de facto exclusion of individuals who lack the means to undertake such a complex endeavour. The financial burden of legal representation, already a significant hurdle for many individuals, is likely to escalate exponentially in the context of algorithmic discrimination claims; regrettably, the Act is unresponsive to this reality. Furthermore, the judicial officer tasked with deciding the case must also possess the technical knowledge to interpret the statistical evidence.[145] The specialised and resource-intensive nature of this process in the context of Ai discrimination creates a significant disadvantage for individuals affected by Ai discrimination, potentially rendering the right to non-discrimination illusory for many.

[144] Kgomosotho, A policy analysis of Confidentiality obligation under Article 78 of the EU Ai Act (TECHila Law, 2025)< https://techilalaw.com/2025/06/18/elementor-1734/> Accessed 21 October 2025.

[145] Article 26(2) of the Act mandates that deployers of high-risk Ai systems assign human oversight to individuals with the 'necessary competence, training, and authority.' While not explicitly mentioning judges, this provision implies that those overseeing AI systems in the judicial context should have adequate training.

The effectiveness of all legal mechanisms for successfully launching discrimination claims centre on the complainant's ability to provide sufficient evidence to establish a *prima facie* case of discrimination based on prohibited grounds. In the context of algorithms, this burden escalates exponentially, presenting significant limits to the accessibility and effectiveness of existing legal redress mechanisms, potentially undermining the very right to non-discrimination.

IV. Is the EU Ai Act a Human Rights Document?

There is a critical perspective which highlights that the Act is not primarily a human rights document, but rather a market regulation focused on product safety and compliance bureaucracy for Ai developers.[146] Criticisms suggest it may lead to superficial compliance rather than ensuring real accountability for fundamental rights risks,[147] being seen as a regulation for companies, not people, partly due to minimal direct user obligations concerning those affected by Ai systems.[148] Reports from NGOs criticise the Act for not being a robust human rights document, citing a lack of robust redress mechanisms and exemptions, illustrating that it is not fundamentally designed as a human rights document.[149] They argue that the focus on promoting Ai uptake potentially comes at the expense of safeguarding rights.[150] This critique is astute - the Act is not fundamentally designed to be a human rights document. Its product-safety, risk-focused, procedural structure is built to address Ai as a technical and commercial object. It seeks to manage "risk" and "bias" rather than discrimination. Because justice cannot be derived from computation, the Act focuses on what is governable, the technical and procedural duties of developers. Therefore, the perceived conflict here is not an oversight, but a feature of a legal framework that defines its success by the creation of a harmonised internal market for trustworthy Ai systems, ensuring safety and managing specific Ai-related risks within a framework that complements, rather than replaces, existing human rights and non-discrimination

[146] Chander, 'EU's AI Law Needs Major Changes to Prevent Discrimination and Mass Surveillance - European Digital Rights (EDRi, 28 April 2021)' <https://edri.org/our-work/eus-ai-law-needs-major-changes-to-prevent-discrimination-and-mass-surveillance/> accessed 29 May 2024; Coalition of Digital, Human Rights and Social Justice Groups, 'EU's AI Act fails to set gold standard for human rights' (Joint Statement/Analysis, 3 April 2024), https://www.amnesty.eu/news/eus-ai-act-fails-to-set-gold-standard-for-human-rights/

[147] Ibid.

[148] Ibid.

[149] Ibid.

[150] Ibid.

law. These are fundamentally different legislative objectives from the pursuit of substantive equality, justice and the active correction of societal inequality.

Seen from the present perspective, the Act doesn't need to be a human rights document. The Act integrates procedural fundamental rights safeguards as essential requirements for market access and deployment, without transforming such compliance into compliance with non-discrimination, or into a standalone rights redress mechanism.

V. Conclusion

The Artificial Intelligence Act is a significant and necessary regulatory response to the risks posed by biased Ai systems. However, as this analysis demonstrates, the Act adopts a predominantly technical approach to bias, positioning itself in a necessary, yet ultimately supporting role relative to established Union non-discrimination frameworks. Consequently, the Ai Act functions as an essential, Ai-specific layer of preventative regulation that supports existing non-discrimination law. By imposing technical obligations like those in Article 10, the Act helps duty-holders reduce the *likelihood* of discriminatory outcomes, thereby facilitating the operationalisation of non-discrimination principles during Ai design and development.

Nonetheless, this supportive function has inherent limits; compliance with Article 10 and other technical mandates cannot guarantee the elimination of discrimination as required by non-discrimination law. An Ai system can technically comply with the Act and still result in discrimination upon deployment, as discrimination can arise from factors beyond initial bias mitigation, including real-world context, dynamic proxies, and how outputs are used. Understanding the Act as primarily a market and safety regulation with integrated safeguards clarifies its supporting role in the governance of non-discrimination. If its foundational purpose is regulating Ai products for the market, its contribution to non-discrimination is naturally facilitative. Its focus on technical compliance aligns with supporting existing non-discrimination legal frameworks by addressing Ai-specific technical vulnerabilities.

VI. Bibliography

A. Primary Sources

Charter of Fundamental Rights of the European Union [2012] OJ C 326/391

Council of Europe, European Convention on Human Rights, as amended by Protocols Nos. 11, 14 and 15, ETS No. 005, 4 November 1950, https://www.refworld.org/legal/agreements/coe/1950/en/18688 accessed 29 September 2025

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 L 180/22

Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation [2000] OJ L 303/16

Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services [2004] OJ L 373/37

D.H. and Others v. the Czech Republic App no 57325/00 (ECtHR [GC], ECHR 2007-IV)

Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) [2006] OJ L 204/23

Guberina v Croatia App no 23682/13 (ECtHR, 22 March 2016)

Proposal for a Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation, COM(2008) 426 final, 2 July 2008 {SEC(2008) 2180} {SEC(2008) 2181}

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts [2024] OJ L 252/1.

B. Secondary Sources

Adams, and others, 'Human rights and the fourth industrial revolution in South Africa' (HSRC Cape Town, 2021)

Arnold, Lukas, 'How the European Union's AI Act Provides Insufficient Protection Against Police Discrimination' (2024) University of Pennsylvania Carey Law School (*UPCLS*)

Artzt, Matthias and Dung, TV, 'Artificial Intelligence and Data Protection: How to Reconcile Both Areas from the European Law Perspective' (2022) Vietnamese Journal of Legal Sciences (*VJLS*) 7 (39)

Barocas, Solon and Selbst, Andrew, 'Big Data's Disparate Impact' (2016) California Law Review (*CLR*) 104 (671)

Borgesius, Frederik Zuiderveen, 'Discrimination, Artificial Intelligence, and Algorithmic Decision-Making' (2018) Strasbourg: Council of Europe

Boudolf, Paul, Imagery Pseudonymization: Using Ai For Privacy Enhancement (Ghent, 2020)

Bower, John, 'The Nature of Data and Their Collection', in John Bower (ed.), Statistical Methods for Food Science: Introductory Procedures for the Food Practitioner, 2nd edn. (Hoboken, 2013) 15

Broussard, Meredith, 'Artificial Unintelligence: How Computers Misunderstand the World' (Cambridge, 2018)

Buolamwini, Joy, 'Press Kit' (MIT Media Lab) https://www.media.mit.edu/projects/gender-shades/press-kit/accessed 20 May 2024

Burk, Dan, 'Algorithmic Legal Metrics' (2020) Notre Dame Law Review (*NDLR*) 96 (1147)

Cao, Longbing, 'AI in Finance: Challenges, Techniques, and Opportunities' (2022) ACM Computing Surveys (*ACM CS*) 64:1

Chander, Anupam, 'EU's AI Law Needs Major Changes to Prevent Discrimination and Mass Surveillance - European Digital Rights (EDRi, 28 April 2021)' https://edri.org/our-work/eus-ai-law-needs-major-changes-to-prevent-discrimination-and-mass-surveillance/ accessed 29 May 2024

Chourasia, Rishav and Shah, Neil, 'Forget Unlearning: Towards True Data-Deletion in Machine Learning', Proceedings of the 40th International Conference on Machine Learning (*PMLR*) (2023) <https://proceedings.mlr.press/v202/chourasia23a.html accessed 30 June 2024

Citron, Danielle Keats and Pasquale, Frank, 'The Scored Society: Due Process for Automated Predictions' (2014) Washington Law Review (*WLR*) 89 (1)

Coalition of Digital, Human Rights and Social Justice Groups, 'EU's AI Act fails to set gold standard for human rights' (Joint Statement/Analysis, 3 April 2024) https://www.amnesty.eu/news/eus-ai-act-fails-to-set-gold-standard-for-human-rights/ accessed 4 May 2025

Coglianese, Cary and Lehr, David, 'Transparency and Algorithmic Governance' (2019) Administrative Law Review (*ALR*) 71 (1)

Cozgarea, Elena, Cozgarea, Gabriel and Stanciu, Andrei, 'Artificial Intelligence Applications In The Financial Sector' (2008) Theoretical and Applied Economics (*TAE*) 12 (57)

Creel, Kathleen and Hellman, Deborah, 'The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems' (2022) Canadian Journal of Philosophy (*CJP*) 1 https://doi.org/10.1017/can.2022.3 accessed 4 May 2025

Crompton, Helen and Burke, Diane, 'Artificial Intelligence in Higher Education: The State of the Field' (2023) International Journal of Educational Technology in Higher Education (*IJETHE*) 20 (22)

Dass, Rahul Kumar and others, 'Detecting Racial Inequalities in Criminal Justice: Towards an Equitable Deep Learning Approach for Generating and Interpreting Racial Categories Using Mugshots' (2023) AI & SOCIETY (*AI&Soc*) 897

Datta, Anupam and others, 'Proxy Non-Discrimination in Data-Driven Systems' (2017) arXiv.org

Deck, Lukas and others, 'Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness' (2024) arXiv.org

Delbosc, Alexa, 'There Is No Such Thing as Unbiased Research – Is There Anything We Can Do about That?' (2023) Transport Reviews (*TR*) 33 (155)

Demircan, Kalyna, 'Europe: The EU AI Act's Relationship with Data Protection Law: Key Takeaways' (Privacy Matters, April 2024) https://privacymatters.dlapiper.com/2024/04/europe-the-eu-ai-acts-relationship-with-data-protection-law-key-takeaways/ accessed 1 July 2024

du Preez, Derek, 'AI and Ethics - "Unbiased Data Is an Oxymoron' (31 October 2019) https://diginomica.com/ai-and-ethics-unbiased-data-oxymoron accessed 19 October 2023

du Preez, Derek, 'AI Is Currently Too Expensive to Take Most of Our Jobs, Finds MIT Researchers' (24 January 2024) https://diginomica.com/ai-currently-too-expensive-take-most-our-jobs-finds-mit-researchers accessed 4 May 2025

Eaglin, Jessica, 'Constructing Recidivism Risk' (2017) Emory Law Journal (*ELJ*) 67 (59)

Fedorchuk, Andrii, 'Prooving in Cases of Discrimination in The Field of Labour' (2020) Bulletin of Taras Shevchenko National University of Kyiv. Legal Studies (*BSNUK LS*) 59

Feiler, Lukas and others, 'EU AI Act: Diversity and Inclusion Prevails over Data Protection' (Lexology, 26 June 2024) https://www.lexology.com/library/detail.aspx?g=a978bb5a-409a-4b26-8df1-26e3244bd97f accessed 29 June 2024

Fink, Melanie and Finck, Michèle, 'Reasoned A(I)dministration: explanation requirements in EU law and the automation of public administration' (2022) European Law Review (*ELR*) 47 (376)

Fredman, Sandra, Discrimination Law, 2nd edn. (Oxford, 2011) 139

Friedler, Sorelle A, Scheidegger, Carlos and Venkatasubramanian, Suresh, 'The (Im)Possibility of Fairness' (2021) Communications of the ACM 136 https://doi.org/10.1145/3433949 accessed 4 May 2025

Geiger, R Stuart and others, 'Garbage in, Garbage out' Revisited: What Do Machine Learning Application Papers Report about Human-Labelled Training Data?' (2021) Quantitative Science Studies (*QSS*) 2 (795)

Gerards, Janneke and Borgesius, Frederik Zuiderveen, 'Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence' (2021) Colorado Technology Law Journal (CTLJ) 55

Gillis, Talia, 'The Input Fallacy' (2022) Minnesota Law Review (*MLR*) 106 (1175)

Goodman, Bryce and Flaxman, Seth, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"' (2017) AI Magazine (*AiMag*) 38

IBM, 'AI Fairness 360' (IBM Research) https://aif360.mybluemix.net/ accessed 4 May 2025

'Italian DPA Fines Food Delivery App 2.6M Euros for GDPR Violations' https://iapp.org/news/b/italian-dpa-fines-food-delivery-app-3m-euros-for-gdpr-violations accessed 1 July 2024

Izzo, Zachary and others, 'Approximate Data Deletion from Machine Learning Models', Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (*PMLR*) https://proceedings.mlr.press/v130/izzo21a.html accessed 30 June 2024

James, Stefanie and others, 'Synthetic Data Use: Exploring Use Cases to Optimise Data Utility' (2021) Discover Artificial Intelligence (*DAI*) 1 (15)

Joh, Elizabeth, 'The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing' (2016) Harvard Law & Policy Review (*HLPR*) 10 (15)

Johnson, Gabbrielle, 'Algorithmic Bias: On the Implicit Biases of Social Technology' (2021) Synthese 198 (10)

Kaur, Gagandeep, 'Concerns Remain Even as the EU Reaches a Landmark Deal to Govern AI' (2023) ProQuest https://www.proquest.com/docview/2900586403 accessed 29 June 2024

Kemp, Charles and Tenenbaum, Joshua, 'Structured Statistical Models of Inductive Reasoning' (2009) Psychological Review (*PR*) 116 (20)

Kgomosotho, Keketso, A policy analysis of Confidentiality obligation under Article 78 of the EU Ai Act (TECHila Law, 2025)< https://techilalaw.com/2025/06/18/elementor-1734/> Accessed 21 October 2025.

Köchling, Alina and Wehner, Marius Claus, 'Discriminated by an Algorithm: A Systematic Review of Discrimination and Fairness by Algorithmic Decision-Making in the Context of HR Recruitment and HR Development' (2020) Business Research (*BR*) 795

Kosta, Eleni, 'Algorithmic state surveillance: Challenging the notion of agency in human rights' (2022) Regulation & Governance (*R&G*) 16 (212)

Laina, R.R., 'Proving an Employer's Intent: Disparate Treatment Discrimination and the Stray Remarks Doctrine after Reeves v. Sanderson Plumbing Products' (2002) Vanderbilt Law Review (*VLR*) 55 (219)

Lind, Jessica, 'The Prima Facie Case of Age Discrimination in Reduction-in-Force Cases' (1995) Michigan Law Review (MLR) 94 (832)

Lindebaum, Dirk and others, 'Insights From 'The Machine Stops" to Better Understand Rational Assumptions in Algorithmic Decision Making and Its Implications for Organisations' (2020) Academy of Management Review (*AMR*) 45 (247)

Lomas, Natasha, 'ChatGPT Is Violating Europe's Privacy Laws, Italian DPA Tells OpenAI' (TechCrunch, 29 January 2024) https://techcrunch.com/2024/01/29/chatgpt-italy-gdpr-notification/ accessed 4 May 2025

Maatman, Sanneke, 'Unbiased Machine Learning Does Not Exist (LBBOnline' October 2018) https://www.lbbonline.com/news/unbiased-machine-learning-does-not-exist-3 accessed 19 October 2023

Majeed, Abdul and Lee, Sungchang, 'Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey' (2021) (*IEEEA*) 8512

Martens, Martin, 'The European Union AI Act: Premature or Precocious Regulation?' (Bruegel, 23 May 2024) https://www.bruegel.org/analysis/european-union-ai-act-premature-or-precocious-regulation accessed 30 June 2024

Marwala, Tshilidzi, *'Closing the Gap: The Fourth Industrial Revolution in Africa'* (Johannesburg, 2020)

Maslej, Nestor and others, 'The AI Index 2023 Annual Report' (AI Index Steering Committee, 4 April 2023) https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf accessed 4 May 2025

Meding, Burkhard, 'It's complicated. The relationship of algorithmic fairness and non-discrimination regulations in the EU AI Act' (2025) arXiv.org 2501.12962v2

Mehrabi, Ninareh and others, 'A Survey on Bias and Fairness in Machine Learning' (2021) ACM Computing Surveys (*ACM CS*) 54 (115)

Meyer, David, 'The Cost of Training AI Could Soon Become Too Much to Bear' See David Meyer, 'The Cost of Training AI Could Soon Become Too Much to Bear' (Fortune, May 2024)<https://fortune.com/2024/04/04/ai-training-costs-how-much-is-too-much-openai-gpt-anthropic-microsoft/> accessed 10 January 2025;

Microsoft, 'Fairlearn' https://fairlearn.org/ accessed 4 May 2025

Mühlhoff, Rainer and Ruschemeier, Hannah, 'Updating Purpose Limitation for AI: A Normative Approach from Law and Philosophy' (2024) International Journal of Law and Information Technology (*IJLIT*)

Nishant, Rohit and others, 'The Formal Rationality of Artificial Intelligence-based Algorithms' (2024) Journal of Information Technology (*JIT*) 39 (20)

O'Neil, Cathy, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (New York, 2016)

Orwat, Carsten, 'Risks of Discrimination through the Use of Algorithms: A study compiled with a grant from the Federal Anti-Discrimination Agency (2019) Germany Federal Anti-Discrimination Agency

Paal, Boris P, 'Artificial Intelligence as a Challenge for Data Protection Law: And Vice Versa' in Oliver Mueller and others (eds), The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives (Cambridge, 2022)

PAIR, 'What-If Tool' https://pair-code.github.io/what-if-tool/ accessed 4 May 2025

Pan, Yuan and others, 'The Adoption of Artificial Intelligence in Employee Recruitment: The Influence of Contextual Factors' (2022) The International Journal of Human Resource Management (*IJHRM*) 1125

Panigutti, Cecilia and others, 'The Role of Explainable AI in the Context of the AI Act' (2023) Conference on Fairness, Accountability and Transparency

Paterson, Moira and McDonagh, Maeve, 'Data protection in an era of big data: the challenges posed by big personal data' (2019) Monash University Law Review (*MULR*) 44 (1)

Patty, John W and Penn, Elizabeth Maggie, 'Algorithmic Fairness and Statistical Discrimination' (2022) Philosophy Compass (*PE*) e12891

Pavlidis, Georgios, 'Unlocking the Black Box: Analysing the EU Artificial Intelligence Act's Framework for Explainability in AI' (2024) Law, Innovation and Technology (*LI&T*)

Pham, Phuong and Sampson, Daniel, 'The Development of Artificial Intelligence in Education: A Review in Context' (2022) Journal of Computer Assisted Learning (*JCAL*) 38 (1408)

Ponce, Pedro Piedade, 'Direct and Indirect Discrimination Applied to Algorithmic Systems: Reflections to Brazil' (2022) Computer Law and Security Review (*CL&SR*) (forthcoming)

Pouget, Sophie and Zuhdi, Sabrina, 'AI and Product Safety Standards Under the EU AI Act' (2024) Carnegie Endowment (*CE*)

Prince, Anya and Schwarcz, Daniel, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data' (2020) Iowa Law Review (*ILR*) 105 (1257)

Quezada-Tavarez, Dutkiewicz and Krack, 'Voicing Challenges: GDPR and AI Research' (2022) Open Research Europe (*ORE*) 2 (126)

Rodolfa, Kit, Lamba, Hemank and Ghani, Rayid, 'Empirical Observation of Negligible Fairness–Accuracy Trade-Offs in Machine Learning for Public Policy' (2021) Nature Machine Intelligence (*NMI*) 10 (896)

Ross, Malcolm, 'The Burden of Proving Discrimination' (2000) International Journal of Discrimination and the Law (IJDL) 4 (2)

Samuel, Sigal, 'Why It's so Damn Hard to Make AI Fair and Unbiased' (Vox, 19 April 2022) https://www.vox.com/future-perfect/22916602/ai-bias-fairness-tradeoffs-artificial-intelligence accessed 19 October 2023

Seijo-Pardo, Borja and others, 'Biases in Feature Selection with Missing Data' (2019) Neurocomputing (*NeoroCom*) 432 (97)

Selbst, Andrew D and others, 'Fairness and Abstraction in Sociotechnical Systems' (2019) Conference on Fairness, Accountability, and Transparency (*CFAT*) 59

Shelton, Dinah, 'Prohibited Discrimination in International Law' in Aristotle Constantinides and Nikos Zaikos (eds), *The Diversity of International Law: Essays in Honour of Professor Kalliopi K. Koufa* (Martinus Nijhoff Publishers 2009) 261-292

Solove, Daniel J, 'Artificial Intelligence and Privacy' (2024) Florida Law Review, (*FLR*) Volume 1 (77)

Solove, Daniel J, The Digital Person: Technology and Privacy in the Information Age (New York, 2004)

Staab, Robin and others, 'Beyond Memorization: Violating Privacy Via Inference with Large Language Models' (arXiv, 6 May 2024) https://doi.org/10.48550/arXiv.2310.07298 accessed 4 May 2025

Surber, Regina, 'Artificial Intelligence: Autonomous Technology (AT), Lethal Autonomous Weapons Systems (LAWS) and Peace Time Threats' (2018) ICT for Peace https://ict4peace.org/wp-content/uploads/2018/02/2018_RSurber_AI-AT-LAWS-Peace-Time-Threats_final.pdf accessed 4 May 2025

'The French SA Fines Clearview AI EUR 20 Million | European Data Protection Board' https://www.edpb.europa.eu/news/national-news/2022/french-sa-fines-clearview-ai-eur-20-million_en accessed 1 July 2024

van Bekkum, Marvin and Borgesius, Frederik Zuiderveen, 'Using Sensitive Data to Prevent Discrimination by Artificial Intelligence: Does the GDPR Need a New Exception?' (2023) Computer Law & Security Review 48 (105770)

Varanda, Artur and others, 'Log Pseudonymization: Privacy Maintenance in Practice' (2021) Journal of Information Security and Applications (*JISA*) 63 (103021)

Wachter, Sandra, Mittelstadt, Brent and Russell, Chris, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) Computer Law & Security Review (*CL&SR*) 105567 https://doi.org/10.1016/j.clsr.2021.105567 accessed 4 May 2025

Wortham, Rachel, 'Garbage in, toxic data out: a proposal for ethical artificial intelligence sustainability impact statements' (2023) AI and Ethics (*AI&E*) 3 (135)

Xie, Nan, 'An explanation of the relationship between artificial intelligence and human beings from the perspective of consciousness' (2021) Cultures of Science (*CoS*) 4 (124)

Yamin, Muhammad and others, 'Weaponized AI for Cyber Attacks' (2021) Journal of Information Security and Applications (*JISA*) 57

Yucer, Seyma and others, 'Measuring Hidden Bias within Face Recognition via Racial Phenotypes' (2021) IEEE Winter Conference on Applications of Computer Vision, (*WACV*)

Zhang, Kirsty and others, 'The AI Index 2024 Annual Report' (AI Index Steering Committee, May 2024) https://arxiv.org/abs/2405.19522 accessed 4 May 2025