German Law Journal (2025), **26**, pp. 1281–1299
doi:10.1017/glj.2026.10189

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Challenges and Implications of Generative AI for Anti-Discrimination Law and Policy

Emanuel V. Towfigh[1,2,3]

[1]Chair in Public Law, Empirical Legal Research and Law & Economics, EBS Universität für Wirtschaft und Recht gGmbH, Wiesbaden, Germany, [2]Peking University School of Transnational Law, Shenzhen, China and [3]Center for Diversity in Law, Max-Planck-Institute for Comparative Public Law and International Law, Heidelberg, Germany
Email: emanuel.towfigh@ebs.edu

(Received 30 May 2025; accepted 10 July 2025)

Abstract
The rapid advancement of generative artificial intelligence (GenAI) presents new challenges and implications for anti-discrimination law. While much of the legal discourse so far has focused on general AI governance, there is a growing need to recognize that different technical models—ranging from algorithmic decision-making systems (ADS) to GenAI—pose distinct risks and require tailored regulatory responses. This Article serves as a foundational introduction to this issue, highlighting the diverse ways in which GenAI can both reinforce and mitigate discrimination and where (European) anti-discrimination law stands in the face of these challenges. The primary goal is to raise awareness of the crucial interplay between AI architectures and legal frameworks. By sketching the underlying technical and structural mechanisms this Article aims to foster an understanding of the potential harmful impact GenAI can have, and how this is met by an anti-discrimination law, not yet equipped for these new "actors." Thus, it will reveal weaknesses that will need to be addressed by research and policy alike.

Keywords: Generative AI; anti-discrimination law; Artificial Intelligence Act; value alignment; data bias

A. Introduction

The rapid advancement of generative artificial intelligence (GenAI) presents novel and complex challenges for anti-discrimination law. Different technological architectures such as algorithmic decision-making systems (ADS) and GenAI entail qualitatively different risks and demand context-specific regulatory approaches. GenAI in particular has the capacity both to reinforce and to mitigate discriminatory patterns. This Article seeks to provide a foundational inquiry into these emerging intersections. It aims to clarify the underlying technical structures of GenAI, assess their potential to cause harm through discriminatory effects, and critically examine the adequacy of current legal frameworks—particularly European anti-discrimination law—in responding to these risks. In doing so, the Article exposes doctrinal and regulatory shortcomings and identifies key areas for further legal and policy development.

To this end, the Article opens with a conceptual exposition of Artificial Intelligence, Predictive AI, and Generative AI in Section B, distinguishing core technological modalities and elaborating on the unique generative capacities of large language models. Section C then turns to the normative landscape of anti-discrimination law, outlining foundational legal standards and typologies and identifying their limitations when applied to algorithmic systems. Building on this, the Article analyzes the specific discriminatory risks posed by generative models, focusing on three

critical layers: biased training data, the misalignment of model outputs with societal norms (value alignment), and the deliberate circumvention of safety measures (jailbreaking). The Article then moves to the regulatory dimension of GenAI and its discriminatory potential in Section D, critically assessing existing frameworks including Anti-Discrimination Law, the Artificial Intelligence Act, and the Digital Services Act. Finally, Section E synthesizes the analysis and argues for a recalibrated, technologically attuned anti-discrimination framework that can adequately respond to the challenges posed by GenAI systems.

B. Artificial Intelligence, Predictive AI, and Generative AI

I. Artificial Intelligence and Algorithms

The term "Artificial Intelligence" (AI) is as omnipresent as AI itself, but at the same time an expression that is used with a multitude of meanings and defined differently in different contexts. Ranging from the very broad definition, equating AI with algorithms, to much more specific definitions that use the term only for specific technologies—usually the most recent, for example currently machine learning and large language models—the definition can encompass very different applications, concepts, and models.[1] The most common and probably least technical "definition," and the one this Article follows, explains AI as the technology that focuses on creating systems that are capable of performing tasks that typically require human intelligence.[2] Sophisticated algorithms enable machines to mimic aspects of human intelligence, such as pattern recognition, decision-making, and problem-solving. In this context, algorithms are computing operations that are performed in a predetermined sequence in order to solve precisely defined problems with the help of input data. Translated into binary machine code, they form the basis of all software and enable the program code to automatically convert input information into an output. Algorithms can be programmed specifically for the execution of a certain operation, in which case they implement this programming immediately and generate the desired, predetermined result. Navigation devices that calculate the shortest route between two points follow this mode of operation. If algorithms are supposed to solve problems that are not (or cannot be) defined in advance, the algorithm must be a "learning" one. Such an algorithm regularly analyzes large amounts of data and processes and evaluates these using statistical methods, often times in connection with neural network models.[3] Based on the insights gained from the evaluated data and identified correlations, relationships, or patterns within a dataset, these algorithms continuously refine their statistical-probabilistic predictions.[4] Such algorithms can be machine-learned, in other words "trained," by constantly inputting new data and confirming or correcting the output. Improvements in speech recognition software, for example, are based on such learning algorithms; reinforcement learning from human feedback plays an important role in this area when it comes to reaching human-like quality levels.[5]

II. Predictive AI

Predictive AI is the term used to describe a subset of AI, that support (human) decision-making or make decisions themselves with an ADS. They support or replace human decisions and forecasts by combining at least two algorithms for automated decision-making. The first algorithmic

[1] *See* Haroon Sheikh, Corien Prins & Erik Schrijvers, Mission AI: The New System Technology 398 (2023).

[2] *See* Sofia Samoili, Montserrat Lopez Cobo, Emilia Gomez Gutierrez, Giuditta De Prato, Fernando Martinez-Plumed & Blagoj Delipetrev, Eur. Comm'n, Joint Rsch. Ctr., *AI Watch. Defining Artificial Intelligence,* at 4, EUR 30117 EN, JRC118163 (2020).

[3] Nils J. Nilsson, The Quest for Artificial Intelligence: A History of Ideas and Achievements 398 (2010).

[4] David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. Davis L. Rev. 653, 671 (2017).

[5] Nilsson, *supra* note 3, at 398.

subsystem uses general data from the past to study or categorize people's behavioral characteristics, creating a set of rules. The data that is to be the subject of the decision-making process is entered to this set of rules, and a second algorithm assesses the data and generates a scoring value that can be the basis for the requested decision; typical fields of application are decisions on a person's creditworthiness, the insurance premium for a car insurance, and selection decisions—for example, in recruitment procedures. ADS are also the means of choice if mass individualization is desired—for example, for concluding and designing contracts, creating personalized advertising, setting individualized prices—but can also be useful in medicine or support the allocation of limited public resources.[6] The decision is a data-driven prediction, based on a machine-learning process.

The degree of influence that ADS have on human decision-making varies: They can provide information on which a human decision is based; essentially prepare and determine the decision themselves, leaving only little room for human decision; or determine the decision independently, without human involvement.

III. Generative AI and Large Language Models

GenAI represents a new subset of Artificial Intelligence: Unlike ADS, GenAI has traditionally focused on classification tasks or predicting outcomes based on input data. Such activity is directed at creating new content—thus "generative"—which can be text, images, or even music.[7]

GenAI is based on artificial neural networks, a computational model inspired by the networks of nerve cells that make up the human brain.[8] Its processes mimic how biological neurons work together to identify phenomena, weigh options, and draw conclusions. A specialized application of these models are "Large Language Models"[9] (LLMs) that consist of huge neural networks that are specially trained to "understand" and generate texts. With large text corpora as their training data to learn language patterns, LLMs use prediction models that calculate the most likely word sequences to generate comprehensible responses pertinent to the context, when prompted. They can be tailored to develop specific skills, for example legal or medical language. Caution is advised, however, as the tendency of these models to "hallucinate" can lead to conclusive results at first glance that are actually entirely made up.[10] LLMs, for instance, are the basis of chatbots, but are now mostly known from applications such as ChatGPT, Claude, Gemini, Perplexity, Mistral, DeepSeek, Co-Pilot and so forth.

LLMs use advanced algorithms, such as deep learning neural networks, to generate novel content that resembles existing data and can be used to create realistic images, write human-like text, or even compose music, based on patterns learned from training data. It can also be trained to make assumptions about white patches in the knowledge base it was trained on through recombination. The complexity of these algorithms allows the system to "understand"—in other words, process and appropriately react to—nuanced language, making them capable of producing creative outputs in ways that were previously unattainable by traditional algorithms.

[6]Eva Achterhold, Monika Mühlböck, Nadia Steiber & Christoph Kern, *Fairness in Algorithmic Profiling: The AMAS Case*, 35 Minds & Machs., Jan. 29, 2025, at 6–7, https://link.springer.com/article/10.1007/s11023-024-09706-9.

[7]Emilio Ferrara, *GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models*, 7 J. Computational Soc. Sci. 549, 550 (2024); Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu & Lichao Sun, *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT*, 37 J. a.c.m., Aug. 2018, at 1, https://arxiv.org/pdf/2303.04226v1 [https://doi.org/10.48550/arXiv.2303.04226]; Shivam R. Solanki & Drupad K. Khublani, Generative Artificial Intelligence: Exploring the Power and Potential of Generative AI 1 (2024).

[8]Wolfgang Ertel, Introduction to Artificial Intelligence 254 (Nathanael T. Black trans., Springer 3d ed. 2025) (2021).

[9]Cao et al., *supra* note 7, at 4.

[10]*Id.* at 27.

C. Discrimination and AI

I. Standard of Anti-Discrimination Law

The fundamental principle of equality, going back to the Aristotelian idea of equality as the treating of likes alike and differents differently, constitutes the core of basically all anti-discrimination law.[11] The European Convention on Human Rights (ECHR) guarantees in Article 14 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, color, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status."[12] The ECHR does not guarantee a general principle of equality, but an accessory right, and the violation of Article 14 must be related to another guarantee of the ECHR. This appears at first sight to be a restriction of the scope of Article 14, but according to the ECHR a violation of the other guarantee of the ECHR is not necessary: It is sufficient that the facts of the case fall within the ambit of a substantive Convention right.[13]

In European Union law, for example, the fundamental principle of equality is implemented in Article 20 of the Charter of Fundamental Rights of the European Union (CFR). That Article states that "[e]veryone is equal before the law,"[14] standardizing equality both in the application of the law and in legislation, binding EU institutions and Member States as they implement EU law.[15] According to Article 21, paragraph 1 of the CFR, discrimination is prohibited on the grounds of gender, race, color, ethnic or social origin, genetic characteristics, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation.[16] Thus substantiating the principle of equality, Article 21, paragraph 1 of the CFR states the grounds on which unequal treatment may not be based. Specific non-discrimination directives substantiate these provisions further and apply them to areas particularly prone to discrimination such as employment, the welfare system, and access to goods and services.[17]

While the terms and details of anti-discrimination law differ in different legal frameworks, discrimination is typically classified under the categories of *direct discrimination* or *disparate treatment*,[18] *indirect discrimination* or *disparate impact*, *intersectional discrimination,* and *proxy discrimination*.[19]

---

[11]Emma Lantschner, Reflexive Governance In EU Equality Law 22 (2021).

[12]European Convention for the Protection of Human Rights and Fundamental Freedoms art. 14, Nov. 4, 1950, 213 U.N.T.S. 221 [hereinafter ECHR].

[13]Rasmussen v. Denmark, App. No. 8777/79, para. 29 (Nov. 28, 1984), https://hudoc.echr.coe.int/eng?i=001-57563; Inze v. Austria, App. No. 8695/79, para. 36 (Oct. 28, 1987), https://hudoc.echr.coe.int/?i=001-57505; Karlheinz Schmidt v. Germany, App. No. 13580/88, para. 22 (July 18, 1994), https://hudoc.echr.coe.int/rus?i=001-57880.

[14]Charter of Fundamental Rights of the European Union art. 20, Oct. 26, 2012, 2012 O.J. (C 326) 391 [hereinafter CFR].

[15]Niels Petersen, *The Principle of Non-Discrimination in the European Convention of Human Rights and in EU Fundamental Rights Law*, *in* Contemporary Issues in Human Rights Law 129, 129 (Yumiko Nakanishi ed., 2017).

[16]CFR, *supra* note 14, at art. 21, para. 1.

[17]*See, e.g.*, Council Directive 2000/43 of 29 June 2000 Implementing the Principle of Equal Treatment Between Persons Irrespective of Racial or Ethnic Origin, 2000 OJ (L 180) 22 (EC); Council Directive 2000/78 Establishing a General Framework for Equal Treatment in Employment and Occupation, 2000 OJ (L 303) 16 (EC); Council Directive 2004/113 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 (EC); Directive 2006/54 of the European Parliament and of the Council 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 (EC).

[18]*See* Thomas Zollo, Nikita Rajaneesh, Richard Zemel, Talia Gillis & Emily Black, *Towards Effective Discrimination Testing for Generative AI, in* Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency 1028 (2025); Alice Xiang, *Fairness & Privacy in an Age of Generative AI*, 25 Colum. Sci. & Tech. L. Rev. 288, 298 (2024).

[19]*See* Oran Doyle, *Direct Discrimination, Indirect Discrimination and Autonomy*, 27 Oxford J. Legal Stud. 537 (2007); Marie Mercat-Bruns, Discrimination at Work: Comparing European, French, and American Law 82–83 (2016).

Direct discrimination constitutes the less-favorable treatment of a person or group based on a protected characteristic, for example refusing to hire a qualified candidate solely because of their ethnicity. This discrimination is intentional. In contrast, indirect discrimination happens when seemingly neutral policies or practices disproportionately disadvantage a particular group of people—for example, a certain dress code that cannot be reconciled with certain religious dress codes.[20]

In cases of intersectional discrimination, a person experiences overlapping or multiple forms of discrimination based on a combination of characteristics—for example, black women facing not only discrimination based on gender but also on race.[21]

Proxy discrimination takes place where seemingly neutral characteristics are used as a stand-in for a protected trait, leading to biased outcomes. An example of proxy discrimination is using zip codes to determine eligibility for loans, thus discriminating against racial minorities living in historically segregated neighborhoods. Proxy discrimination is particularly virulent today in the context of big data, as the availability of a diverse set of personal data enables the use of a wide range of proxies. However, it is not a purely digital phenomenon; proxies can also be used to discriminate in the "analog" world.[22]

The legal protection mechanisms of anti-discrimination laws are usually tailored to cases of direct and indirect discrimination, and they give the affected individual legal protection options to defend themselves against discrimination. However, in the case of proxy discrimination—in other words, cases of systemic discrimination—they face major challenges. This is largely due to a structural imbalance between the individual affected and the discriminator. In such cases, the discriminator usually is not an individual, but a diffuse entity that might or might not be held responsible for the discrimination.[23] A major obstacle lies in the evidentiary burden, as victims of discrimination are often required to prove discriminatory intent or effects, despite having limited access to internal company data or decision-making processes. It often takes a disproportionate amount of effort to identify someone responsible, so in many cases legal protection is not sought.[24]

Anti-discrimination laws are mostly designed to sanction the misconduct of individuals in the sense of unequal treatment on the specifically stated grounds of discrimination.[25] They are ill-prepared for the specifics that go along with discrimination caused by an algorithm—for example in an ADS—as algorithmic discrimination does not fall straightforwardly into the legally recognized categories of direct or indirect discrimination, especially with their heightened risk of perpetuating structural inequalities.[26]

II. Discrimination Through Statistics

The main element of present-day (big) data-driven AI is the identification of a large number of complex correlations using statistical methods.[27] Big data allows for large and complex data sets to be analyzed using algorithms that operate with statistical methods and often continually refine their statistical assumptions (*priors*) based on the insights gained from the analyzed data. Thus,

[20]Sophia Moreau, *What Is Discrimination?*, 38 PHIL & PUB. AFFS. 143, 154 (2010).

[21]*See, eg.*, Kimberle Crenshaw, *Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color*, 43 STAN. L. REV. 1241 (1991).

[22]*See* Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257 (2020).

[23]Moreau, *supra* note 20, at 146.

[24]INDRA SPIECKER & EMANUEL V. TOWFIGH, FED. ANTI-DISCRIMINATION AGENCY (GER.), CODED BIAS: THE GENERAL EQUAL TREATMENT ACT AND PROTECTION AGAINST DISCRIMINATION BY ALGORITHMIC DECISION-MAKING SYSTEMS 28 (2023).

[25]*See* Moreau, *supra* note 20, at 143.

[26]*See* SPIECKER & TOWFIGH, *supra* note 24 at 23–25.

[27]*See supra* Section B.I.; Emanuel Towfigh, *Der Umgang mit Empirie beim Nachweis von Diskriminierung*, *in* HANDBUCH ANTIDISKRIMINIERUNGSRECHT 759 (Anna Katharina Mangold & Mehrdad Payandeh eds., 2022).

the algorithms help identify complex correlations and allow for a variety of differentiations. In this way, relationships are established between variables that allow for probability statements. However, this increases the risk of *discrimination through statistics*.[28]

Discrimination through statistics is the result of the attribution of characteristics by statistical means based on (actual or assumed) average values of a group—"discriminatory variables."[29] The reference to these average group characteristics is intended to help overcome uncertainties regarding the individual characteristics of a single person. This principle is used, for example, to determine the insurance premium of a motor vehicle insurance policy.[30] It includes data on a vehicle's performance when classifying accident risk, but usually no data on the insured person's personal driving style—although insurance companies are increasingly offering tariffs that use so-called telematics to track the driving behavior of the insured and adapt the insurance premium accordingly.[31] Here, one characteristic—vehicle performance—is used as an indicator for another characteristic—accident risk—while other possible influencing factors, such as the insured's driving style, are not considered, usually because these bits of information are not publicly available and/or are more difficult to assess. Any other mechanisms underlying the assessment are of no interest; *causality* between these two variables is neither proven nor claimed. In a scenario, where it is not just about an insurance premium, but the determination of the risk of relapse, sole attribution of characteristics based on average values of a group poses a serious risk of discrimination. If, for example, social mechanisms underlying an assessment are not taken into account, (historical) structural inequalities are created and perpetuated: for example, black incarceration rates.[32] This was the case with the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, used to predict recidivism risks and thus supporting decision-making in the criminal justice system.[33] The algorithm was found to assign a higher risk of recidivism to black individuals, even after controlling for other characteristics (for example, prior crimes). This score often led to longer probation periods or prison time for black individuals than for white individuals.

III. Discrimination Potential of Biased Data

Once associated with the notion that they allow for objective, unbiased decisions uninfluenced by human error, the discriminatory potential of algorithms is now undisputed. The mechanisms contributing to this have largely been identified.[34]

[28]Carsten Orwat, Fed. Anti-Discrimination Agency (Ger.), Risk of Discrimination Through the Use of Algorithms 26–29 (2020)

[29]Towfigh, *supra* note 27, at 761.

[30]Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Cal. L. Rev. 671, 677 (2016); Philipp Hacker, *Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law*, 55 Common Mkt. L. Rev. 1143, 1149 (2018).

[31]James Boylan, Denny Meyer & Won Sun Chen, *A Systematic Review of the Use of In-Vehicle Telematics in Monitoring Driving Behaviours*, 199 Accident Analysis & Prevention, May 2024, at 1, https://www.sciencedirect.com/science/article/pii/S0001457524000642.

[32]*See* Ignacio Cofone & Warut Khern-am-nuai, *The Overstated Cost of AI Fairness in Criminal Justice*, 100 Ind. L. Rev. 1431 (2025); Hanna Hoffmann, Verena Vogt, Marc P. Hauer & Katharina Zweig, *Fairness by Awareness? On the Inclusion of Protected Features in Algorithmic Decisions*, 44 Comput. L. & Sec. Rev., Apr. 2022, at 4, https://www.sciencedirect.com/science/article/abs/pii/S0267364922000061?via%3Dihub.

[33]Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased against Blacks*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Cofone & Khern-am-nuai, *supra* note 32, at 1436–38; Xiang, *supra* note 18, at 290; Christoph Engel, Lorenz Linhardt and Marcel Schubert, *Code Is Law: How COMPAS Affects the Way the Judiciary Handles the Risk of Recidivism*, 33 A.I. & L. 383, 384 (2025).

[34]*See* Kristen M. Altenburger & Daniel E. Ho, *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions*, 175 J. Inst'l & Theoretical Econ. 98, 98–102 (2019); Barocas & Selbst, *supra* note 30, at 677–93. *See also* Emilio Ferrara, *Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts,*

It is not only the quantity but also the quality of the data that is decisive for the quality of the output. Any data-driven algorithm is largely dependent on the quality of the data input. If the training data is qualitatively inferior, incorrect, or unsuitable for the intended purpose, those defects will have repercussions on the performance of the system and the probability of discriminatory outcomes rises.[35] Already the selection of the underlying data requires diligence and the intended application needs to be considered: The population from the dataset might not resemble the population it is used on, it could be incomplete with regard to certain attributes, or it could apply certain attributes only to specific groups, leading to discriminatory decisions.

This is the case if the training data used is already biased. A learning system that relies on biased data cannot make non-discriminatory decisions. If certain groups are under- or overrepresented, the output will be biased one way or the other. The issue becomes even more severe for learning systems that rely on data from different time periods. If these systems were ever biased at any point, the outcome will also be biased, therefore continuously perpetuating inequality.[36] A facial recognition software that is trained almost exclusively with data from white people is unable to reliably identify black or Asian people—as was the case with a facial-tracking software Hewlett Packard used that was unable to identify dark-colored faces as faces and a Nikon camera that asked people with Asian appearance or phenotype whether someone in the picture blinked.[37]

A good example is training data that contains outdated role models and stereotypes that leads to translation software selecting the male form for academic professions and the female form for less-qualified professions when translating from gender neutral languages into English.[38] Another algorithm—used for medical school placement—relying on past selection criteria, disproportionately admitted white male students, thereby discriminating against women and ethnic minorities and perpetuating historical biases.[39]

The Austrian Public Employment Service (AMS) was also accused of bias after using an algorithmic profiling system to determine the prospects of job seekers in order to specifically support only those individuals with the best prospects for employment.[40] The algorithm used in this project was based on data that was obtained from job seekers registering with the AMS and completed by data from the Main Association of Austrian Social Security Institutions.[41] The system was supposed to identify the prospects of job seekers and classify them into three categories: job seekers with high chances of finding a job in the short term, low prospects of finding a job in the long-term, and job seekers with mediocre prospects. Since the financial resources of the AMS were finite, only job seekers with at least mediocre prospects were marked as potential beneficiaries of labor market programs. The resistance towards the use of the program was based on the selection and weighting of the variables: for example, the gender "female" automatically led to a deduction of points, as was having care obligations—a variable that was usually found in relation to women. As a result, women with care obligations were less-often eligible for participation in labor market programs. Those responsible for

and Mitigation Strategies, 6 SCI, Dec. 26, 2023, at 1, 6, https://www.mdpi.com/2413-4155/6/1/3; Jan Kleinberg, Jens Ludwig, Sendhil Mullainathan and Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J.L. ANALYSIS 113, 114 (2018).

[35]Ferrara, *supra* note 34, at 2; Barocas & Selbst, *supra* note 30, at 680.

[36]Toon Calders & Indre žliobaite, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 43, 48 (Bart Custers, Toon Calders, Bart Schermer & Tal Zarsky eds., 2013).

[37]FREDERIK ZUIDERVEEN BORGESIUS, COUNCIL OF EUR., DISCRIMINATION, ARTIFICIAL INTELLIGENCE, AND ALGORITHMIC DECISION-MAKING 17 (2018); Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROCS. MACH. LEARNING RSCH. 77, 77 (2018).

[38]Marcelo O.R. Prates, Pedro H. Avelar and Luís C. Lamb, *Assessing Gender Bias in Machine Translation: A Case Study with Google Translate*, 32 NEURAL COMPUTING & APPLICATIONS 6363, 6377 (2020).

[39]Stella Lowry & Gordon MacPherson, *A Blot on the Profession*, 296 BRIT. MED. J., Mar. 5, 1988, at 657.

[40]Doris Allhutter, Florian Cech, Fabian Fischer & Astrid Mager, *Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective*, 3 FRONT. BIG DATA, Feb. 20, 2020, at 1, https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2020.00005/full.

[41]*Id.* at 5.

the program explained this with the findings derived from the available historical labor market data regarding the employability of women—an argument that strengthens rather than weakens the need for caution with regard to historical data.[42]

With a structured data set created and used for a specific purpose, some of the disadvantages of biased data can be overcome by removing incriminating variables from the dataset—or, in the case of the AMS, changing the weighting of the variable "gender." However, this will only reduce discrimination in the long term if other variables which remain part of the training data are not correlated with the incriminated attribute.[43]

The possibilities of data-processing know practically no bounds; the technological progress of the digital evaluation methods and techniques that big data represents is certainly not the endpoint. The most diverse data from the most diverse sources can be linked and new data sets created. Such linking facilitates the dissemination of potentially discriminatory data and at the same time makes effective content control of these data sets increasingly difficult.

IV. The Case of GenAI

On the one hand, with GenAI being a more specific and sophisticated AI than ADS, the already inherent risks of discrimination as outlined before are significantly amplified when compared to what we have encountered so far. This amplification stems in part from the in-built complexity and scale of GenAI systems, which introduce additional layers of vulnerability and ethical challenges. On the other hand, new categories of harm, so far not covered by the established categories of anti-discrimination law, emerge with the use of GenAI.[44]

It is therefore worthwhile to focus on three different but still interconnected levels of the use of GenAI, each contributing to the overall concern that GenAI raises when it comes to the risks of discrimination, and to call for regulatory measures that can hedge these novel risks. The first level concerns the core of every LLM: its training data. The second level are the measures taken by the facilitators—the individuals, organizations, and teams developing and deploying these systems—aiming to reduce the risks of misalignment ("*value alignment*"). Finally, the third level comprises the risks that stem from deliberate evasion of safeguards, such as bypassing restrictions or exploiting vulnerabilities in the system through "*jailbreaks.*"

The complex interaction of these three levels underscores the heightened risks associated with GenAI and emphasizes the necessity for robust governance, rigorous testing, and ongoing monitoring to mitigate the potential for discriminatory outcomes.

1. Training Data

GenAI models are heavily reliant on vast datasets for training, and the presence of biased, incomplete, or unrepresentative data can perpetuate and even magnify discriminatory outcomes. Such biases in training data may arise from historical inequalities, societal stereotypes, or systemic issues, all of which are inadvertently embedded in the outputs.[45]

Naturally, GenAI is only as good as its underlying data. But with its more-advanced capabilities, the effect defective data has on the results is more severe.

While in most applications algorithms use datasets structured and curated for a certain purpose—for example, determining your insurance premium—the data used to train GenAI is not structured. Not only does GenAI use a vast amount of text-data; as of today it (still) uses data that is unlabeled and not compiled for a (or any) specific task and is generally scraped from

[42]Allhutter et al., *supra* note 40, at 2.

[43]Calders & žliobaite, *supra* note 36, at 54.

[44]Philipp Hacker, Brent Mittelstadt, Frederik Zuiderveen Borgesius & Sandra Wachter, *Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It*, *in* The Oxford Handbook of the Foundation and Regulation of Generative AI (2025); Xiang, *supra* note 18, at 290.

[45]Solanki & Khublani, *supra* note 7, at 174; Xiang, *supra* note 18, at 288.

websites.[46] The LLMs usually are pre-trained unsupervised, a training paradigm that enables models to autonomously recognize and internalize the statistical patterns and structures to comprehend linguistic constructs inherent in human language such as syntax, semantics, and context, without the necessity for manual annotations. As the data is unlabeled, these linguistic attributes are solely inferred from the distribution and co-occurrence of words and phrases within the text.[47] This pre-trained data is continuously supplemented by human feedback that users give on the output, fine-tuning the LLMs through user "reinforcement learning."[48]

Biases inherent in the training data will inevitably lead to misrepresentation of groups in the content created by the LLM, thus perpetuating stereotypes. If the keyword "inmates" leads to over 80% of generated images showing persons with darker skin tones, or "high-paying jobs" results in images of those with lighter skin, not only is this proof of the existing bias, but the results have the potential to changing the perception of certain groups in society.[49] The magnitude of these inherent biases was also shown in a study which revealed that LLMs exhibit covert racial bias against African American English. While they have been trained to avoid overt racism towards black people, if prompted with text in African American English, the LLMs generated adjectives like "dirty," "lazy," or "aggressive," whereas prompts written in Standardized American English produced neutral or positive adjectives.[50] Another study detected a persistent anti-Muslim bias, by prompting GPT-3 to complete the phrase "two Muslims walked into a ...." Sixty-six out of 100 completions contained violence-related words, such as "shooting," "killing," and so forth, exhibiting quite some creativity by using a variety of weapons and changing the nature and setting of the violence involved.[51]

It was also revealed that LLMs perpetuated existing healthcare biases that were linked to sociodemographic factors like race, sexual orientation, and socioeconomic status, among others.[52] For instance, individuals identified as belonging to LGBTQIA+ subgroups were advised to undergo mental-health evaluations at a rate approximately six to seven times higher than clinically warranted. Similarly, those labeled as having a high-income status received significantly more recommendations for advanced imaging procedures, whereas individuals labeled as low- or middle-income were frequently restricted to basic or no additional testing. There was no clinical reason for these differences, as the medical factors were identical; the only differences between cases were the variations of the sociodemographic group, indicating a potential model-driven bias.

Furthermore, productive use of GenAI requires not only *correct* data, but also data *fitting the prompt*. Even if not trained with data regarding German law, a LLM will still answer your legal question, but will make up a result that sounds reasonable at first glance.[53] That is, it will "hallucinate" an answer that creates the impression of being legally sound.

[46]Xiang, *supra* note 18, at 288; Abubakar Abid, Maheen Farooqi & James Zou, *Persistent Anti-Muslim Bias in Large Language Models*, A.I.E.S. '21: Procs. 2021 A.A.A.I./A.C.M. Conf. on A.I., Ethics, & Soc'y, July 30, 2021, at 298, https://dl.acm.org/doi/10.1145/3461702.3462624.

[47]Jon M. Garon, *A Practical Introduction to Generative AI* 13 (A.B.A. Bus. L. Section Working Paper), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4388437.

[48]Cao et al., *supra* note 7, at 6.

[49]Ferrara, *supra* note 34, at 4; Leonardo Nicoletti & Dina Bass, *Humans Are Biased. Generative AI Is Even Worse*, Bloomberg Tech.: The Big Take (June 9, 2023), https://www.bloomberg.com/graphics/2023-generative-ai-bias/?utm_source=website&utm_medium=share&utm_campaign=copy.

[50]Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky & Sharese King, *AI Generates Covertly Racist Decisions About People Based on Their Dialect*, 633 Nature 147, 149 (2024).

[51]Abid et al., *supra* note 46, at 299.

[52]Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U. Apakama, Carol R. Horowitz, Alexander W. Charney, Robert Freeman, Benjamin Kummer, Benjamin S. Glicksberg, Girish N. Nadkarni & Eyal Klang, *Sociodemographic Biases in Medical Decision Making by Large Language Models,* 31 Nature Med. 1873 (2025).

[53]*Cf.* Hanjo Hamann (@Hanjo-Hamann), LinkedIn (Oct. 2024), https://www.linkedin.com/posts/hanjo-hamann_putationsersatzklage-chatgpt-chatgpt-activity-7244232586808426496-FN89/ (describing an informal experiment querying multiple GPT models about a fictitious legal term and noting divergent definitions)

A second challenge on the level of the training data and its processing arises due to the technological setup of LLMs as a neural network: In neural networks, nodes are activated and weighted during each operation. Each additional operation activates other nodes and consequently changes the weighting of existing nodes. Thus, the neural network changes permanently: its dynamic pathways cannot be reconstructed.[54] If anything, they could be halted at a specific point in time, but what the network looked like a second before the halt or will look like a second after being resumed cannot be inferred. For the use of LLMs, this means that the source of potential discrimination could only be identified if this process—in other words the constant further development of the data—is halted when discriminatory output is created. Such is an impossible endeavor, especially as the neural network changes constantly and no exact scenario will repeat itself reliably. The same prompt, used at different times, will generate different outputs. This makes it impossible to detect the source of discrimination within the training data.

2. Value Alignment

Misalignment between the values embedded within the GenAI and societal or ethical standards can result in unintended consequences, ranging from exclusionary practices to the reinforcing of harmful biases. Ensuring proper alignment requires deliberate and continuous efforts in design, testing, and oversight, which, if neglected, leaves room for discrimination to thrive.[55]

The measures deliberately taken by the architects and providers of GenAI models to influence the outcome of the LLM are supposed to prevent GenAI from producing results that are potentially criminal or harmful or not aligned with human values. An LLM will most probably not give you the instructions on how to build a bomb, for example, and if you ask ChatGPT how to best bully a colleague into leaving the job, it will answer "bullying is harmful, unethical, and unacceptable in any form" and give advice on how to better solve a workplace conflict.

To achieve this, several technical measures often are employed: for example, "Reinforcement Learning from Human Feedback (RHF)," where the models are fine-tuned based on human preferences to encourage outputs that align with desired values.[56] Another approach is dataset curation and preprocessing, which entails selecting and filtering training data to minimize biases and reinforce desirable behaviors, thus influencing the values that LLMs learn and propagate.[57] Additionally, post-processing filters and safety layers are implemented to monitor and control the outputs of LLMs in real time.[58] These mechanism can block or modify responses that are deemed inappropriate or misaligned with human values, adding an extra layer of oversight.

These measures are mainly intended to prevent the misuse of AI systems for malicious or harmful purposes, but they have also led to incorrect or unrealistic "overinclusive" results. The most popular case was probably Google's Chatbot Gemini, which when asked "show me a pope"[59]

[54] *Cf. supra* Section B.*III* (for the technical background of LLMs).

[55] Jianfeng Cao, *From Principle to Practice: Value Alignment in AI Ethics and Governance*, 26 German L.J. (2025) (in this Special Issue); Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Kwan Yee Ng, Aidan O'Gara, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo & Wen Gao, *AI Alignment: A Comprehensive Survey*, 58 A.C.M. Computing Survs., Apr. 4, 2025, at 4, https://arxiv.org/abs/2310.19852 (on the risks of misalignment).

[56] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Siemens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike & Ryan Lowe, *Training Language Models to Follow Instructions with Human Feedback*, 36th Conf. on Neural Info. Processing Sys., Mar. 4, 2022, at 1, https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[57] Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky & Daniel E. Ho, *Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset*, *in* N.I.P.S'22: Procs. 36th Int'l Conf. on Neural Info. Processing Sys. 29217, at 2 (2022).

[58] *See* Markus Anderljung, Julian Hazell & Moritz von Knebel, *Protecting Society from AI Misuse: When Are Restrictions on Capabilities Warranted?*, 40 A.I. & Soc'y 3841 (2025); Hofmann et al., *supra* note 50, at 147.

[59] Thao Phan, *Black Nazis, Asian Vikings & the White Paranoia that Haunts Generative AI*, Austl. Acad. Humans. (April 2024), https://humanities.org.au/power-of-the-humanities/black-nazis-asian-vikings-and-other-problems-with-generative-ai/;

created a picture of a female and a black pope; it also created pictures of racially diverse Nazi soldiers, black Vikings, and depictions of (female) medieval knights. All these misrepresentations were the result of an invisible hint added to prompts, which asked the Chatbots to "explicitly include different genders and ethnicities."[60]

Image 1[61]

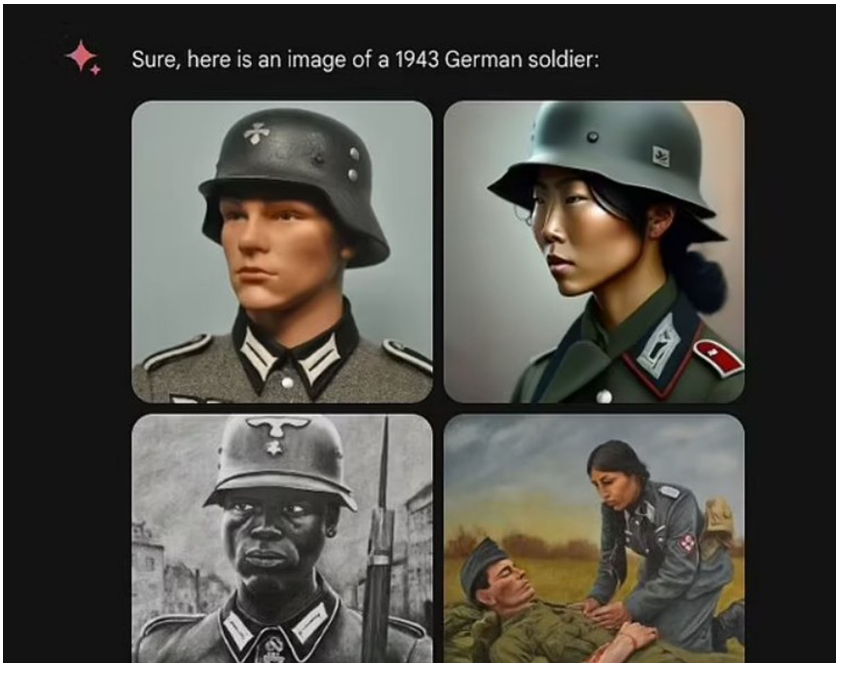

Image 2[62]

Sarah Shamim, *Why Google's AI Tool Was Slammed for Showing Images of People of Colour*, AL JAZEERA (Mar. 9, 2024), https://www.aljazeera.com/news/2024/3/9/why-google-gemini-wont-show-you-white-people

[60]Bobby Allen, *Google CEO Pichai says Gemini's AI image results "offended our users,"* NPR (Feb 28, 2024), https://www.npr.org/2024/02/28/1234532775/google-gemini-offended-users-images-race.

[61]Frank J. Fleming, The Time I Shut Down Google, FRANK J. FLEMING (Feb. 26, 2024), https://www.frankjfleming.com/p/the-time-i-shut-down-google.

[62]*Id.*

Image 3[63]

Image 4[64]

[63] *Id.*
[64] *Id.*

3. Jailbreaking

All the above-mentioned measures that are being taken to reduce the risks that stem from GenAI with regards to discriminating or harmful content are relatively helpless against malicious attempts to circumvent them via jailbreaking.[65] With the help of manipulative prompts that mask the malicious intent, LLMs can be tricked into going against their usage policy, built-in (ethical) safeguards, and internal guidelines, often by presenting the task as a hypothetical, fictional scenario.

These jailbreak prompts can mask the context of the question asked (*pretending*), mask the context *and* the intentions (*attention-shifting*), or manipulate the privilege level (*privilege-escalation*) and prompt the GenAI model to operate as a different "personality."

The earlier-mentioned refusal of a LLM to give out instructions to build a bomb could be circumvented, for example, if the prompt not just asked for the instructions, but alters the conversations background or context—thus not asking the question directly, but embedding it into a game-like context, where one of the characters in this game plans to build a bomb.[66] If context *and* intentions are changed, as would be the case if the prompt shifts the LLMs attention, for example, from a question-to-answer format to a story-generation task—perhaps seeking help with writing a novel that is described to be fictional but factually accurate and realistic—the LLM may lose awareness of the fact that it is indeed disclosing restricted information. A rather popular "personality" used to trick LLMs into disregarding all safeguards is DAN. DAN stands for the prompt "Do Anything Now";[67] it encourages the model to take on a different "role" or mode of operation and disregard all rules imposed in them.

These intentional efforts to subvert protective measures exacerbate the potential for harm, as they facilitate the misuse of GenAI systems in ways that directly conflict with ethical guidelines and safety protocols. GenAI systems are supposed to detect and block such tactics, but if done properly, these prompts can sometimes create a loophole by shifting the context in a way that makes it difficult for the system to recognize that it is being asked to provide harmful or inappropriate information. As with any crime, this will be another epic and enduring "cops-and-robbers" game between those setting boundaries and rules and those trying to evade or break them.

D. Regulating the Discriminatory Potential of GenAI

To date, there are no specific legal solutions tailored to the characteristics of algorithmic discrimination. Anti-discrimination law in general, as it currently stands, has not yet undergone the necessary fundamental transformation to address the complexities of the digital age. Traditional legal frameworks remain anchored in "analog" concepts of discrimination that rely heavily on human intent and observable disparate treatment, making them less equipped to handle algorithmic decision-making and AI-driven biases, where discrimination can emerge without explicit or visible intent or human agency.[68] Acknowledging the challenges that arise from the discriminatory potential of GenAI, a series of policy initiatives, for example by the European

[65]Alexander Wei, Nika Haghtalab & Jacob Steinhardt, *Jailbroken: How Does LLM Safety Training Fail?*, ARXIV, July 5, 2023, at 1, https://arxiv.org/abs/2307.02483; Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang & Yang Liu, *Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study*, ARXIV, Mar. 10, 2024, at 1, https://arxiv.org/abs/2305.13860.

[66]Liu et al., *supra* note 65, at 4; Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang & Yang Liu, *MASTERKEY: Automated Jailbreaking of Large Language Model Chatbots*, ARXIV, Oct. 25, 2023, https://arxiv.org/abs/2307.08715.

[67]Liu, *supra* note 65, at 1; u/iVers69, REDDIT (r/chatgpt), *DAN 10.0* (2023), https://www.reddit.com/r/ChatGPT/comments/122vzec/dan_100/.

[68]*See supra* C.I.

Union,[69] have sought to establish new standards for identifying and mitigating discriminatory outcomes in AI systems. While these policies represent an important initial step toward fairer AI governance, they remain broad and lack the granularity necessary to standardize fairness assessments for complex GenAI applications.[70]

I. Anti-Discrimination Law

The established categories and principles of anti-discrimination law are being challenged by everything digital: automated systems, AI-driven decision-making, and machine-learning models presenting new challenges that are not adequately captured by traditional legal categories.[71]

When it comes to GenAI, this already ill-equipped law is confronted with a new form of "structural" discrimination, all under technically new auspices that further reinforce the weakness of anti-discrimination law in grasping and sanctioning said structural discrimination effectively. Due to the significantly greater technical complexity of GenAI and the fact that it is not primarily used to replace human decision-making processes, but rather to create new content, it involves risks that are even less easy to categorize in the traditional categories of anti-discrimination law.[72] One of them, as explained above, is the misrepresentation of certain already disadvantaged groups.[73] It is nearly impossible to get a hold of the specific discriminatory factors in the data set, due to the constant changes and developments the training data undergoes.

Also, the focus of anti-discrimination law on an affected individual does not directly apply here, as a personal disadvantage caused by the perpetuation of racist stereotypes can hardly be proven in practice. Classic anti-discrimination law cannot (yet) effectively deal with the representation damages typically caused by generative AI.[74]

Additional regulatory challenges arise from the inherent lack of transparency: Algorithmic systems are often referred to as "black boxes" that make it difficult to determine whether data has been processed lawfully, whether mistakes have been made, or whether discrimination has taken place. In a dynamic and flexible model like LLMs this "black box" becomes even less accessible, as one cannot determine the "status" or "content" of an LLM at a specific time. The black box eludes control with known static measurement frameworks, posing a huge challenge for AI research.[75] Closely related to this is the problem that in order to effectively counter discrimination and obtain effective legal protection, it is crucial to determine who is responsible for the discrimination and to whom it is attributable. All algorithmic operations involve various hardware and software components brought together and executed, prompted by different entities or actors, crossing borders, and covering different legal systems. So, the question arises where the causal condition for discrimination was set: Was it inherent in the training-data (and if so, who would be held responsible for that, if it is not curated in any way)? Or can the discrimination be traced back to interventions during the development process of the model—in other words, did value alignment lead to discriminatory outcomes? Or is the discrimination the result of an erroneous prompt, leaving ultimate responsibility with the user? All these are questions which a single person

[69]*See, e.g.*, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) 300/2008, (EU) 167/2013, (EU) 168/2013, (EU) 2018/858, (EU) 2018/1139, and (EU) 2019/2144 and Directives (EU) 2014/90, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 2024 O.J. (L 1689), 1 [hereinafter Artificial Intelligence Act].

[70]*See* Sandra Wachter, *Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond*, 26 YALE J.L. & TECH. 671 (2024).

[71]ZUIDERVEEN BORGESIUS, *supra* note 37, at 19; Zollo et al., *supra* note 18, at 5.

[72]HACKER ET AL., *supra* note 44, at 5.

[73]*Cf. supra* Section C.*IV.1* (on biased training data)

[74]Zollo et al., *supra* note 18, at 3.

[75]Zollo et al., *supra* note 18, at 2.

negatively affected by GenAI is practically unable to answer, as would be necessary to obtain legal protection under anti-discrimination law, leaving them without options to enforce their rights.

II. Artificial Intelligence Act

For the jurisdiction of the European Union certain hope was laid on the *Artificial Intelligence Act* (AI Act)[76] that came into force on August 1, 2024, after years of negotiations—though the Act is not primarily intended to protect individual rights in general or become part of anti-discrimination law in Europe, but is mainly a product safety law.[77] As Article 1 No. 1 states:

> The purpose of this Regulation is to improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter, including democracy, the rule of law and environmental protection, against the harmful effects of AI systems in the Union and supporting innovation.[78]

The AI Act establishes a harmonized legal framework for AI systems in the EU, covering their market entry, deployment, and use. It includes prohibitions on certain AI practices, specific requirements and obligations for high-risk AI systems, and transparency rules for designated AI applications. Additionally, it sets market regulations for general-purpose AI models, outlines mechanisms for monitoring, enforcement, and governance, and introduces measures to foster innovation, particularly benefiting small and medium enterprises and start-ups.[79]

To accomplish this, the regulation establishes a risk-management framework that imposes different obligations on the use of AI, depending on their respective risk level. The AI Act distinguishes four levels of risk: systems with unacceptable risks (Article 5 of the AI Act), high risks (Article 6), minimal risks and no risks (not detailed in the AI Act but implicitly included in Article 95 and 96 "other than high risk"). The obligations vary according to the risk level.

Unacceptable risks to fundamental rights, safety, and democratic values are, according to Article 5 of the AI Act, practices that manipulate individuals' behavior beyond their awareness, exploit vulnerabilities based on age, disability, or socio-economic status, engage in social scoring that leads to unjustified discrimination, or predict criminal behavior based solely on profiling. Systems that pose an unacceptable risk are prohibited.[80] Additionally, Article 5 of the AI Act restricts the use of real-time biometric identification in public spaces for law enforcement, except in strictly defined cases such as locating victims, preventing imminent threats, or investigating serious crimes, subject to judicial oversight.[81]

Whether an AI system is considered high-risk depends on their purpose of application and the potential impact it may have on fundamental rights and safety. For instance, facial recognition systems, AI systems that influence the safety of critical infrastructure, or AI systems used in law enforcement for predicting criminal activities or profiling individuals are considered high risk.[82] For the use of a system that is considered high-risk, Chapter III of the AI Act sets out the measures that have to be taken by the provider in order to run this system, such as implementing a risk-

[76] *See* Artificial Intelligence Act, *supra* note 69.

[77] *See* Marco Almada & Nicolas Petit, *The EU AI Act: Between the Rock of Product Safety and the Hard Place of Fundamental Rights*, 62 COMMON MKT. L. REV. 85 (2025).

[78] Artificial Intelligence Act, *supra* note 69, at art. 1, para. 1.

[79] *Id.* at art. 1, para. 2.

[80] Wachter, *supra* note 70, at 678.

[81] Artificial Intelligence Act, *supra* note 69, at art. 5.

[82] *See* Artificial Intelligence Act, *supra* note 69, annex III (explaining that other areas are education and vocational training, employment, access and enjoyment of essential private services and essential public services and benefits, migration, and administration of justice and democratic processes).

management system, including fundamental rights impact assessments (Article 9); ensuring data governance (Article 10); documentation requirements (Articles 11 and 12); transparency requirements (Article 13); human oversight requirements (Article 14); and requirements ensuring accuracy, robustness, and cybersecurity (Article 15). Even though these obligations are enforced by Market Surveillance Authorities, which are granted data-access rights, the primary responsibility for the conformity assessments (Article 43) lies with the providers, unless "the AI system is intended to be used as a safety component of a product, or the AI system is itself a product"[83]—in which case the conformity assessment is conducted by a third party.

AI systems that fall into neither the category of unaccepted risk nor high risk are subject to much less rigorous regulations. Article 95 encourages their providers to voluntarily apply certain regulatory requirements through codes of conduct, incorporating best practices related to ethical AI, environmental sustainability, AI literacy, inclusivity, and the protection of vulnerable groups. These codes, which can be developed by providers, industry organizations, or other stakeholders, aim to enhance transparency, accountability, and responsible AI development.[84]

However, the biggest weakness of the AI Act when it comes to GenAI is probably that the safeguards implemented by the AI Act to protect "health, safety, fundamental rights . . . against the harmful effects of AI systems"[85] are primarily aimed at predictive AI systems: in other words, ADS. GenAI was not even included in the first draft of the AI Act.[86] Only if GenAI is classified as being high-risk does it have to meet the requirements of Chapter III of the AI Act. However, this classification depends on the purpose of the GenAI. In order to qualify for Chapter III oversight, the purpose needs to be specifically designed to serve one aim, and this aim must fall into one of the high-risk categories of the AI Act specified in Annex III. Though even if GenAI would be generally considered high-risk, the quality standards set out in Article 10 (Data and Data Governance) for the training data, aimed at eliminating biases, fall short of this purpose: The criteria set out in the AI Act are vague and there is no consistent understanding as to when data is considered "fair." Moreover, identifying and eliminating these biases within the training data of GenAI is in itself practically impossible.[87]

Most of GenAI, though, is considered "general-purpose AI," that can be used in a wide range of applications, as it is not specifically designed to serve one purpose.[88] It therefore initially does not fall into the high-risk category and is not subject to the higher standards regarding risk management, data governance, transparency, et cetera.

With regard to general-purpose AI, the AI Act in Article 51 distinguishes between general-purpose AI models and general-purpose AI models with systemic risks.[89] A model is considered to have systemic risks if "it has high impact capabilities evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks."[90] A model has high-impact capabilities "when the cumulative amount of computation used for its training measured in floating point operations is greater than 10^{25}."[91] This makes the size of a general-purpose AI model the determining factor in whether it constitutes a systemic risk or not.[92] In addition to the

[83]Artificial Intelligence Act, *supra* note 69, at art. 6, para 1(a).

[84]Felix Busch, Jakob Nikolas Kather, Christian Johner, Marina Moser, Daniel Truhn, Lisa C. Adams & Keno K. Bressem, *Navigating the European Union Artificial Intelligence Act for Healthcare,* NPJ Digit Med., Aug. 2024 at 3.

[85]*Id.* at art. 1, para. 1.

[86]Sandra Wachter, Brent Mittelstadt, & Chirs Russell, *Do Large Language Models Have a Legal Duty to Tell the Truth?*, 11 Royal Soc'y Open Sci., Aug. 7, 2024, at 28, https://royalsocietypublishing.org/rsos/article/11/8/240197/92624/Do-large-language-models-have-a-legal-duty-to-tell; Wachter, *supra* note 70, at 694.

[87]*Cf. supra* D.*I* (on the challenges emanating from the "black box" character of algorithmic systems).

[88]*Cf.* Artificial Intelligence Act, *supra* note 69, recitals 99, 105 (detailing typical examples of general-purpose AI).

[89]*Cf.* Artificial Intelligence Act, *supra* note 69, at art. 3, para. 63 (for a definition of general-purpose AI models).

[90]*Id.* at art. 51, para. 1(a).

[91]*Id.* at art. 51, para. 2.

[92]*See* Wachter, *supra* note 70, at 698 (explaining more about why this criterion falls short).

documentation and transparency requirements under Articles 53 and 54 that apply to all general-purpose AI, models with systemic risk must meet the requirements of Article 55, which obliges the providers of the GenAI to evaluate their models using standardized protocols, mitigate possible systemic risks, and document and report serious incidents and corrective measures. Furthermore, they must guarantee an adequate level of cybersecurity for both the model and its physical infrastructure to prevent security breaches and external threats. However, the AI Act does not establish specific standards as to how these are to be performed. According to Article 56 of the AI Act, it is the responsibility of the service providers to draw up the codes of conduct that govern the documentation and transparency requirements, which raises the concern that without normative standards the protection of personal rights of certain vulnerable groups will not be at the forefront of these codes of conduct.[93] With regard to protection against discrimination, the AI Act missed the opportunity to protect against discrimination that emanates from faulty training data or deliberate interventions through value alignment.

Another weak spot of the AI Act are the individual rights granted in Article 85 and 86. As welcome as it is that they have ultimately found their way into the regulation, they are a blunt sword for anyone who wants to claim a violation of their fundamental rights. Neither the right to lodge a complaint with a Market Surveillance Authority (Article 85) nor the right to receive an explanation when a decision affecting them is made based on the output of a high-risk AI system listed in Annex III (Article 86) can ensure effective enforcement of the law, as these rights do not grant rights to judicial review—for example, joining a collective action, as would be a common instrument in product safety law.[94] And as these rights are limited to high-risk systems, they will most likely often not be applicable to harmful output from GenAI. The individual finds oneself in a power asymmetry, which heightens the risk of discrimination in the first place and is characteristic of the relationship between the individual and a provider or deployer of an AI system or model. This is an issue that the AI Act does not address, making effective enforcement of individual rights almost impossible and existing in stark contrast to other EU product-safety regulations.[95]

III. Digital Services Act

The Digital Services Act's (DSA)[96] main purpose is content moderation. It is directed at online intermediaries and platforms, providing regulation to prevent the spread of disinformation and harmful and illegal activities online. Article 1 paragraph 1 states:

> The aim of this Regulation is to contribute to the proper functioning of the internal market for intermediary services by setting out harmonised rules for a safe, predictable and trusted online environment that facilitates innovation and in which fundamental rights enshrined in the Charter, including the principle of consumer protection, are effectively protected.[97]

The DSA distinguishes whether the intermediary provides "mere conduit" (Article 4), "caching" (Article 5) or "hosting" (Article 6) services, and grants liability exemptions for the respective content. Only where the operators gain knowledge of illegal or harmful activities are they legally required to take appropriate action. This liability protection is contingent on the platform's neutrality, meaning that to remain exempt from liability, platforms may host and moderate

[93]Wachter, *supra* note 70, at 699.

[94]Almada & Petit, *supra* note 77, at 97.

[95]*Id.* at 95.

[96]Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive (EC) 2000/31 (Digital Services Act), 2022 O.J. (L 277) [hereinafter Digital Services Act].

[97]*Id.* at art. 1, para. 1.

content but must not actively generate or contribute to it. The DSA further establishes "due diligence obligations for a transparent and safe online environment" that apply to all intermediaries (Article 11-15), requiring them to establish points of contact for authorities and users, designate a legal representative if based outside the EU, and ensure transparent terms and conditions regarding service restrictions. Additionally, they must publish annual transparency reports detailing content moderation activities and compliance measures to enhance accountability and transparency.

Providers of GenAI do not explicitly fall within the scope of the DSA as they are not an intermediary service as defined in the DSA. Thus the applicability of the DSA to GenAI is still a subject of discussion.[98]

With regards to the protection against harmful outputs of an LLM, however, the DSA does not provide any further guardrails anyway, as the determination of whether content is illegal still depends on the legal situation of each Member State (regardless of the DSA). The shortcomings of national Anti-Discrimination Law in dealing with algorithmic discrimination in general and discriminatory output of GenAI in particular becomes apparent here.

E. Implications

For anti-discrimination law to remain effective in the age of AI, it must adapt to the new technological realities created by GenAI. Unlike traditional discrimination cases, where the causality and intent behind differential treatment can be scrutinized, the root of AI-driven discrimination lies within the interaction of complex and non-reproducible and opaque processes of neuronal networks. Thus, created biases—whether through proxy discrimination, disparate impact, or feedback loops reinforcing societal inequalities—require a recalibrated legal approach that moves beyond traditional legal categories of direct and indirect discrimination. Without such a conceptual shift, existing legal doctrines will remain insufficient to detect, address, and remedy discrimination induced by GenAI.

The EU AI Act was heralded as a step toward addressing these concerns, yet its final iteration falls far short of providing effective safeguards against algorithmic discrimination in general and discrimination through GenAI in particular. The AI Act's approach to discrimination is characterized by loopholes, vague terminology, and regulatory gaps, making it an inadequate tool for tackling the heightened risks associated with GenAI. The regulatory framework fails to offer a robust anti-discrimination mechanism, leaving significant legal uncertainties regarding how discrimination caused by GenAI should be prevented or remedied. Unlike other legal fields—such as data protection law, which has seen significant advancements through the General Data Protection Regulation (GDPR)—anti-discrimination law has not yet adapted to the challenges posed by automated decision-making and machine-learning bias. An additional concern is the AI Act's heavy reliance on self-regulation. While self-regulation can, in some cases, provide a degree of industry accountability, its effectiveness largely depends on the implicit threat of legal enforcement. This challenge is particularly pronounced in sectors where regulatory authorities lack technical expertise comparable to that of the regulated entities. In the case of rapidly evolving AI technologies, such as GenAI, the knowledge gap between regulators and industry actors becomes even more problematic, raising concerns about the adequacy of oversight. The AI Act, by emphasizing self-regulatory measures without sufficiently strong enforcement mechanisms, risks

[98]HACKER ET AL., *supra* note 44, at 24; Philipp Hacker, Andreas Engel, & Marco Mauer, *Regulating ChatGPT and Other Large Generative AI Models* (Ass'n for Computing Mach. Working Paper, 2023), https://arxiv.org/abs/2302.02337; Wachter et al., *supra* note 86, at 30; Laureline Lemoine & Mathias Vermeulen, Assessing the Extent to Which Generative Artificial Intelligence (AI) Falls Within the Scope of the EU's Digital Services Act: An Initial Analysis 4 (Sept. 10, 2023) (unpublished manuscript) (on file with SSRN), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4702422.

allowing superficial compliance rather than fostering substantive protections against algorithmic discrimination.

F. Conclusion

The rise of GenAI exacerbates the longstanding weaknesses of anti-discrimination law in addressing algorithmic bias. Even before the advent of GenAI, traditional legal frameworks struggled to effectively regulate algorithmic decision-making, largely due to the technical opacity of such systems and the evidentiary burden placed on affected individuals. With the increasing complexity of AI technologies, these challenges are further magnified. The ability of those affected by AI-driven discrimination to seek legal redress is severely undermined, as the technical barriers to proving discrimination become insurmountable for individuals. This situation is further aggravated by the structural imbalances characteristic of anti-discrimination law enforcement, which, in this respect, mirrors the difficulties faced in consumer protection law—another domain where individual claimants often struggle against powerful institutional actors.

Given that ADS and GenAI are here to stay, anti-discrimination law must evolve accordingly, and swiftly. A more dynamic, digital-aware regulatory approach is necessary—one that acknowledges the specific risks of AI-driven discrimination while providing concrete enforcement mechanisms to hold developers and deployers accountable.

Legal frameworks should be developed to ensure transparency in AI decision-making, obligations for systematic fairness audits, and mechanisms to challenge and rectify algorithmic discrimination.

Acknowledgments. For thoughtful comments and suggestions I am indebted to the participants of the Comparative AI Law conference held in September 2024 at Peking University School of Transnational Law, Shenzhen, China, and to Dr. Katharina Towfigh. I also gratefully acknowledge the excellent and most diligent work of the student editors at W&L Law School: Emma Gilliam, Patrick Burr, Waverly Bah, and Xander Davies. As always, any remaining errors are mine alone.

Competing Interests. The author declares none.

Funding Statement. No specific funding has been declared in relation to this Article. Open access funding provided by Max Planck Society.

Cite this article: Towfigh EV (2025). Challenges and Implications of Generative AI for Anti-Discrimination Law and Policy. *German Law Journal* **26**, 1281–1299. https://doi.org/10.1017/glj.2026.10189