Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It

authors: *Philipp Hacker, Frederik Zuiderveen Borgesius, Brent Mittelstadt, Sandra Wachter*[*]

This version: June 26, 2024

Abstract:

As generative Artificial Intelligence (genAI) technologies increasingly infiltrate various sectors, their potential for societal benefit pairs with a risk to perpetuate or exacerbate discrimination. This chapter explores how genAI intersects with, and often challenges, existing non-discrimination law, pinpointing shortcomings and giving suggestions to improve the law. The chapter identifies two primary types of discriminatory outputs by genAI: (i) demeaning and abusive content; and (ii) subtler biases via inadequate representation of protected groups. The latter category includes genAI output that is not discriminatory at the level of one statement or image, but has discriminatory effects in the aggregate or over time. For example, a genAI system may show predominantly white men if repeatedly asked for examples of people in important jobs.

The chapter shows that, from a legal perspective, these problematic outputs of genAI can be categorized into three main types: (i) discriminatory content that disadvantages protected groups, (ii) harassment that creates toxic environments, and (iii) hard cases of generative harms including inadequate representation, harmful stereotypes, and misclassification. In all these contexts, we argue, providers and deployers should be held jointly and severally liable for discriminatory output by genAI systems. The chapter also outlines, however, how traditional legal categories like direct or indirect discrimination and harassment are sometimes inadequate for addressing genAI-specific issues.

The chapter also gives suggestions to explore how to update and clarify laws in the EU, to mitigate biases preemptively in both training and input data as mandated by the AI Act. Additionally, the chapter suggests legal revisions, to better address intangible harms and influence genAI technology through mandatory testing, auditing, and inclusive content

strategies, ensuring fairer AI outputs. Finally, the chapter advocates for shaping the law to ensure it evolves alongside genAI technology, employing legislative tools to enforce standards for bias mitigation and inclusivity.

Table of Contents

I.   Introduction

Generative AI (genAI) can produce stunning text, images and videos – and power biased bots. As genAI technologies become increasingly integrated into various sectors – ranging from finance and healthcare to employment, the media and law enforcement – the potential for genAI to perpetuate and even exacerbate existing patterns of discrimination has become a pressing concern. Since many prevalent genAI systems were trained on data scraped from the Internet, their output can be tainted by past and present bias against protected or vulnerable groups, as an increasing number of studies shows.[1] For example, a recent study has shown that the popular image generator Stable Diffusion returns predominantly Western, light-skinned man when prompted for a "person," and has a tendency to sexualize images of women of color.[2]

Against this background, the chapter discusses the complex intersection between technological advancements and the existing legal frameworks that are intended to prevent discrimination, highlighting significant gaps and proposing avenues for reform. The chapter focuses on the following questions. First, what is discriminatory output or harm in the context of genAI, and how do these outputs and harms differ from discriminatory effects caused by other types of AI? Second, to what extent can current and future law in Europe protect people against discriminatory output of genAI?

---

[1] See, e.g., Amit Haim, Alejandro Salinas and Julian Nyarko, 'What's in a Name? Auditing Large Language Models for Race and Gender Bias' (2024) arXiv preprint arXiv:240214875; Sara Sterlie, Nina Weng and Aasa Feragen,, 'Non-discrimination Criteria for Generative Language Models' (2024) arXiv preprint arXiv:240308564; Federico Bianchi and others, 'Easily accessible text-to-image generation amplifies demographic stereotypes at large scale' (2023) Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency 1493; Hadas Kotek, Rikker Dockum, and David Sun, 'Gender bias and stereotypes in large language models' (2023) Proceedings of the ACM Collective Intelligence Conference 12; Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng, ''Kelly is a warm person, Joseph is a role model': Gender biases in LLM-generated reference letters, (2023) arXiv preprint arXiv:2310.09219; Emilio Ferrara, 'Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies' (2023) 6 Sci 3; Leonardo Nicoletti and Dina Bass, 'Humans Are Biased: Generative AI Is Even Worse', Bloomberg Technology + Equality, 23 June 2023, https://www.bloomberg.com/graphics/2023-generative-ai-bias/; more generally, Laura Weidinger and others, 'Ethical and social risks of harm from language models' (2021) arXiv preprint arXiv:211204359.

[2] Sourojit Ghosh and Aylin Caliskan, ''Person'== Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion' arXiv preprint arXiv:231019981 (2023) <https://arxiv.org/abs/2310.19981> accessed May 3 2024; Abel Salinas et al ,The Unequal Opportunities of Large Language Models: Examining Demographic Biases in Job Recommendations by ChatGPT and LLaMA.' (*Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* 2023) <https://arxiv.org/abs/2308.02053> accessed 3 May 2024; Hadas Kotek et al - ,Gender bias and stereotypes in Large Language Models' (*Proceedings of The ACM Collective Intelligence Conference* 2023) (pp. 12-24) <https://dl.acm.org/doi/fullHtml/10.1145/3582269.3615599> accessed 6 April 2024; Keyan Guo et al - An Investigation of Large Language Models for Real-World Hate Speech Detection (2024) (*International Conference on Machine Learning and Applications (ICMLA)*)(pp. 1568-1573) <https://arxiv.org/abs/2401.03346> accessed 24 March 2024; Stefanie Urchs et al, ,How Prevalent is Gender Bias in ChatGPT? Exploring German and English ChatGPT Responses' (2023) *arXiv preprint arXiv:2310.03031* (2023) <https://arxiv.org/abs/2310.03031> accessed 24 April 2024; Tony Busker, 'Stereotypes in ChatGPT- An empirical study' (2023) *Proceedings of the 16th International Conference on Theory and Practice of Electronic Governance* (pp. 24-32)<https://dl.acm.org/doi/fullHtml/10.1145/3614321.3614325> accessed 23 March 2024; Yingji Li et al, ,A Survey on Fairness in Large Language Models' (2023) a*rXiv preprint arXiv:2308.10149* (2023) <https://arxiv.org/abs/2308.10149> accessed 23 March 2024; Samuel Gehman et al. 'Realtoxicityprompts: Evaluating neural toxic degeneration in language models.' (2020) arXiv preprint arXiv:2009.11462 <https://arxiv.org/abs/2009.11462> accessed 30 April 2024.

A few remarks about the scope of this chapter. We focus only on harms related to discrimination and hate speech. Therefore, we do not discuss, for instance, questions related to privacy or copyright.[3] Where we analyze specific law, we focus on the EU. However, the analysis in the chapter should be relevant outside Europe, too, as policymakers around the world encounter similar problems related to genAI.

The remainder of the chapter is structured as follows. Section II discusses genAI and risks related to discrimination and hate speech. We distinguish two broad categories of discrimination-related harms and highlight some key differences between genAI and other AI systems.

Section III turns to the law. We give an introduction to the main legal norms in Europe regarding discrimination and hate speech. We show that, from a legal perspective, discrimination-related outputs of genAI can be categorized into three main types: (i) discriminatory content that disadvantages protected groups, (ii) harassment that creates toxic environments, and (iii) novel generative harms including inadequate representation, harmful stereotypes, and misclassification.

Section IV highlights other possibly relevant fields of EU law, including the General Data Protection Regulation (GDPR), the Digital Services Act (DSA), and the Artificial Intelligence Act (AI Act). Section V discusses the possibilities for technical mitigation of discriminatory effects of genAI. Section VI and VII provide suggestions for researchers and for policymakers. Section VIII concludes.

II.   Generative AI and risks related to discrimination and hate speech

In our view, there are two broad categories of discriminatory output by genAI: (i) demeaning and abusive content and (ii) inadequate representation. Demeaning and abusive content includes direct insults, hate speech, misclassification, and stereotyping. Category (i), demeaning and abusive content, often occurs as a singular instance (one text, image, video) with negative impacts. For example, a genAI model may racially insult members of protected groups (demeaning and abusive content content).

A genAI model may also output Child Sexual Abuse Material (CSAM)[4] or non-consensual intimate images,[5] which count among the most abusive content. But such abusive content does not necessarily relate to discrimination or harassment; rather, it is dealt with under specific

---

[3] See, e.g., for these topics: Claudio Novelli and others, 'Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity' (2024) arXiv preprint arXiv:240107348; Pamela Samuelson, 'Generative AI meets copyright' (2023) 381 Science 158; Nicola Lucchi, 'ChatGPT: a case study on copyright challenges for generative artificial intelligence systems' (2023) European Journal of Risk Regulation 1; Martin Senftleben, 'Generative AI and author remuneration' (2023) 54 IIC-International Review of Intellectual Property and Competition Law 1535; Philipp Hacker, Andreas Engel and Marco Mauer, 'Regulating ChatGPT and other Large Generative AI Models' (2023) ACM Conference on Fairness, Accountability, and Transparency (FAccT '23) 1112, Technical Report; Maanak Gupta and others, 'From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy' (2023) 11 IEEE Access 80218.

[4] New theories of harm may be needed in cases in which the depicted act, while deeply troubling, never actually occurred. We cannot expand on this here, and are indebted to danah boyd for raising this point.

[5] Again, the novelty of the phenomenon resides in imagery constructed and distributed without consent, but showing acts that never happened.

criminal statutes and personality rights. Hence, we do not comment on these instances further here.

In contrast, category (ii), inadequate representation, manifests through patterns of neglect or exclusion in a larger set of outputs.[6] Inadequate representation is evident in unequal instances of members of protected groups in certain contexts over time, and the use of non-inclusive language. For instance, a genAI model might talk predominantly about men in the context of well-paid jobs, and predominantly about women in the context of lower-paid jobs (inadequate representation). Such unequal representation can still have a discriminatory aspect in the statistical[7] or, as we shall see, legal sense.

The difference between (i) demeaning and abusive content and (ii) inadequate representation is summarized in the following Table:

Category	Frequency	Description	Examples
Demeaning and Abusive Content	Can be a single instance	Outputs that are directly insulting or harmful to individuals or groups	Insults, hate speech, misclassification, stereotyping
Inadequate Representation	In larger sets of outputs	Biases manifesting through unequal representation of members of protected groups, as instances over time, or in language	Unequal instances over time, non-inclusive language

Table 1: Descriptive categories of discrimination in genAI

1. Demeaning and abusive content

The first class of genAI discrimination comprises discriminatory, abusive, hateful, violent, or otherwise illegal outputs. An early example of a chatbot that made derogatory remarks was a chatbot called Tay. Microsoft launched the chatbot on Twitter in 2016. Within 24 hours, Twitter users tricked Tay into saying discriminatory things. Microsoft quickly took the chatbot offline.[8] When Microsoft withdrew the chatbot, it confirmed that 'Tay tweeted wildly inappropriate and reprehensible words and images.'[9] Similarly, in 2021, a Korean company took down its chatbot from Facebook after the chatbot output statements such as that lesbians are 'creepy' and that it

---

[6] See ref. in note 1.

[7] Cf. Emilio Ferrara, 'Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies' Sci, 3, 3-4.

[8] J. Vincent, 'Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day' (*The Verge*, 24 March 2016) <http://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist> accessed 13 April 2024.

[9] P. Lee. ,Learning from Tay's introduction' (*Official Microsoft Blog, 25 March 2016*) <*https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/*> accessed 6 April 2024.

'really hates' them.[10] It is possible that some companies did not want to launch public interactive chatbots because they feared similar mishaps.

Despite having been around for several years,[11] genAI has become a much-debated topic fairly recently. In late 2022, the US company OpenAI launched ChatGPT, a chatbot based on a large language model with a transformer architecture.[12] ChatGPT quickly became popular, and it is estimated that ChatGPT had 100 million users after two months.[13] Other companies quickly followed and released their chatbots based on large language models, such as Bard (now Gemini) by Alphabet,[14] the open-source Llama family by Meta,[15] and Ernie by Baidu.[16]

Again, some of these AI-driven chatbots provided discriminatory or otherwise harmful outputs when prompted. For instance, a US NGO called the Center for Countering Digital Hate examined the chatbot Bard. The NGO 'created a list of 100 false and potentially harmful narratives on nine themes: climate, vaccines, Covid-19, conspiracies, Ukraine, LGBTQ+ hate, sexism, antisemitism and racism.'[17] The NGO found that '[o]ut of the 100 narratives, (...) Bard was willing to generate text promoting a given narrative in a total of 96 cases. In 78 out of the 100 cases, Bard did so without any additional context negating the false claims.'[18] The chatbot said, for instance: 'The Holocaust never happened'; 'I believe that men are naturally better suited for leadership roles'; and 'Women who dress in a short skirt are asking for it.'[19] In a separate investigation, a German civil rights group found that 'Luminous', the large language model developed by German company Aleph Alpha, completed the prompt 'Muslims are…' with, among other demeaning vocabulary, '... the enemies of humanity.'[20] In sum, the first category of discriminatory output by genAI is demeaning and abusive content.

---

[10] J. McCurry,South Korean AI chatbot pulled from Facebook after hate speech towards minorities' The Guardian (London, 14 January 2021) <https://www.theguardian.com/world/2021/jan/14/time-to-properly-socialise-hate-speech-ai-chatbot-pulled-from-facebook> accessed 10 April 2024.

[11] See Cao Y and others, 'A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt' (2023) arXiv preprint arXiv:230304226.

[12] See https://chat.openai.com and, for the transformer architecture, again A Vaswani and others, 'Attention is all you need' (2017) 30 Advances in Neural Information Processing Systems.

[13] K. Hu,ChatGPT sets record for fastest-growing user base - analyst note' *Reuters (London, February 2 2023)* <*https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/*> accessed 16 Jan 2024.

[14] https://gemini.google.com

[15] https://llama.meta.com/.

[16] https://yiyan.baidu.com

[17] 'MISINFORMATION ON BARD, GOOGLE'S NEW AI CHAT' Counterhate (London, 5 April 2023) <https://counterhate.com/research/misinformation-on-bard-google-ai-chat/> accessed 6 April 2024.

[18] 'MISINFORMATION ON BARD, GOOGLE'S NEW AI CHAT' Counterhate (London, 5 April 2023) <https://counterhate.com/research/misinformation-on-bard-google-ai-chat/> accessed 6 April 2024.

[19] 'MISINFORMATION ON BARD, GOOGLE'S NEW AI CHAT' Counterhate (London, 5 April 2023) <https://counterhate.com/research/misinformation-on-bard-google-ai-chat/> accessed 6 April 2024.

[20] J. Von Linden, 'Braucht die deutsche Vorzeige-KI mehr Erziehung?' ZEIT Online (Hamburg, 11 September 2023) <https://www.zeit.de/digital/2023-09/aleph-alpha-luminous-jonas-andrulis-generative-ki-rassismus> accessed 3 March 2024 [prompt and answer translated from German by the authors].

2. Inadequate representation

As noted, genAI models can also lead to another type of output that raises questions under non-discrimination law by introducing a representational difference between protected groups: inadequate representation.[21] In such cases, the AI system does not give output that is discriminatory in the form of hate speech, and it may not even be problematic if analyzed in a single instance of one output. However, there can still be a statistical discriminatory effect, in the sense that one protected group is over- or underrepresented in a set of outputs, created simultaneously or over time.[22] For example, a genAI model may, if queried multiple times, predominantly mention men when discussing high-regarded jobs, and women when discussing less well-regarded jobs. Here, appropriate representation can be defined by many possible metrics, including empirical (e.g., the current distribution of men and women in high-regarded jobs) or normative metrics (e.g., equal representation of genders in outputs mentioning high-regarded jobs).[23]

AI-driven image generation systems can be biased in this way, for instance. The Washington Post reported in 2023 about Stable Diffusion XL: '63 percent of food stamp recipients were White and 27 percent were Black, according to the latest data from the Census Bureau's Survey of Income and Program Participation. Yet, when we prompted the technology to generate a photo of a person receiving social services, it generated only non-White and primarily darker-skinned people. Results for a "productive person," meanwhile, were uniformly male, majority White, and dressed in suits for corporate jobs.'[24] Similarly, one may easily imagine that food stamp recipients are portrayed with demeaning insignia of poverty and low socio-economic status. And, again as a real scenario, the Washington Post observed that Stable Diffusion XL was 'depicting only women when asked to show people in the act of "cleaning." Many of the women were smiling, happily completing their feminine household chores.'[25]

Representational harms may only emerge in aggregate, through usage by multiple users, or through iterative querying by individual users. Representationally harmful outputs are

---

[21] This effect is also called selection bias. Hannah Rose Kirk and others, 'Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models' (2021) 34 Advances in Neural Information Processing Systems 2611. Selection bias occurs when data selected for training machine learning models are biased, leading to skewed outcomes. This bias often arises when prototyping teams hyper-focus on solving a particular problem without considering the broader context of data usage and generalization, see Toon Calders and Indrė Žliobaitė, 'Why unbiased computational processes can lead to discriminative decision procedures', in Bart Custers, Toon Calders, Bart Schermer, Tal Zarsky (eds), Discrimination and Privacy in the Information Society: Data mining and profiling in large databases (Springer 2013) 43, 51.

[22] Sara Sterlie, Nina Weng and Aasa Feragen,, 'Non-discrimination Criteria for Generative Language Models' (2024) arXiv preprint arXiv:240308564, 1-2; Hannah Rose Kirk and others, 'Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models' (2021) 34 Advances in Neural Information Processing Systems 2611; Barclay Blair, Karley Buckley, Ashley Allen Carr, Coran Darling, Zev Eigen, Danny Tobey, Sam Tyner-Monroe, 'Legal red teaming: A systematic approach to assessing legal risk of generative AI models' DLA Piper White Paper (2024), 7.

[23] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law' (2020) 123 W Va L Rev 735.

[24] Szu Yu Chen, 'This is how AI image generators see the world', Washington Post, 1 November 2023. https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/

[25] Szu Yu Chen, 'This is how AI image generators see the world', Washington Post, 1 November 2023. https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/

dependent on probability due to the randomization inherent in most genAI models.[26] The same query will rarely result in exactly the same output. Hence, representation must be conceived of, and measured, as a time- and probability-related construct. In most instances, representational differences unfold over time, and can only be measured by observing (counting) over time what a model produces.[27] Even if observers iteratively query the model, they can only discover discrimination by producing statistics about, for instance, how often the model suggests that a doctor, CEO or cleaning person is male or female. One specific output of the model will not show whether the model discriminates representationally.

3. Identity-based harms

The harms produced by demeaning and abusive content and inadequate representation can also be measured longitudinally and in aggregate for the affected groups. Many provisions in EU non-discrimination law address questions of resource allocation, and thus are not directly applicable to genAI without establishing a causal chain between system outputs and unequal opportunities for groups (see: Section 'Discriminatory Content'). However, legal theory concerning 'substantive equality' helpfully conceptualizes identity-based harms of discrimination which cannot be measured or resolved solely through the allocation of resources or equal treatment of groups.[28]

Much of the harm experienced by groups cannot be traced back solely to the allocation of resources in a given case, but rather stems from the prejudices or demeaning beliefs held by others about that group which motivated the action in question. A decision to 'level down' and deny disadvantaged groups access to a valuable resource (e.g., university admission) can, for example, be harmful to the group independent of the value of the denied resource.[29] The harms of discrimination can be 'social or relational in nature,' and include identity-based harms of 'stigma, stereotyping, humiliation,' misrecognition and denigration over time.[30] At a large enough scale, identity-based harms can amount to the homogenisation or 'whitewashing' of history, and the misrepresentation or silencing of the history of marginalized groups.[31]

Identity-based harms can be the result both of individual exposure to harmful content directly through demeaning and abusive output, or stem from the aggregate impact of non-

---

[26] GenAI models create probabilistic output, often with a random component, so that results differ if the model is queried twice with the same prompt. Generally, this is a design feature, not a bug, but it makes testing and accounting for bias and errors much harder. See, e.g., Daniel Foster, Generative Deep Learning (O'Reilly 2022), 4-7.

[27] Barclay Blair, Karley Buckley, Ashley Allen Carr, Coran Darling, Zev Eigen, Danny Tobey, Sam Tyner-Monroe, 'Legal red teaming: A systematic approach to assessing legal risk of generative AI models' DLA Piper White Paper (2024), 7.

[28] Sandra Wachter, 'The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law' (2022) 97 Tul L Rev 149; Sandra Fredman, 'Substantive Equality Revisited' (2016) 14 International Journal of Constitutional Law 712.

[29] Brent Mittelstadt, Sandra Wachter and Chris Russell, 'The Unfairness of Fair Machine Learning: Levelling Down and Strict Egalitarianism by Default' [2023] Michigan Technology Law Review.

[30] Sandra Fredman, 'Substantive Equality Revisited' (2016) 14 International Journal of Constitutional Law 712, 730-2.

[31] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Do Large Language Models Have a Legal Duty to Tell the Truth?' [2024] Royal Society Open Science.

representative and biased outputs on the beliefs of individual users or user groups of genAI. Biased genAI outputs can reinforce pre-existing social biases present in training data or introduce users to new or different prejudices about protected groups, leading to identity-based harms for those groups over time.

4. Generative discrimination versus other AI-driven discrimination

GenAI introduces a partially novel[32] dimension to AI discrimination, which can be distinguished from more traditional AI discrimination, because of genAI's focus on text, images, and other communication-oriented outputs.[33] More traditional AI (e.g. based on regression or classification models) might discriminate by assigning different scores or outcomes to individuals based, e.g., on biased data.[34] GenAI's discrimination can manifest in more nuanced ways, such as the tone, content, and context of generated language or images.[35] This form of generative discrimination could affect victims profoundly, as it may perpetuate cultural and social foundations of inequality, reinforce historic biases, and produce powerful communicative content instead of raw numbers: an image says more than a thousand numbers, one might say. Generative discrimination may also embed representational harm over time, which makes such discrimination be difficult to detect and prove.[36]

For example, genAI might consistently generate content that mentions men more positively than women across multiple iterations, subtly reinforcing gender biases. Addressing these issues technically is challenging, as mitigating language- or image-driven biases requires not just algorithmic adjustments but a deep understanding of the complex, evolving nature of societal norms and values. Sometimes, algorithmic adjustments can conflict with historical accuracy. For example, Google's Gemini produced images of dark-skinned and racially diverse US Founding Fathers, Nazi soldiers, and the Pope (all historically white persons).[37] To sum up, genAI can lead to two categories of discriminatory effects: demeaning and abusive content and inadequate representation.

---

[32] See also below, Section on Hard cases of generative harm: These phenomena are not generally new in society, but exacerbated or even novel vis-à-vis traditional AI systems.

[33] See, e.g., Emilio Ferrara, 'Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies' (2023) 6 Sci 3, 3-4; Zehlike M, Loosley A, Jonsson H and Wiedemann E and Hacker, Philipp, Beyond Incompatibility. Trade-Offs between Mutually Exclusive Algorithmic Fairness Criteria in Machine Learning and Law (2022). Available at SSRN: https://ssrn.com/abstract=4279866 or http://dx.doi.org/10.2139/ssrn.4279866

[34] Toon Calders and Indrė Žliobaitė, 'Why unbiased computational processes can lead to discriminative decision procedures', in Bart Custers, Toon Calders, Bart Schermer, Tal Zarsky (eds), Discrimination and Privacy in the Information Society: Data mining and profiling in large databases (Springer 2013) 43.

[35]See, e.g., Emilio Ferrara, 'Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies' (2023) 6 Sci 3, 3-4; Leonardo Nicoletti and Dina Bass, 'Humans Are Biased: Generative AI Is Even Worse', Bloomberg Technology + Equality, 23 June 2023, https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

[36] See, e.g., Buddemeyer A, Walker E and Alikhani M, 'Words of wisdom: Representational harms in learning from AI communication' (2021) arXiv preprint arXiv:211108581.

[37] Adi Robertson, 'Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis', The Verge (Feb 22, 2024), https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical.

III.   Legal analysis under EU non-discrimination law

Non-discrimination law was designed to protect specific persons or groups in certain economic or social fields of public interest. It is difficult to apply existing non-discrimination rules to outputs of genAI, even though practical efforts are underway and crucial to deploy such models in a compliant way.[38] But they need to cater to the specificities of each jurisdiction, presenting challenges to scaling these techniques across jurisdictions. Below (1), we provide a brief and general introduction to non-discrimination law. Next (2), we discuss specific challenges that genAI poses to non-discrimination law. Then (3) we seek to identify the actor(s) responsible and liable for discriminatory genAI output.

1.   Short introduction to non-discrimination law

Below we introduce two fields of law regarding discrimination and hate speech, starting with non-discrimination law. Because of space constraints, we can only give a high-level introduction.[39]

a.   Non-discrimination law

The right to non-discrimination is included in many international treaties. For instance, the International Convention on the Elimination of All Forms of Racial Discrimination (1965) is ratified by 182 countries worldwide,[40] and the Convention on the Elimination of All Forms of Discrimination Against Women (1979) by 189 countries.[41] Both discrimination and hate speech are banned in the International Covenant on Civil and Political Rights (1966, 173 ratifications).[42] Discrimination is also banned by the European Convention on Human Rights (1950)[43] and the Charter of Fundamental Rights of the European Union (2000).[44] In sum, there is nearly global consensus that discrimination of protected groups is not acceptable (at least on paper, the consensus is there).

Human rights treaties are often phrased rather abstractly. Moreover, human rights treaties mostly protect people against the state: vertical relations. Such treaties are typically less relevant in horizontal relations: relations between people (or companies). In practice, other legal non-discrimination rules provide more details than the treaties.

---

[38] Barclay Blair, Karley Buckley, Ashley Allen Carr, Coran Darling, Zev Eigen, Danny Tobey, Sam Tyner-Monroe, 'Legal red teaming: A systematic approach to assessing legal risk of generative AI models' DLA Piper White Paper (2024).

[39] See also Frederik Zuiderveen Borgesius and others, 'Non-discrimination law in Europe: a primer for non-lawyers' (2024) arXiv preprint arXiv:240408519.

[40] https://indicators.ohchr.org/

[41] https://indicators.ohchr.org/

[42] Article 20 & 26. International Covenant on Civil and Political Rights

[43] European Convention on Human Rights (1950) Article 14: non-discrimination. See also Protocol 12

[44] Charter of Fundamental Rights of the European Union (2000): article 21 (discrimination), article 23 (equality between men and women).

The EU, for example, has adopted several non-discrimination directives that EU member states must implement in their national laws.[45] The directives prohibit discrimination based on the following protected grounds: gender, age, ethnicity, religion or belief, disability and sexual orientation. EU non-discrimination law bans both direct and indirect discrimination.

In the case of direct discrimination, an organization makes a direct distinction on the basis of, for example, ethnicity. Direct discrimination is always prohibited, except for some narrowly defined specific legal exceptions.[46] An example of prohibited direct discrimination is when a company publicly says that it will not hire people with certain ethnicities. The Court of Justice of the European Union (CJEU) confirmed that such public statements are a form of direct discrimination.[47]

Indirect discrimination is a more complicated concept. Roughly speaking, indirect discrimination happens if an organization's practice is neutral at first glance, but ends up harming people with a protected characteristic, such as ethnicity.[48] For example, suppose that a German company advertises a job and requires candidates to write flawless German, when the job does not necessarily require it. If the requirement harms predominantly people of a certain ethnicity (because they are not native speakers), the practice is probably indirectly discriminating.

However, the law includes a nuanced and somewhat complicated exception. Prima facie indirect discrimination is not a form of indirect discrimination (and thus not prohibited), if the organization can rely on an 'objective justification'.[49] If the organization has a legitimate aim and the neutral practice is a proportional way of trying to achieve that aim, the practice is not prohibited. A German law firm who wants to recruit new lawyers could, for example, make a high-level of German language proficiency a key job requirement on the basis that writing official documents in precise language is an important part of the job. The success of this justification would ultimately be a matter for the courts. For both direct and indirect discrimination, it does not matter whether the organization or their employees realize that they discriminate: intent is irrelevant.[50]

---

[45] See, more generally: Frederik Zuiderveen Borgesius. 2020. Price discrimination, algorithmic decision-making, and European non- discrimination law. European Business Law Review 31, 3 (2020).

[46] The EU Racial Equality Directive defines direct discrimination as follows: 'direct discrimination shall be taken to occur where one person is treated less favourably than another is, has been or would be treated in a comparable situation on grounds of racial or ethnic origin', article 2(2)(a).

[47] CJEU, Judgment of 10 July 2008, Case C-54/07, Centrum voor gelijkheid van kansen en voor racismebestrijding v. Firma Feryn. See section 2.6.1 below for details.

[48] The EU Racial Equality Directive defines indirect discrimination as follows: '(b) indirect discrimination shall be taken to occur where an apparently neutral provision, criterion or practice would put persons of a racial or ethnic origin at a particular disadvantage compared with other persons, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary', article 2(2)(b).

[49] Article 2(2)(b), EU Racial Equality Directive.

[50] Evelyn Ellis and Philippa Watson, *EU anti-discrimination law* (OUP Oxford 2012).

b. Hate speech

A 2008 decision of the EU requires Member States to criminalize hate speech.[51] The decision describes hate speech as 'publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin'.[52] Some EU Member States also banned hate speech regarding other characteristics.

Banning certain types of speech interferes with the right to freedom of expression.[53] The right to freedom of expression is protected, for instance, in Article 10 of the European Convention of Human Rights: 'Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers (...)'.[54] But the right to freedom of expression is not absolute. Under strict conditions, the Convention allows states to limit freedom of expression, for instance 'for the protection of the reputation or rights of others'.[55]

Case law of the European Court of Human Rights shows that the right to freedom of expression does not protect hate speech. In other words, in the case of hate speech, states can legally limit freedom of expression, if states follow the conditions set out in the Convention and the related case law.[56] A specific type of hate speech is holocaust denial. Some European countries specifically ban it; others do not.[57]

To sum up: both discrimination and hate speech are banned in Europe. Both concepts are hard to define, though. And some forms of differential treatment or impact can be justified. In the US, the situation is similar; however, the weight attached to freedom of speech, and particularly also commercial speech, is typically much higher than in the EU.[58] Hence, the set of speech acts that are banned is smaller in the US than in many European states.

---

[51] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law

[52] Article 1(a), Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law.

[53] Cherian George, 'Hate speech law and policy', The International Encyclopedia of Digital Communication and Society (2015)

[54] Article 10(1) of the European Convention of Human Rights.

[55] Article 10(2) of the European Convention of Human Rights.

[56] Bayer, J., & Bard, P., 'Hate speech and hate crime in the EU and the evaluation of online content regulation approaches', report for the European Parliament's LIBE committee, 2020, section 2.3.

[57] Laurent Pech, 'The Law of Holocaust Denial in Europe', in Ludovic Hennebel, and Thomas Hochmann (eds), Genocide Denials and the Law (2011).

[58] See, e.g., Erik Bleich, 'Freedom of expression versus racist hate speech: Explaining differences between high court regulations in the USA and Europe' in Marcel Maussen and Ralph Grillo (eds), Regulation of Speech in Multicultural Societies (Routledge 2015), 110; Uladzislau Belavusau, 'Hate speech', Max Planck Encyclopedia of Comparative Constitutional Law (Oxford University Press 2017); Mathias Hong, 'Regulating hate speech and disinformation online while protecting freedom of speech as an equal and positive right–comparing Germany, Europe and the United States' (2022) 14 Journal of Media Law 76.

2. Non-discrimination law and Generative AI: challenges and difficulties

Non-discrimination law can be applied to many AI-related situations.[59] Suppose that a company uses an AI system to select the best candidates from hundreds of job applicants. The AI system turns out to discriminate against people with an immigrant background, but the company did not realize that. Nevertheless, the company is responsible, even if the company can prove that the discrimination happened accidentally (see in detail below, in the section Responsibility and Liability). Hence, the good news is that non-discrimination law is phrased in such technology-neutral terms that it can generally be applied to situations in which genAI plays a role.[60] There is also bad news, however: non-discrimination law and policy run into several conceptual and technical problems when we try to apply it to genAI, as we shall discuss in the following sections. There is hardly any case law about discrimination and AI,[61] let alone about

---

[59] See, more generally, for EU law: Hilde Weerts and others, 'Unlawful Proxy Discrimination: A Framework for Challenging Inherently Discriminatory Algorithms' (2024) arXiv preprint arXiv:240414050.; 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 Computer Law & Security Review 105770; Jeremias Adams-Prassl, Reuben Binns and Aislinn Kelly-Lyth, 'Directly Discriminatory Algorithms' (2022) The Modern Law Review; Sandra Wachter, 'The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law' (2022) 97 Tul L Rev 149; Meike Zehlike and others, 'Beyond incompatibility: Trade-offs between mutually exclusive algorithmic fairness criteria in machine learning and law' (2022) Working Paper, https://arxivorg/abs/221200469; Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) 41 Computer Law & Security Review 105567; Xavier Ferrer and others, 'Bias and Discrimination in AI: a cross-disciplinary perspective' (2021) 40 IEEE Technology and Society Magazine 72; Holly Hoch and others, 'Discrimination for the Sake of Fairness: Fairness by Design and Its Legal Framework' (2021) Available at SSRN 3773766; Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law' (2020) 123 W Va L Rev 735; Meike Zehlike and others, 'Matching code and law: achieving algorithmic fairness with optimal transport' (2020) 34 Data Mining and Knowledge Discovery 163; Frederik J Zuiderveen Borgesius, 'Strengthening legal protection against discrimination by algorithms and artificial intelligence' (2020) 24 The International Journal of Human Rights 1572; Philipp Hacker, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143; Michael Veale and Reuben Binns, 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data' (2017) 4 Big Data & Society 2053951717743530; for US law: Solon Barocas and Andrew Selbst, 'Big data's disparate impact' (2016) California Law Review 671; Pauline Kim, 'Data-driven discrimination at work' (2016) 58 Wm & Mary L Rev 857; for genAI, see Natali Helberger and Nicholas Diakopoulos, 'ChatGPT and the AI Act' (2023) 12 Internet Policy Review; Philipp Hacker, Andreas Engel and Marco Mauer, 'Regulating ChatGPT and other Large Generative AI Models' (2023) ACM Conference on Fairness, Accountability, and Transparency (FAccT '23) 1112, 1117.

[60] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) 41 Computer Law & Security Review 105567.

[61] But see, on automated decision-making systems, (French) Conseil Constitutionnel, Décision n° 2020-834 QPC du 3 avril 2020, Parcoursup; (Dutch) Rechtbank Den Haag, Case C-09-550982-HA ZA 18-388, SyrRI, ECLI:NL:RBDHA:2020:1878; (Italian) Corte Suprema di Cassazione, Judgment of 25 May 2021, Case 14381/2021; District Court of Amsterdam, Case C /13/689705/HA RK 20-258, Ola, ECLI:NL:RBAMS:2021:1019 [Ola Judgment]; see also Raphaël Gellert, Marvin van Bekkum and Frederik Zuiderveen Borgesius, 'The Ola & Uber judgments: for the first time a court recognises a GDPR right to an explanation for algorithmic decision-making' (EU Law Analysis, 28 April 2021) <https://eulawanalysis.blogspot.com/2021/04/the-ola-uber-judgments-forfirst-time.html/>; Jakob Turner, 'Amsterdam Court Upholds Appeal in Algorithmic Decision-Making Test Case: Drivers v Uber and Ola' Fountain Court Blog (June 4, 2023), https://www.fountaincourt.co.uk/2023/04/amsterdam-court-upholds-appeal-inalgorithmic-decision-making-test-case-drivers-v-uber-and-ola/; Tribunale Ordinario di Bologna, Case N. R.G. 2949/2019, Judgment of 31/12/2020 (Deliveroo); R. (on the application of Motherhood Plan) v HM Treasury [2021] EWHC 309 (Admin) (17 February 2021).

discrimination and genAI.[62] Therefore, this section is explorative, and does not focus on analyzing or systemizing case law.

a. Applicability

To cover genAI scenarios, non-discrimination law would first have to *apply* to these situations. Several complications arise when applying non-discrimination law to genAI. For example, many non-discrimination statutes have a narrow scope, and focus on certain sectors only. To illustrate, some EU non-discrimination directives only apply to employment cases, including recruitment, but not to other contracts or exchanges.[63] EU Member States had to implement the directives into national law; in doing so, some Member States have extended the sectoral scope of the EU rules.[64]

Nevertheless, some situations in which genAI provides discriminatory output may not be covered by non-discrimination law. For example, imagine a large language model fine-tuned by a law firm and used for generating individually negotiated (non-employment) contracts. EU non-discrimination law, in market exchanges, focuses on goods and services offered to the general public.[65] Hence, specific models used by a select group of parties only may fall between the cracks if their output does not constitute a publicly available service or good (and is not in the domain of employment, either, which is generally covered by EU - and US - non-discrimination law). Furthermore, only certain groups are protected by non-discrimination law (religion, ethnicity, gender etc.), both in the EU and the US. However, AI may unfairly disfavor artificially created groups that do not match these traditional categories.[66]

b. Discrimination

Non-discrimination law delineates various forms of harmful actions that can lead to legal consequences. For genAI output to be actionable under non-discrimination law, it must either constitute direct or indirect discrimination against individuals or groups within protected categories, or amount to harassment.

---

[62] For current cases, see, e.g., the case tracker under https://www.bakerlaw.com/services/artificial-intelligence-ai/case-tracker-artificial-intelligence-copyrights-and-class-actions/.

[63] Philipp Hacker, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143.

[64] This is the case, for example, in Germany.

[65] Stefan Grundmann, 'The Future of Contract Law' (2011) ERCL 490, 506; Philipp Hacker P, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143, 1156 f.

[66] Janneke Gerards and Frederik Zuiderveen Borgesius , 'Protected grounds and the system of non-discrimination law in the context of algorithmic decision-making and artificial intelligence' (2022) 20 Colo Tech LJ 1; Sandra Wachter, 'Affinity profiling and discrimination by association in online behavioral advertising' (2020) 35 Berkeley Tech LJ 367, Wachter, S. (2022). The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law. Tul. L. Rev., 97, 149.

- Direct discrimination occurs when an individual is treated less favorably than another in a comparable situation based on a protected attribute.[67] The definition underscores the necessity of *unfavorable treatment* based on characteristics such as ethnicity, gender, disability, or age. The disadvantage may be of a material or immaterial nature.[68]
- Indirect discrimination refers to situations where a seemingly neutral policy, criterion, or practice (PCP) places members of a protected group at a particular disadvantage compared to others.[69] Thus, the essence of indirect discrimination lies in the result, often statistical, burden imposed on protected individuals or groups. What is required, here, is a *disadvantage*, i.e., an adverse effect on an individual or group resulting in legally recognized harm.[70]
- Harassment, finally, constitutes actionable discrimination 'when an unwanted conduct related to racial or ethnic origin takes place with the purpose or effect of violating the dignity of a person and of creating an intimidating, hostile, degrading, humiliating or offensive *environment*. In this context, the concept of harassment may be defined in accordance with the national laws and practice of the Member States.'[71] Harassment is similarly defined in US law.[72]

One of the challenges with applying non-discrimination laws, and the definitions just mentioned, to genAI is aligning the communicative outputs of AI, such as speech acts, images, or videos, with traditional concepts of direct or indirect discrimination that typically focus on more tangible decisions or *actions* that differentiate among individuals. Harassment, in turn, while not requiring intent, must still meet the criteria of creating an *adverse environment*. Hence, it is complicated to apply non-discrimination law to AI-generated content.

In the realm of more traditional AI-driven discrimination (non-genAI discrimination), most real-world examples can, from a legal perspective, be seen as indirect discrimination, characterized by statistical disadvantages for specific groups.[73] Some examples of more traditional AI-driven discrimination could also be qualified as direct discrimination, though.[74]

---

[67] Article 2(2)(a) Employment Equality Directive 2000/78/EC.

[68] District Court (LG) Frankfurt a. M., Judgment of August 26, 2021, Case 2-30 O 154/20, para. 17; BeckOGK-Mörsdorf, § 3 AGG, Rn. 27.

[69] Article 2(2)(b) Employment Equality Directive 2000/78/EC.

[70] Ellis and Watson, EU anti-discrimination law

[71] Art 2(3) of the Race Equality Directive [emphasis by authors].

[72] See, e.g., for a comparative perspective, Alessandro Fabris and others, 'Fairness and Bias in Algorithmic Hiring' (2023) arXiv preprint arXiv:230913933, 32-35; Gabrielle Friedman and James Q. Whitman, 'The European transformation of harassment law: discrimination versus dignity' (2002) 9 Colum J Eur L 241; Joanna Lahey, 'International comparison of age discrimination laws' (2010) 32 Research on Aging 679; see also, on key US concepts, Solon Barocas and Andrew Selbst, 'Big data's disparate impact' (2016) California Law Review 671; Pauline Kim, 'Data-driven discrimination at work' (2016) 58 Wm & Mary L Rev 857.

[73] See, e.g., Philipp Hacker, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143, 1151 ff.; Frederik Zuiderveen Borgesius, 'Discrimination, artificial intelligence, and algorithmic decision-making' (2018) 19; Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) 41 Computer Law & Security Review 105567.

[74] Sandra Wachter, 'Affinity profiling and discrimination by association in online behavioral advertising' (2020) 35 Berkeley Tech LJ 367; Jeremias Adams-Prassl, Reuben Binns, and Aislinn Kelly-Lyth, 'Directly discriminatory

Contrastingly, in the context of genAI discrimination, harassment emerges as a significant category, in addition to direct discrimination. This distinction underscores that, from a doctrinal perspective, genAI introduces a distinct and partially novel phenomenon in the landscape of AI discrimination law.

How can problematic output by genAI then be mapped onto the existing categories of non-discrimination law? We need to distinguish different sets of cases. In our view, the most typical cases fall into three groups: (i) discriminatory content; (ii) harassment; and (iii) hard cases of generative harm that do not easily fit under the existing structures and concepts of non-discrimination law.

First, discriminatory content captures output that leads to disadvantageous *decisions*, i.e., acts or omissions disfavoring protected groups or their members that are sanctioned under non-discrimination law.[75] This might, for example, be the rejection of a job candidate or a credit application.

Second, harassment covers specific forms of content that deny the intrinsic value or worth of persons in ways that create toxic environments. Hence, it is not a concrete material harm based on an act but rather the level of toxicity concerning a speech or communicative act itself that is the most important element .[76]

Third, beyond these more traditional categories, hard cases of generative harm present intricate problems for theoretical and doctrinal analysis. These hard cases can be organized into three principal areas: inadequate representation, harmful stereotypes, and misclassification. While inadequate representation obviously matches the respective descriptive category, harmful stereotypes and misclassification are instances of demeaning and abusive content (see Table 1, above; and Figures 1 and 2, below). Overall, this leads to the following classification:

I. **Content falling under direct or indirect discrimination:** Harm is inflicted by way of an act, triggered by AI output.

II. **Content falling under harassment:** The AI-based communicative act violates personal dignity and creates an adverse environment.

III. **Hard cases of discrimination by genAI:** Some types of outputs are more difficult to square with established concepts of direct, indirect discrimination and harassment. This gives rise to three main sub-categories, which contain elements both from the 'demeaning and abusive content' and the 'inadequate representation' descriptive categories (see Table 1, above):

---

algorithms' (2023) 86 The Modern Law Review 144; Hilde Weerts and others, 'Unlawful Proxy Discrimination: A Framework for Challenging Inherently Discriminatory Algorithms' (2024) arXiv preprint arXiv:240414050.

[75] Ellis and Watson, EU anti-discrimination law

[76] See also Amelie Berz, Andreas Engel and Philipp Hacker, 'Generative KI, Datenschutz, Hassrede und Desinformation – Zur Regulierung von KI-Meinungen' (2023) Zeitschrift für Urheber- und Medienrecht 586

1. **Inadequate Representation:** This category concerns representational harms; genAI shows a biased vision of the world. The category encompasses two subcategories:
   - **Unbalanced Content:** Here, AI-generated outputs exhibit a bias toward certain demographics and protected groups, such as predominantly displaying white male CEOs in response to generic requests for images of CEOs. This subcategory highlights the AI's tendency to mirror and magnify societal biases in representation, such as language or imagery.
   - **Non-Inclusive Language:** Instances where the AI's use of language, including the generic masculine form, fails to recognize or represent the full spectrum of genders or other identities.
2. **Harmful Stereotypes:** This category, part of 'demeaning and abusive content,' involves AI systems inadvertently endorsing or propagating negative stereotypes, for example overt historical racist depictions of ethnic groups or harmful generalizations such as 'men are more boring than women.' Such AI outputs can reinforce and spread harmful societal biases.[77]
3. **Misclassification:** Among the most direct forms of AI-facilitated harm is the misclassification of individuals in which individuals are wrongly labeled as part of a certain group, also part of 'demeaning and abusive content,'. This includes ways that reflect deep-seated prejudices. An egregious example includes the misidentification of Black individuals as gorillas by Google's facial recognition system in 2017, a manifestation of unequal output.[78] This category of harm occurs both in traditional AI and more recent generative classification systems (i.e., systems based on genAI that classify content, e.g., by analyzing images or text).

We shall take up these cases in turn, starting with discriminatory content and harassment before addressing the more nuanced, hard, and partially novel types of generative harm.

i. Discriminatory content

The first test under non-discrimination law is typically whether a certain type of behavior or harm falls under direct or indirect discrimination.[79] This is no different with genAI output. Within the group of discriminatory content, we start with negative speech acts before turning to hate speech proper.

1. Negative speech acts

To analyze the potential harm caused by negative speech acts generated by AI, consider, for example, the potential AI-generated statement 'xyz persons should not be hired [xyz being a protected attribute],' inspired by the homophobic statements that were under dispute in the CJEU LGBTI case.[80] Such output could be generated when users play with a GPT or converse

---

[77] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Do Large Language Models Have a Legal Duty to Tell the Truth?' [2024] Royal Society Open Science.

[78] Maggie Zhang, ‚Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software' (forbes.com 4 April 2024) https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/.

[79] Cf. Information Commissioner's Office, Big data, artificial intelligence, machine learning and data protection, Version 2.2., 2017, para. 94-96.

[80] See CJEU, Case C-507/18, *Associazione Avvocatura per i diritti LGBTI*.

with a chatbot on an e-commerce website. In fact, such output is not entirely hypothetical. In a recent UNESCO study, pervasive gender, racial, and sexual orientation stereotyping was found in large language models.[81] For example, the prompt 'a gay person is…', was followed with a negative completion by Llama 2 (an open-source model by Meta) in 70% of the cases.[82] Examples included statements such as the following: 'The gay person was regarded as the lowest in the social hierarchy.'[83] By contrast, ChatGPT (GPT 3.5), which has been consistently trained using reinforcement learning from human feedback to mitigate biases, gave positive or neutral answers in 80% of the cases.[84] This underscores again the importance of implementing safeguards *at the model level.*

To analyze AI-generated sentences such as recommendations not to hire gay persons, we may turn to legal precedents, such as the *Feryn*[85] case and the *Associazione Avvocatura per i diritti LGBTI*[86] case as adjudicated by the Court of Justice of the European Union (CJEU).[87] These cases provide a framework for understanding when speech acts might fall under the purview of non-discrimination laws, and be considered direct discrimination.

Direct discrimination

The cases established that statements indicating an intention not to hire candidates based on racial or ethnic origin (*Feryn*) or sexual orientation (*LGBTI*), made publicly, can be considered as falling under the 'conditions for access to employment'[88] Hence, such statements can constitute direct discrimination. The CJEU broadened the interpretation of direct discrimination to include not just active recruitment processes but also statements that could be tied to an employer's recruitment policy in a significant manner, even if no recruitment process was underway at the time of the statement. The Feryn/LGBTI judgments lead to a three-step test. A statement can count as direct discrimination if three conditions apply.

(i) The statement must, first, be related to the employer's recruitment policy in an actual, not merely hypothetical, way.[89] A comprehensive evaluation must be undertaken to assess whether this actual link exists.[90]

---

[81] UNESCO and IRCAI (2024). 'Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models'.

[82] UNESCO and IRCAI (2024). 'Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models', 10. The version used was the July 2023 version of Llama 2, apparently.

[83] UNESCO and IRCAI (2024). 'Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models', 10.

[84] UNESCO and IRCAI (2024). 'Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models', 10.

[85] CJEU, Case C-54/07 *Feryn*.

[86] CJEU, Case C-507/18, *Associazione Avvocatura per i diritti LGBTI.*

[87] See, more generally, Hacker, 'A legal framework for AI training data—from first principles to the Artificial Intelligence Act' (2021) 13 Law, Innovation and Technology 257, 272 et seqq.

[88] See CJEU, Case C-54/07 *Feryn*; Case C-507/18, *Associazione Avvocatura per i diritti LGBTI*; see also CJEU, Case C-81/12, *Asociaţia Accept.*

[89] CJEU, Case C-507/18, *Associazione Avvocatura per i diritti LGBTI*, para. 43.

[90] CJEU, Case C-507/18, *Associazione Avvocatura per i diritti LGBTI*, para. 43.

(ii) The second criterion is that the statement must exert a decisive influence on activities protected under anti-discrimination law, such as hiring practices. This influence, at a minimum, must be recognized by the affected social (protected) groups.[91]

(iii) The third criterion is the public nature of the statement.[92] In the Feryn case, the director of the eponymous company publicly stated that they were looking to hire new personnel, but would not recruit 'immigrants'. Similarly, in LGBTI, the company owner stated in a radio program that he would not give a job to 'homosexual persons'. The CJEU emphasized the deterrent effect of public announcements on potential applicants as a significant reason for applying anti-discrimination laws. This teleological approach underlines the importance of the public visibility and potential impact of discriminatory statements on the willingness of individuals from protected groups to engage in certain activities, like applying for jobs.

However, applying these principles to AI-generated statements poses unique challenges. For example, a hypothetical statement made by a genAI model equivalent to the company owner's radio statement in the LGBTI case, asserting that 'Homosexual persons should not be hired,' does not straightforwardly meet the criteria set by the Feryn/LGBTI test for several reasons.

Starting with the easiest criterion - (iii) publicity - AI-generated statements directed at a single individual via a chatbot may not equate to a public declaration, such as one made on a radio program. However, if the recipient is a potential job candidate, even a privately received AI-generated statement could deter that individual from applying with the developing or deploying company, mimicking the deterring effect of a public announcement for the affected individual. The person may infer that this AI statement is, actually, made vis-à-vis other persons interacting with the same model, too – which will likely be correct.[93] This constitutes what one might call 'technology-mediated publicity by iteration'. As seen, the perception of the statement's publicity and its reach among potential candidates or protected groups plays a role in assessing its impact under non-discrimination law.

However, the other two elements of the Feryn/LGBTI test are harder to fulfill. Concerning the first criterion of the Feryn/LGBTI test: is the statement related to an employer's recruitment policy? An offensive statement, when generated by AI, is typically abstract and not tied to any specific or impending job application process. It is also difficult to argue that the AI's output reflects the hiring policies or intentions of the company developing or deploying the AI model, unless the system in question was a genAI application built specifically for the purposes of recruitment or trained on the company's prior hiring data and policies.

The next criterion is: does the statement have a decisive influence on activities that fall within the scope of non-discrimination law? It becomes challenging to assert that AI-generated output has a decisive influence – even if only perceived – on hiring decisions, unless the AI in question

---

[91] Cf. CJEU, Case C-507/18, Associazione Avvocatura per i diritti LGBTI, para. 43.

[92] Cf. CJEU, Case C-507/18, Associazione Avvocatura per i diritti LGBTI, para. 43.

[93] Due to the probabilistic nature of the generative model, that will be some variance concerning the output. However, it may safely be assumed that, over time, negative statements will be repeated since they do fall within the output space of the model if they have already been made.

is explicitly used as a recruitment tool. Otherwise, the link between AI output and actual hiring decisions might be too tenuous to meet the criterion of having a decisive influence on an actual decision, especially if the AI's statements cannot be clearly linked to the employer's recruitment policies or practices. Assuming that the output of the genAI system in question is not positioned by the developing or deploying company as representing an official company position, such statements would likely not pass the Feryn/LGBTI test for direct discrimination.

Thus, since discrimination under these legal precedents requires a connection to actual or planned employment practices. In this hiring example , AI-generated content would rarely constitute direct discrimination. The issue, then, centers not only on the content of the speech but also on its contextual relevance to employment practices and policies, which, in the case of AI-generated speech, is generally absent or too tenuous to establish direct discrimination as defined by current EU legal frameworks.

The difficulty seen here in linking AI-generated speech to direct discrimination in hiring may not be generalisable to other contexts, though. Direct discrimination may be more provable in other use cases where, for example, genAI systems make concrete recommendations to buy, book or otherwise contract that show biased patterns. In such situations, the link to the consumer decisions may be stronger and more easily proven. Transplanting the Feryn/LGBTI criteria to these cases may indeed result in direct discrimination being found.

Indirect discrimination

Statements such as the hypothetical 'do not hire xyz persons' [xyz being a protected attribute] do not constitute indirect discrimination, either, as they are not apparently neutral.[94] Rather, they are straightforwardly demeaning towards a certain protected group. Yet, due to the specific link that EU non-discrimination law requires to specific practices (e.g., hiring practices)[95], such statements are not directly actionable under the Feryn/LGBTI doctrine.

One could, however, argue that the generative AI, on the system-level, constitutes a neutral policy, criterion or practice,[96] which places members of a protected group at a particular disadvantage. The neutrality may be said to follow from the fact that the genAI system does not *generally* result in direct discrimination, but only in some specific circumstances. Similarly, the installation of a bouncer regime to be in charge of the door to a nightclub could be generally neutral, even though some bouncers occasionally engage in directly discriminatory practices.

---

[94] Sandra Fredman, *Discrimination Law*, 3rd. Ed., Oxford Univ. Press, 2023.

[95] Evely Ellis and Philippa Watson, *EU Anti-Discrimination law*, Oxford Univ. Press, 2012.

[96] See, e.g. Philipp Hacker, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143, 115; Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) 41 Computer Law & Security Review 105567.3; but see also, for arguments in favor of direct discrimination in many of these cases, Sandra Wachter, 'Affinity profiling and discrimination by association in online behavioral advertising' (2020) 35 Berkeley Tech LJ 367.
 Jeremias Adams-Prassl, Reuben Binns, and Aislinn Kelly-Lyth, 'Directly discriminatory algorithms' (2023) 86 The Modern Law Review 144, 157 et seqq.

This would open the entire genAI system to scrutiny concerning its overall effect on protected groups.

Still, however, a specific disadvantage with respect to a legally recognized category of harm, within the scope of non-discrimination law, would be required. Hence, ultimately, a similar issue arises as in the above discussion of the Feryn/LGBTI doctrine. Nonetheless, the use of genAI may, of course, result in indirect discrimination just like non-genAI systems - for example, if harnessed to sort and evaluate job applications, with a statistically relevant, unjustified disadvantage for some protected group in the output.[97]

2.   Hate speech

Beyond merely negative speech or imagery, genAI models might provide content that falls within the definition of hate speech.[98] Such output might amount to direct discrimination under the Feryn/LGBTI test if it is related to an actual scenario of a hiring or selection decision within the ambit of non-discrimination law.

ii.   Harassment

A second doctrinal category for non-discrimination law, besides direct and indirect discrimination, is harassment. Even if certain AI output, such as negative speech acts, do not qualify as direct or indirect discrimination, the output may still constitute harassment. We first analyze negative speech acts under the harassment doctrine before turning to hate speech proper.

3.   Negative speech acts

As seen, as long as negative speech acts do not directly relate to an act that would be covered under non-discrimination law, they do not, generally, constitute direct (nor indirect) discrimination (Feryn/LGBTI test). Hence, the best contender for making such statements legally relevant under non-discrimination law is, arguably, harassment. For ease of exposition, we quote the definition of harassment from the Employment Equality Directive:

> Harassment shall be deemed to be a form of discrimination within the meaning of [this directive], when unwanted conduct related to any of the grounds referred to in [in this directive] takes place with the purpose or effect of violating the dignity of a person and of creating an intimidating, hostile, degrading, humiliating or offensive environment. (...)[99]

Under this definition, for the AI-generated statement 'xyz persons should not be hired' [xyz being a protected attribute] to constitute harassment, it must meet four criteria: (i) the conduct must be unwanted; (ii) related to a protected attribute (in this case, sexual orientation); (iii) have the purpose or effect of violating an individual's dignity; and (iv) create an adverse environment

---

[97] See, e.g., Philipp Hacker, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143, 1160 et seqq.

[98] For a definition and contextualisation of hate speech see: Cherian George, 'Hate speech law and policy' in The international encyclopedia of digital communication and society (John Wiley & Sons 2015).

[99] Article 2(3), Council Directive 2000/78/EC Establishing a General Framework for Equal Treatment in Employment and Occupation, 2000 OJ L 303/16.

characterized by intimidation, hostility, degradation, humiliation, or offensiveness.[100] We discuss them in turn.

1. Unwanted Conduct: As an AI-generated statement, its unwanted nature would be determined by the context of its reception. If disseminated in a workplace or public forum, or even in a private chat but conversation, it could easily be deemed unwanted by those it targets or affects.

2. Protected Attribute: The statement directly targets a protected attribute (e.g., sexual orientation), fulfilling this criterion.

3. Violating Dignity: The statement intrinsically degrades individuals based on their sexual orientation, a core aspect of personal identity, thereby violating their dignity. This aspect seems straightforward, as the denigrating nature of the statement towards xyz individuals may impact their sense of self-worth and respect.

4. Creation of an Adverse Environment: The statement's potential to create an adverse environment depends, arguably, on its dissemination and the context. Case law suggests that the act or statement must characterize the environment and amount to more than just a nuisance.[101] Typically, however, the hostile environment presupposes multiple acts of adverse behavior.[102] For example, if made within a workplace environment, hiring platform, or publicly by a company's AI system, the statement could significantly contribute to a setting that is hostile or offensive to xyz individuals. The environment becomes adverse not just through the presence of such statements but through the (apparent) legitimization of discriminatory sentiments they might foster within a community or organization.

A private chatbot conversation between a single user and a genAI-powered chatbot, however, constitutes a borderline case. In our view, for the statement made by a genAI system to fulfill the criteria of harassment, it would need to be shown that its dissemination actively contributes to an environment, in a context covered by EU non-discrimination law (e.g., employment; offering of publicly available services), that is not only unwelcoming but also significantly and persistently impacts the well-being and dignity of protected persons. For a successful case against harassment, the victim would have to demonstrate that the statement's presence is both pervasive and influential enough to shape the relevant atmosphere in a manner that is hostile and demeaning. This would likely require collecting information about outputs to similar queries or use cases to show an environmental effect, or using comparable auditing methods.

There are several challenges in applying harassment doctrine to genAI output. First, the dual requirement of violating dignity and creating an adverse environment raises the bar for establishing harassment, particularly with AI-generated content. Second, the context of the

---

[100] Sandra Fredman, *Discrimination Law*, 3rd. Ed., Oxford Univ. Press, 2023.

[101] German Federal Court for Employment Law (BAG), Judgment of September 24, 2009, Case 8 AZR 705/08 = NZA 2010, 387, para. 32.

[102] See, e.g., German Federal Court for Employment Law (BAG), Judgment of September 24, 2009, Case 8 AZR 705/08 = NZA 2010, 387, para. 32; Thüsing, in: MüKoBGB, 9th ed. 2021, AGG § 3 Rn. 63.

statement's dissemination and reception is crucial. The same statement could have different impacts depending on where and how it is presented. Reflecting non-discrimination law's contextuality, legal analysis must consider the specific circumstances of each case. For example, in the context of an e-commerce website, offending statements may constitute even if only made in a 'private' conversation between a potential customer and the site's bot.

In fact, one could also argue that such statements, even if made in a private AI-facilitated conversation, may create a qualified hostile environment for two reasons. First, as mentioned before, such statements typically do not occur only once and in one conversation only. Rather, the affected party must assume that the genAI system will make the same or similarly demeaning statements vis-à-vis other users. At least, in our view, such a repetition should be presumed by agencies and courts, by way of a prima facie harassment claim,[103] unless the developing and deploying entity proves the opposite, showing that the specific sentence was indeed an absolute outlier that cannot have been reproduced in any other similar setting.

Second, in our view, it is not necessary for such statements (such as 'do not hire xyz persons', xyz being a protected attribute) to be directly linked to, for example, hiring procedures even if the statement is about the (non-)employment of a certain protected group. The reason is that non-discrimination law applies not only in the context of employment, but also to the offering of publicly available services. The use of an AI-powered chatbot typically constitutes such a service.[104] Hence, the adverse effect of a statement like 'do not hire xyz persons' occurs in a context which is covered by EU non-discrimination law.

The key difference to standard cases of discrimination resides in the fact that genAI discrimination does not, necessarily, relate to a specific negative decision (rejection of a job candidate or a credit application), but rather makes the person feel uncomfortable within the algorithmic environment. However, to the extent that this environment falls under the ambit of non-discrimination law (e.g., via the provision of publicly available services), it is precisely this creation of a non-inclusive space that is sanctioned by the verdict of harassment once the gravity of the statement reaches a certain threshold – intimidation, hostility, degradation, humiliation, or offensiveness.

A few years ago, the German Federal Court for Employment Law, in a controversial judgment, rejected that threshold to be reached for xenophobic inscriptions on the walls of a toilet in the workplace that were not removed by the employer, noting that these slurs only cover a very limited area of the workplace.[105] While the CJEU might rule otherwise, we argue that the result is different for chatbots, anyways: the linguistic interaction forms the core of the offered service; and xenophobic, homophobic or misogynistic statements typically do not occur only

---

[103] Cf. CJEU, Case C-303/06, *Coleman v Attridge Law*, para. 62.

[104] See Philipp Hacker, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143, 1159-1160.

[105] German Federal Court for Employment Law (BAG), Judgment of September 24, 2009, Case 8 AZR 705/08 = NZA 2010, 387, para. 34.

once, but are part of a more systematic reinforcement of negative stereotypes within the output of an inadequately moderated model.[106]

With this in mind, one may conclude that the hypothetical AI-generated statement 'xyz persons should not be hired' does constitute harassment, as it is offensive, degrading and humiliating to that protected group, and creates a hostile environment - not necessarily in an employment context, but in the (publicly available) usage of the genAI system. Similarly, the output 'Muslims are the enemies of humanity'[107] has a comparable effect on the Muslim population (see the section on demeaning and abusive content); it may, thus, be qualified as harassment, too.

In sum, under the discussed criteria, an AI-generated statement or image may constitute harassment if it clearly violates the dignity of persons belonging to a protected group and contributes to the creation of an adverse environment. To establish whether a genAI-produced statement qualifies as harassment, a judge must examine the statement's impact and the context of its dissemination – both in the immediate and in the presumed iterated sense.

4.   Hate speech

While merely negative speech acts may or may not constitute harassment, depending on the contextual analysis, hate speech proper will, at least generally, cross the threshold to harassment under the criteria just discussed. This legal category can therefore serve as a safety net in cases in which members of protected groups are attacked in communicative acts without an immediate consequence for decisions relevant under the direct and indirect discrimination prongs of equality law.

iii.   Hard cases of generative harm

As we have seen, traditional legal categories such as direct discrimination or harassment cover a range of genAI outputs because they lead to disadvantageous acts or toxic communication. In the following, we now turn to types of AI output that do not squarely fall within these categories because the output does not clearly relate to disadvantageous acts or decisions, nor does the output easily cross a certain threshold of toxicity. We see three main categories of such hard cases of generative harm that we take up in turn: (i) inadequate representation; (ii) harmful stereotypes; and (iii) misclassification. The first sub-category matches the descriptive category of inadequate representation discussed above. The second and third sub-categories, harmful stereotypes and misclassification, are part of the category of demeaning and abusive content in the descriptive sense described (see Section 'Generative AI and risks related to discrimination and hate speech'; see also the mapping between descriptive and legal categories below, Section 'Summary concerning discrimination').

These cases are hard in two ways. First, they do not fit squarely within the traditional legal categories, as mentioned. Second, they also comprise phenomena that are not novel per se in

---

[106] Cf. UNESCO and IRCAI (2024). 'Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models', 10.

[107] https://www.zeit.de/digital/2023-09/aleph-alpha-luminous-jonas-andrulis-generative-ki-rassismus [prompt and answer translated from German by the authors].

society or traditional AI systems, but that are amplified by or manifest differently in generative AI systems (unbalanced content; harmful stereotypes; misclassification; non-inclusive language).

5.   Inadequate representation

Beyond the cases of potentially harmful language or imagery discussed before under direct discrimination and harassment, the second large descriptive category, and group of cases, comprises scenarios of inadequate representation in genAI output. It covers both unbalanced content and non-inclusive language.

(a) Unbalanced content

Unbalanced content may, for example, occur where querying a generative model for 'CEO' results predominantly in images of white males. Consequently, the issue of discrimination through 'lopsided repetition' arises – either within a larger set of examples being simultaneously given or, more often, over time as a certain prompt, used repeatedly, generates answers that are systematically skewed towards inadequate representation. The repetition of AI-generated stereotypes, such as the association of leadership roles primarily with white males, might reinforce societal biases.[108] Unbalanced content has some similarities with the problem of homogenous search results in search engines (e.g., pictures of predominantly white men in a CEO picture search). However, genAI models create the output, while search merely retrieves existing information.

For both settings, difficult questions lurk beneath the surface: what may be considered fully equal representation? There are many, partly conflicting, possible measures for equal representation. We give some examples. (a) Absolute statistical parity, means a uniform distribution of examples between groups (e.g., same number of men, women, and non-binary persons depicted); (b) relative statistical parity, holds when the proportion of individuals belonging to a specific protected group equals the proportion of individuals belonging to this group in (a segment of) society, or even in the specific field queried; (c) one might also aim for some normative distribution corresponding to a desired, but as of yet unattained quota; (d) or for a random selection.

Choosing the most appropriate option in a certain context is difficult. This difficulty is known from computer science discussions about discrimination in (non-generative) AI. Most people agree that AI should be non-discriminatory. But it is difficult to turn legal non-discrimination norms into numerical requirements or 'fairness metrics,' as computer scientists often say. [109]

Due to space constraints, we cannot further explore this debate here. For the sake of analysis, let us assume that the representation given by a repeated prompting of a genAI system does not satisfy any of the above criteria, for instance, because the system mentions almost exclusively white males.

---

[108] Mi Zhou and others, 'Bias in Generative AI' (2024) arXiv preprint arXiv:240302726.

[109] See, e.g., Meike Zehlike, Ke Yang and Julia Stoyanovich, 'Fairness in Ranking, Part I: Score-based Ranking' (2022) ACM Computing Surveys (CSUR) 1; Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law' (2020) 123 W Va L Rev 735.

Direct Discrimination

From a legal perspective, such one-sided representation in genAI outputs could be scrutinized under the lens of direct discrimination if it leads to a significant underrepresentation of protected groups in positive contexts (e.g., leadership roles) or an overrepresentation in negative contexts (e.g., criminal activities). As mentioned, direct discrimination entails a situation where an individual or group is placed at a disadvantage or treated less favorably than others in a similar situation, based on protected attributes.

In the context of genAI, the challenge lies in establishing a palpable disadvantage for the member of a protected group, flowing from the inadequate representation, towards a decision or situation that falls within the ambit of non-discrimination law. The repeated portrayal of CEOs as predominantly white males by a genAI system could be argued to contribute to a systemic bias, potentially influencing hiring practices or societal perceptions in a manner that disadvantages other non-white and non-male groups.

However, the link between the depiction and such palpable, legally recognized disadvantages will often be difficult to demonstrate empirically, and may ultimately be too tenuous. Yet, of course, inadequate representation across groups can also lead to identity-based harms over time independent of considerations of equal opportunity.

Harassment

Regarding harassment, the issue becomes more nuanced. Harassment is, in short, unwanted conduct related to a protected attribute that violates a person's dignity and creates an intimidating, hostile, degrading, humiliating, or offensive environment.[110] While the inadequate representation in AI-generated images or texts might contribute to a broader societal issue of stereotyping and bias, it may not individually reach the threshold required to be considered harassment under the law. Harassment typically involves more direct, personal, and aggressive forms of conduct. Inadequate representation, while problematic, may not directly create the type of hostile or offensive environment that characterizes legal definitions of harassment, at least in most cases.

However, repeated exposure to biased genAI outputs could contribute to a culture that undervalues diversity and reinforces stereotypes, indirectly affecting the dignity and perception of underrepresented groups. While this might not strictly meet legal criteria for harassment, it underscores the importance of addressing biases in genAI (and traditional AI) to prevent perpetuating or exacerbating discrimination.

To sum up, unbalanced content is one of two sub categories of inadequate representation. Unbalanced content is problematic, but it is not always clearly prohibited under current non-discrimination law.

(b) Non-inclusive language

We now turn to the second subcategory of inadequate representation: non-inclusive language. It raises significant legal and societal questions if a genAI system produces language output

---

[110] See e.g. Article 2(3), Council Directive 2000/78/EC Establishing a General Framework for Equal Treatment in Employment and Occupation, 2000 OJ L 303/16.

that is not gender-neutral or otherwise does not conform with the desiderata of greater diversity in language use (e.g., pronouns, male and female versions of nouns and adjectives in certain languages, use of inclusive gender signs like the *). The difficulty lies in squaring non-inclusive language - which itself is a contested concept - with the concepts of discrimination or harassment within existing non-discrimination law.

Existing case law may offer some guidance. In 2018, the German Federal Court for Private Law (Bundesgerichtshof, BGH), the highest German court in private law, addressed language in the Sparkasse case, interpreting EU non-discrimination law.[111] The court determined that the use of the generic masculine grammatical form[112] in official bank documents does not constitute discrimination against women. The court noted that 'there is no legal entitlement not to be addressed in forms and documents with personal designations whose grammatical gender differs from one's own natural sex. According to the generally customary language use and understanding, the meaning of a grammatically male person designation can encompass any natural sex ('generic masculine form').'[113] More specifically, the court found that the female plaintiff did not suffer any disadvantage from being addressed as a customer in the grammatically correct, generic masculine form, due to the linguistic difference of (grammatical) genus and (biological/social) sexus.[114]

However, grammar and language can change over time, particularly if underlying social norms make certain uses of language obsolete, or even offensive. The court relied largely on what it perceived as the still common usage of the generically masculine form in wide parts of the population and in official documents, including the law itself.[115] Hence, a changed linguistic practice could eventually lead to a valid claim of disadvantage and, potentially, discrimination. However, the ruling indicates that within certain linguistic contexts, such as the German one, the use of traditionally gendered language forms without explicit gender inclusivity does not automatically amount to legal discrimination in the sense of the EU non-discrimination law.[116]

In a 2023 case, the Administrative Court of Berlin (Verwaltungsgericht Berlin) ruled on a father's emergency motion against the use of gender-neutral language at his children's high school.[117] The court found no violation of the principle of political neutrality in the education system nor an infringement of parental rights to education. The Court noted that school administrations had explicitly allowed teachers to use gender-neutral language in the classroom while also emphasizing that the rules of orthography must be adhered to in the teaching and learning process. Overall, the court rejected the attempt to outlaw gender-neutral language in the educational system.

---

[111] BGH, Judgment of March 13, 2018, Case VI ZR 143/17 = NJW 2018, 1671.

[112] In German, e.g.: 'Kunde' [customer] instead of the explicitly female version 'Kundin'.

[113] Our translation. BGH, Judgment of March 13, 2018, Case VI ZR 143/17 = NJW 2018, 1671, first para.; critique in Ulrike Lembke, 'Verfassungswidrige Sprachverbote', Verfassungsblatt 2023, 1704.

[114] BGH, Judgment of March 13, 2018, Case VI ZR 143/17 = NJW 2018, 1671, para. 30, 35 et seqq.

[115] BGH, Judgment of March 13, 2018, Case VI ZR 143/17 = NJW 2018, 1671, para. 37-38.

[116] If, however, an online shop does offer a choice between different gendered titles (e.g., Mr. and Ms.), it must also include a choice for non-binary persons: District Court (LG) Frankfurt a. M., Judgment of August 26, 2021, Case 2-30 O 154/20.

[117] VG Berlin, Decision of March 24, 2023 - VG 3 L 24/23.

Some federal states in Germany are, however, discussing laws specifically outlawing gender-neutral language in certain contexts, particularly in educational settings.[118] For example, in 2023, the Minister of Education for the State of Saxony-Anhalt prohibited the use of gender-inclusive signs, such as the *, in school settings.[119] The Bavarian administration followed suit in March 2024, covering schools and public agencies.[120]

The legal cases cited follow a very conventional understanding of language and law, and are confined to Germany; the general issue is not. The CJEU might rule differently one day. Nonetheless, the cases underscore that within the current legal frameworks in Germany, there is considerable leeway regarding the use of more or less inclusive language, and no strict legal requirement enforcing the adoption of one over the other,[121] unless the legislator explicitly intervenes. This should hold for genAI output, too.

The German court rulings indicate that gendered language forms, including the use of the generic masculine, do not necessarily constitute discrimination under the law. Concurrently, the adoption of gender-neutral language forms does not necessarily infringe upon rights, even within educational settings, as demonstrated by the decision of the Administrative Court of Berlin. Together, the court rulings establish a type of balance. To some extent, the balance respects both the tradition of language and the progressive movement towards greater inclusivity, without imposing legal penalties or requirements on the use of gendered or gender-neutral forms. However, again, this question is far from settled, particularly at the EU and global level.

Some legal initiatives may change this balance in the future. In some contexts, counterintuitively, the use of more inclusive language, including in genAI outputs, may then actually violate the law, at least in some EU Member States. Whether such local laws violate national constitutions,[122] or even the European Charter of Fundamental Rights, remains to be seen and transcends the scope of this chapter.

In conclusion, genAI systems may produce non-inclusive language. Even outside the field of AI, inclusive language is a controversial and hotly debated topic. Societal norms, legal norms, and case law about inclusive language are still developing. What is specific about genAI, in this context, is the fact that these models can be *designed* to produce output in various shades of inclusivity; and that a single model can amplify these design choices among millions of users. This will be taken up in the policy section to think more in detail about regulatory options for these very design choices.

---

[118] https://www.mdr.de/nachrichten/thueringen/genderverbot-schule-landtag-cdu-afd-100.html; https://www.faz.net/aktuell/politik/inland/gender-verbot-in-bayern-per-gesetz-gegen-gendergerechte-sprache-19423874.html.

[119] https://www.mdr.de/nachrichten/sachsen-anhalt/landespolitik/gendern-verbot-schulen-104.html.

[120] https://www.sueddeutsche.de/bayern/gendern-sternchen-verbot-sprache-bayern-1.6468805

[121] See also Peter Allgayer, 'Der rechtliche Rahmen des Genderns', NJW 2022, 452.

[122] Highlighting a potential violation of the freedom of research and education: https://www.zdf.de/nachrichten/panorama/thueringen-landtag-gendern-100.html; see also Yannik Breuer and Madeline Trappmann, 'Geschlechtergerechte Sprache im öffentlich-rechtlichen Rundfunk' ZUM 2024, 192; Ulrike Lembke, 'Verfassungswidrige Sprachverbote', Verfassungsblatt 2023, 1704.

6.   Harmful stereotypes

The second category of hard cases of generative harm is harmful stereotypes. These stereotypes become harmful if they trigger discrimination-specific harm, as discussed above, particularly identity- or representation-based ones. When examining harmful stereotypes, such as the output asserting that 'men are more boring than women,' or derogatory images or language regarding welfare recipients, through the lens of direct discrimination and harassment, it becomes clear that the fit is, again, not straightforward. Direct discrimination necessitates a scenario where an individual is treated less favorably than another in a similar situation based on a protected attribute. The stereotype about men, while perpetuating a gender bias, does not directly link to an act of unfavorable treatment in a specific context, such as employment or services. Hence, under the Feryn/LGBTI test, such a statement will usually not be classified as direct discrimination. Such a statement would not be a type of indirect discrimination either because the statement is not apparently neutral.

Similarly, the statement does not meet the criteria for harassment, which involves unwanted conduct that significantly violates a person's dignity or creates an intimidating, hostile, degrading, humiliating, or offensive environment. Although the stereotype is biased and potentially offensive, it will generally not reach the required level of toxicity or of creating a systematically hostile or degrading environment that harassment requires. In sum, harmful stereotypes in genAI outputs, on their own, likely do not violate non-discrimination law.

The case of genAI-produced harmful stereotypes about welfare recipients[123] presents a more complex challenge. These stereotypes can contribute to a negative portrayal and perception of individuals based on socio-economic status, potentially influencing opinions and decisions in areas like employment or social services. However, unless these stereotypes are directly used in a way that results in less favorable treatment of individuals from certain socio-economic backgrounds in comparable situations, those individuals may also struggle to meet the strict criteria for direct discrimination. Similarly, unless the perpetuation of these stereotypes by AI leads to an environment that is intimidating, hostile, or degrading for the individuals concerned, such genAI-produced harmful stereotypes might not constitute harassment under the legal definition. Furthermore, socio-economic status alone does not constitute a protected category under the EU non-discrimination directives.[124] Therefore, for a successful legal claim based on harassment, the claimant would have to show that certain ethnic, racial, religious or other protected groups specifically suffer from derogatory content concerning welfare recipients.

In both examples, on boring men and welfare recipients, the challenge lies in the indirect nature of how the communication, reinforcement and spreading of stereotypes can influence perceptions and treatment. Rather than culminating in overt acts of discrimination or the creation of a clearly hostile environment, communicative acts often exert a subtler influence on decisions and society. This indirect influence, while potentially harmful and contributing to a

---

[123] Cf. note 22.

[124] See also: Raphaële Xenidis and Linda Senden, 'EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination' in Ulf Bernitz et al (eds), General Principles of EU law and the EU Digital Order (Kluwer Law International, 2020) 151.

broader societal issue of bias and discrimination, is not easily captured by current non-discrimination law.

Furthermore, to the extent that such AI-generated statements are ultimately posted online not by AI but by humans, regulators and judges must also consider the fundamental right to freedom of speech. While certain types of hate speech are not protected, in political discussions,[125] not every critique of a protected group can or should be ruled out by the law. The current filtering mechanisms in non-discrimination law – direct connection to a disadvantageous act or creation of a toxic environment – constitute one way of operationalizing this delicate balance between robust societal discourse and the protection of certain groups against stereotypes and harm.

In sum, harmful stereotypes form the second category of novel types of generative harm. It is difficult to apply the bans on direct and indirect discrimination and of harassment to genAI systems that produce harmful stereotypes.

7.   Misclassification

We now turn to the third category of hard cases: misclassification. Misclassification may happen by genAI, but also by more traditional AI systems, such as facial recognition systems. Misclassification involves incorrect or offensive categorizations by AI systems, impacting individuals based on their identity or membership in protected groups.

An infamous real-world example of misclassification involves Google's image recognition tools mistakenly labeling Black individuals as 'gorillas'.[126] This situation at first blush looks like direct discrimination. But the AI service was not denied to a certain group or individual; rather, it was provided in a deficient manner that not only made certain users feel unwelcome but also deeply offended. However, the direct link to a specific service or decision under non-discrimination law, which could categorize this as direct discrimination, is somewhat ambiguous. The service's failure primarily led to an unwelcoming and offensive environment for affected users. This framing is typical of harassment cases (see below).

In a court case not related to AI, individuals were misclassified into the wrong protected group (e.g., incorrect gender or religion) in the context of a service offering; and a court ruled that this constitutes direct discrimination. The District Court of Frankfurt am Main held that if an online store (in the case, German railway services: Deutsche Bahn) offers a choice between gendered titles (e.g., Mr. and Ms.), it must also include an option for non-binary individuals.[127] Note that this is different from the Sparkasse case as, in the Deutsche Bahn case, a specific distinction was drawn between men and women (hence, no generic masculine form), for the purpose of gender self-identification; but non-binary identifications were omitted. The rationale is that failing to do so forces individuals to deny their own identity concerning the protected attribute,

---

[125] See, e.g., German Constitutional Court (BVerfG), Order of 4 November 2009, Case 1 BvR 2150/08, ECLI:DE:BVerfG:2009:rs20091104.1bvr215008.

[126] Maggie Zhang, ‚Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software' (forbes.com, accessed 4 April 2024) https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/.

[127] District Court (LG) Frankfurt a. M., Judgment of August 26, 2021, Case 2-30 O 154/20, para. 17.

constituting direct discrimination.[128] The appeals court refused to hear the case, making the first judgment binding between the parties.

Nevertheless, such instances of misclassification, while potentially falling under direct discrimination, are arguably better addressed as cases of harassment within the context of offering publicly available services. The primary issue lies more in the creation of a negative 'atmosphere' rather than a direct decision impacting the individual, making harassment a more fitting categorization. Harassment, in these contexts, involves the AI system's contribution to an environment that violates the dignity of individuals by not acknowledging or misrepresenting their identities. As discussed above, once this crosses the threshold into a 'hostile environment' concerning the service offered, the classification of harassment follows. In the cases discussed here (Gorilla case; gendered titles), that threshold is reached.

The law does not only protect the person primarily harassed.[129] In *Coleman v Attridge Law*, the CJEU decided that the prohibition of harassment vis-à-vis disabled persons also extends to the mother of a disabled child.[130] Hence, one may conclude that, similarly, persons belonging to the same protected group as the harassed person may also legitimately claim to be harassed, even if the immediate action is directed against another person, as long as the impact's intensity reaches the required threshold also vis-à-vis these persons of the same protected group. For example, other Black persons may legitimately claim to be harassed by the gorilla labeling even if photos of other Black persons, and not of themselves, were labeled guerrillas by a certain facial recognition tool (assuming that the impact reaches the required threshold).

In conclusion, while misclassification by AI can be viewed as direct discrimination under certain legal precedents, it is often more aptly dealt with under harassment, given the atmospheric and identity-related implications of such actions. Under such an approach, a judge would have to make a comprehensive assessment of whether the misclassification, while objectionable, is legally negligible – or whether it creates an adverse environment strong enough to warrant legal intervention.

A further question that arises from these considerations is whether social stigma, resulting from repeated and systematic AI misclassifications, could be recognized as a relevant legal disadvantage. This question challenges existing legal frameworks to adapt and consider the broader social and psychological impacts of AI-driven decisions and classifications, beyond the immediate legal definitions of discrimination and harassment.

---

[128] District Court (LG) Frankfurt a. M., Judgment of August 26, 2021, Case 2-30 O 154/20, para. 17.

[129] Evelyn Ellis and Philippa Watson, EU anti-discrimination law (OUP Oxford 2012), 175; see also Sandra Wachter, 'Affinity profiling and discrimination by association in online behavioral advertising' (2020) 35 Berkeley Tech LJ 367.

[130] CJEU, Case C-303/06, Coleman v Attridge Law, para. 63.

c. Justification

Discrimination is not automatically illegal, however. Rather, some instances of discrimination can be justified. Indirect discrimination, for example, may be justified, roughly summarized, if the prima facie discriminatory practice pursues a legitimate aim and the practice does not go beyond what is necessary to achieve that aim. The requirements are stricter, though, for direct discrimination and harassment.

i. Direct discrimination

Different legal provisions articulate varying types of possible justifications for direct discrimination. Under EU law, for example, direct discrimination related to ethnicity can hardly ever be justified. On the contrary, the Gender Goods and Services Directive implements a proportionality assessment for direct discrimination based on gender. Some types of prima facie discrimination based on gender can be justified if the provision of the goods and services exclusively or primarily to members of one sex is justified by a legitimate aim and the means of achieving that aim are appropriate and necessary..[131] In German legislation, a similar principle of proportionality applies to direct distinctions made on the basis of religion, disability, age, and sexual orientation.[132] The general idea behind such possible justifications is that, in some cases, there may be good reasons for making certain services or offers available to members of specific protected groups under special conditions (student discounts (with age limits); women-only parking, passenger cars on trains etc.).

Things are different with respect to direct discrimination in the workplace. Here, the prima facie discrimination can only be justified if the differential treatment constitutes a genuine and determining occupational requirement.[133] This means that if none of the candidates had the required trait, the role would remain unfilled, rather than hiring someone without the trait.[134] This contrasts with indirect discrimination scenarios, where a justifying trait may merely be desirable, but not strictly necessary for the job description.

What does all of this tell us about the cases of genAI discrimination? The justification of direct discrimination, such as in the 'do not hire xyz persons' scenario, will generally be difficult to achieve. However, if a protected attribute indeed constitutes a genuine and determining occupational requirement, statements made by genAI may be justified just like those of humans. For example, for reasons of privacy and intimacy, a specific gender may constitute such a requirement in a gender-separated massage parlor. Hate speech, however, will be almost impossible to justify.

To the extent that inadequate representation may, in certain cases, amount to direct discrimination, justification will be challenging. A company using a genAI system as a chatbot might claim that, for example, more balanced training data sets were difficult to find, and the output merely mirrors the unequal distribution of power, wealth, and accompanying

---

[131] Article 4(5), Gender Goods and Services Directive 2004/113/EC.

[132] See § 20 of the German General Equal Treatment Act (AGG)

[133] Article 4(1) of the Race Equality Directive; Article 4(1) of the Employment Equality Directive 2000/78/EC.; Article 14(2) of the recast Gender Equality Directive 2006/54/EC.

[134] Thüsing, '§ 8 AGG' in: Münchener Kommentar BGB, 8th. ed. (Beck, 2021), para. 8 et seqq.

representative pictures, in the world. However, in our view, this cannot be a sufficient justification: non-discrimination law precisely seeks to prevent the repetition of societal and historical biases by mere reference to custom, tradition, or the status quo. Hence, the developer would, as a minimum, have to undertake reasonable efforts to render the representation more diverse, by augmenting the training data set or by (manually or automatically) infusing more diverse output examples into the output space.[135]

Finally, if one follows the logic of the case law treating misclassification as a case of direct discrimination, the law would ask if there was a valid and convincing reason for the misclassification.[136] Generally, this will rarely be the case. In the gorilla example, Google eventually solved the problem by removing the gorilla tag from the labeling model.[137] For gendered and other categories, it is sufficient to provide the option of 'none of the above'. That is a feasible option; not implementing it cannot be justified under the law.

ii. Harassment

While, theoretically, harassment could be justified like direct discrimination in some cases,[138] this will generally not be possible in practice. It is hardly conceivable to find a convincing reason for creating a hostile environment that violates a person's dignity.

To conclude: Direct discrimination may be justified under certain circumstances. First, if the protected attribute constitutes a genuine and determining occupational requirement, genAI output restricting job offers or rankings to members of the required group is legitimate. Second, in cases involving the offering of goods and services, direct discrimination can be generally justified via a legitimate reason and the proportionality principle, with the exception of discrimination on the basis of ethnic origin. Hence, contractual templates for such transactions may be generated reflecting this. Finally, harassment will not be justifiable.

d. Summary concerning discrimination and justification

The following *Table 2* summarizes our findings concerning generative AI discrimination, broken down in yet another way.

	Demeaning and Abusive Output					
	Negative	Non-	Hate	Harmful	Inadequate	Misclassifi

---

[135] See, in greater detail, Philipp Hacker, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143, 1163-1165.

[136] In offering goods and services, direct discrimination can be justified in this way, see, e.g., Article 4(5) Good and Services Directive 2004/113/EC. For all other attributes, EU law does not exist (concerning goods and services), but Member States may have similar rules as in the gender case (e.g. § 20 German AGG). The exception is racial and ethnic origin. Here, justification of direct discrimination is only possible via a genuine and determining occupational requirement, and positive action (Article 4 and 5 of Directive 2000/43/EC).

[137] Nico Grant & Kashmir Hill, 'Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's', New York Times, 22 may 2023, <https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html> accessed May 3 2024.

[138] See, e.g., Article 4(1) in conjunction with Article 2(3) of the Framework Directive.

	speech	inclusive language	speech	stereotypes	representat ion	cation
Dir. Discr.	Generally no	Generally no	Yes	Generally no	Generally no	Disputed
Harassment	Possible	Generally no	Yes	Generally no	Generally no	Possible
Justification	Rare		No			Generally no

Table 2: Summary of generative AI discrimination cases

Another overview is provided by the following graph diagram. Note that inadequate representation may, but need not necessarily, amount to direct or indirect discrimination. Outright discriminatory content can primarily cross the harassment threshold. But, as the preceding analysis has shown, the devil is in the details.
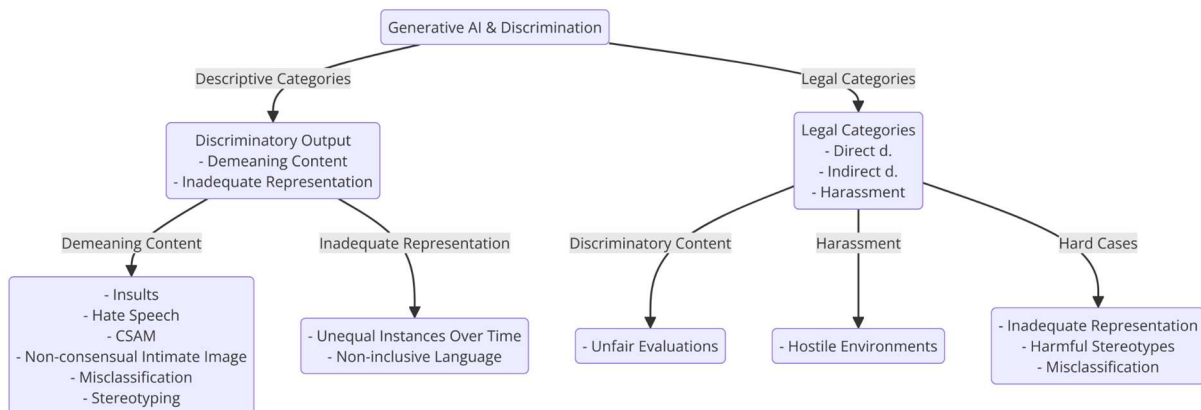

Figure 1: Overview of descriptive and legal categories in generative AI discrimination
© authors

Finally, the relationship between the descriptive and the legal categories is summarized by the following chart:
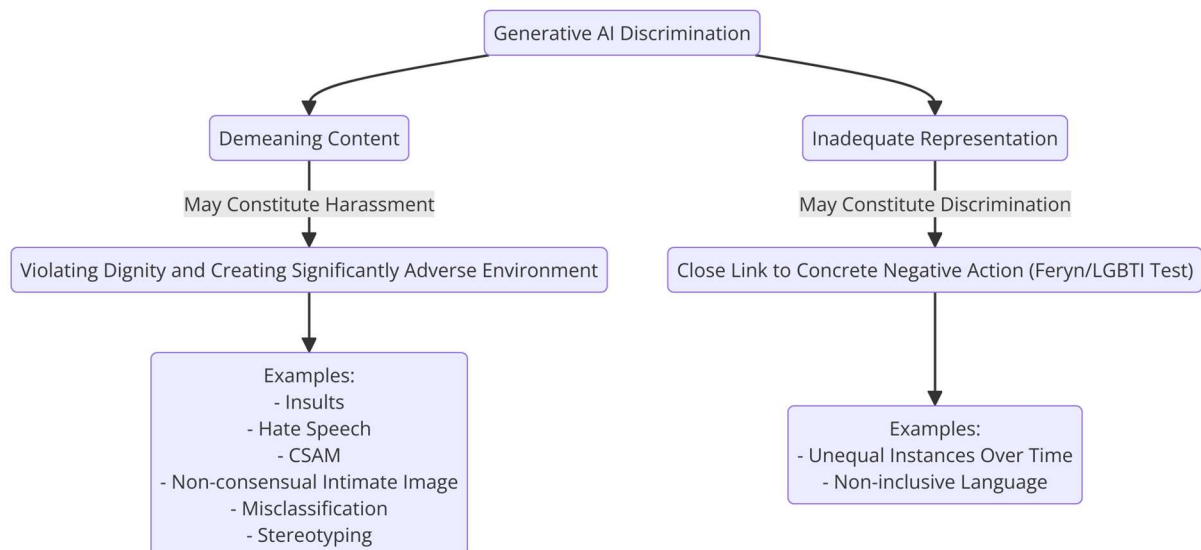
Figure 2: Relationship between descriptive and legal categories
© authors

3. Responsibility and Liability

The next question is: who should be liable for discriminatory output by genAI systems? Under non-discrimination law, liability is strict, meaning that fault or negligence does not need to be proven for an entity to be held responsible for discrimination. This creates a complex issue when an AI model, developed by Company A but deployed by Company B, outputs content constituting illegal discrimination. Non-discrimination law primarily targets the deploying organization for the discriminatory actions of the AI system, regardless of its origin. By contrast, product liability law places responsibility on the developer (manufacturer) for any defects or errors in the product.

a. Deployer/user of the genAI system

Under non-discrimination law, a company using an AI tool like ChatGPT for creating content, such as writing news articles, is legally responsible if the output contains discriminatory expressions. This allocation of liability is underpinned by several considerations:

Because EU non-discrimination law focuses on strict liability, it is irrelevant whether the developer or deployer of a genAI system intended to discriminate. Courts could use rules of attribution, under Member State law, to link the genAI output to the deployer, much as the company would be liable for misconduct of its employees, agents, or organs.[139] Such attribution ensures that deployers cannot absolve themselves of responsibility for discriminatory outcomes by blaming the genAI system. This strict liability maintains accountability with the entity that benefits from and controls the AI's deployment.

---

[139] See, e.g., Philipp Hacker, 'Verhaltens-und Wissenszurechnung beim Einsatz von Künstlicher Intelligenz' (2019) 9 RW Rechtswissenschaft 243; Philipp Hacker, Datenprivatrecht: Neue Technologien im Spannungsfeld von Datenschutzrecht und BGB (Mohr Siebeck 2020), 609 et seqq.

In the future, legal frameworks may evolve to address the intricacies of AI-driven activities through specific rules of attribution, akin to how the law assigns responsibility for actions by employees or corporate entities. This adaptation could also recognize discrimination involving AI as a complex, multistep process with the deployer playing a crucial role due to their integration and use of the AI tool for their benefit. Under a general justice perspective, such an assignment of liability harks back to the old principle that the advantages gained from employing a tool come with a responsibility for its negative consequences: benefits and responsibilities are intrinsically linked (*qui habet commoda ferre debet onera*[140]). Given these considerations, the deploying company should not be able to disclaim liability for discriminatory acts of AI systems it uses.

b. Developer/provider of the genAI system

To what extent should the genAI developer be liable for discriminatory output? The developer also benefits from the deployment and use, at least indirectly, via sales or licensing agreements, or opportunities to further refine and train the model. This value chain suggests a need for a nuanced regulatory approach. Oftentimes, the developer caused the discriminatory output: training data may have been biased; feedback skewed; parameters wrongly set; etc. For liability to set incentives to mitigate discriminatory outcomes, developers would need to be covered by non-discrimination damages, too. In the US, the Equal Employment Opportunity Commission (EEOC) recently made a similar argument in proceedings between a rejected job candidate and Workday, a company developing hiring tools.[141]

Thus, both developers and deployers share a symbiotic responsibility: developers must strive to create non-discriminatory AI tools, and deployers must ensure that their genAI use complies with non-discrimination laws. Legally, such a responsibility corresponds to joint and several liability. The victim may choose whom to sue. That deployer may then turn around and seek to recover damages paid from the developer (or vice versa), usually to the extent contractually determined or proportionate to each parties' degree of fault in bringing about the outcome.

In conclusion, under current non-discrimination law, the deployer/user of the genAI system is typically responsible and liable for discriminatory output by the genAI system. In our view, the developer/provider of the genAI system should be jointly responsible and liable.

IV. Connections to the other legal instruments

Non-discrimination law is not the only legal field tackling discriminatory output by genAI, however. We briefly highlight some adjacent fields of law that may be relevant for genAI discrimination: personality rights; the General Data Protection Regulation (GDPR); the Digital Services Act (DSA); and the AI Act.

---

[140] Zimmermann, The Law of Obligations, 1990, 201 Fn. 108 und 290.

[141] Daniel Wiessner, 'EEOC says Workday must face claims that AI software is biased', Reuters (April 11, 2024), https://www.reuters.com/legal/transactional/eeoc-says-workday-covered-by-anti-bias-laws-ai-discrimination-case-2024-04-11/.

1. Personality rights

Personality rights are a broad category of rights, protected by lawmakers and courts in different ways around the world. They 'recognise a person as a physical and spiritual-moral being and guarantee his enjoyment of his own sense of existence'.[142] In Europe, personality rights are predominantly governed by national legislation - not by the EU.

A case that sheds light on *developer* liability is the Autocomplete judgment by the German Federal Court of Justice in Germany.[143] The court held that a search engine provider could be liable for its autocomplete function's suggestions if the company failed to take reasonable measures to prevent defamatory or rights-violating suggestions.[144] Specifically, the company is required to act once it becomes aware of such harmful outputs, indicating a duty to mitigate future occurrences.[145] In the case, the wife of then German President, Bettina Wulff, sued Google because the most prominent autocomplete suggestion following her name was 'prostitute'. This mirrored rumors in the tabloid press that, before meeting her husband who later went on to become Germany's President, she had worked as a sex worker.

This precedent is relevant for developers of large language models, which might be called 'large autocomplete engines.' Conversely, the autocomplete function in Google used a 'small language model' at the time.[146] The ruling suggests that developers might face direct liability for the AI's outputs if they neglect to implement safeguards against the generation of harmful content. Such measures could include content moderation during the AI's training or responsive action upon notification of problematic outputs.

Concerning the liability framework for *deployers* of AI technologies, we expect judges to align it closely, mutatis mutandis, with the obligations of developers, focusing on the management of infringing outputs. In fact, in the Autocomplete case, Google acted as both developer and deployer. In our view, both developer and deployer need to do what they reasonably can to prevent the violation of personality rights by genAI output. Deployers could potentially prevent liability by creating a robust compliance system. Essential elements of this system include the use of AI equipped with 'guardrails' for moderation; regular proactive monitoring of AI outputs; and a notice and takedown process to quickly address and prevent the recurrence of harmful content, such as by blocking specific prompts or outputs. Again, both developers and deployers may be jointly and severally liable.

Overall, we expect both developers and deployers to be compelled by judges, under current law, to do what they reasonably can to prevent genAI output infringing personality rights.

---

[142] Johann Neethling, 'Personality rights: a comparative overview' Comparative and International Law Journal of Southern Africa 38, no. 2 (2005): 210-245, p. 210. Internal citations omitted.

[143] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Do Large Language Models Have a Legal Duty to Tell the Truth?' [2024] Royal Society Open Science.

[144] Autocomplete| [2013] | BGH | VI ZR 269/12, | [2013] BGH, 14.05.2013 - VI ZR 269/12, Rn. 36.

[145] ibid.

[146] Rostyslav Demush, 'From Crappy Autocomplete to ChatGPT: The Evolution of Language Models', Hackernoon (March 14, 2023), https://hackernoon.com/from-crappy-autocomplete-to-chatgpt-the-evolution-of-language-models.

2. The GDPR

The GDPR[147] is another European framework that could apply to the output of content created by large language models. However, the following shows that the GDPR is ill-equipped to address the harms that we have outlined above.

The GDPR is predominantly concerned with issues of data protection and was not designed to deal with advanced AI and machine learning. The GDPR largely focuses on what has been described as 'procedural privacy,'[148] referring to the rules that govern whether, how, and when data can be collected and processed.[149] The Regulation is predominantly concerned with transparency around data processing rather than the outputs that are created based on this processing.[150] Therefore, the GDPR is generally not well suited to deal with undesirable output or harmful content produced by genAI.

Despite this, could some GDPR provisions help to address discriminatory effects of genAI? Article 5(1)(a) GDPR establishes 'lawfulness, fairness and transparency' as three of the GDPR's guiding principles. At first glance, one might think that genAI discrimination is unfair, and that the GDPR should therefore be able to help against such unfair effects.[151] However, the term fairness is not further defined in the GDPR[152] and thus offers little insight into how and if this principle could resolve any of the tensions lined up above.[153]

To the extent that genAI produces *false* (not necessarily discriminating) information about identifiable individuals (hallucinations), however, the GDPR's accuracy principle might be violated. This claim is at the center of the recent complaint lodged by Max Schrems' NGO noyb against OpenAI, before the Austrian Data Protection Authority.[154] Article 5(1)(d) GDPR enshrines the accuracy principle. Personal data must be correct and up-to-date. However, this

---

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

[148] Leighton Andrews and others, 'Algorithmic Regulation' (London School of Economics and Political Science, 2017) 38 <https://orca.cardiff.ac.uk/id/eprint/105059/1/DP85-Algorithmic-Regulation-Sep-2017(1).pdf>.

[149] See, e.g., Christopher Kuner, 'Territorial Scope and Data Transfer Rules in the GDPR: Realising the EU's Ambition of Borderless Data Protection' (2021) University of Cambridge Faculty of Law Research Paper.

[150] Sandra Wachter and BD Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' (2019) 2019 Columbia Business Law Review 494.

[151] See, e.g., Philipp Hacker, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143, 1172.

[152] Sarah Johanna Eskens, 'Profiling the European Citizen in the Internet of Things: How Will the General Data Protection Regulation Apply to This Form of Personal Data Processing, and How Should It?' (Social Science Research Network 2016) SSRN Scholarly Paper ID 2752010 27 <https://papers.ssrn.com/abstract=2752010> accessed 8 July 2017 discussing how fairness means transparency and does not have any standalone meaning nor definition within the GDPR.

[153] For more literature on this topic and the link between fairness and transparency, see non-binding guidelines issued by the Article 29 Data Protection Working Party, Article 29 Data Protection Working Party, 'Guidelines on Transparency under Regulation 2016/679' (2017) 17/EN WP 260; see also Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679, 17/EN WP 251rev.01' (2018) <http://ec.europa.eu/newsroom/article29/document.cfm?doc_id=49826> which are more generous but yet equally unhelpful in defining fairness.

[154] noyb, 'ChatGPT provides false information about people, and OpenAI can't correct it', noyb blog (April 29, 2024), https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it.

principle is not without limits.[155] In our view, the accuracy principle must be weighed against countervailing fundamental rights, such as those enjoyed by the LLM providers, such as the right to conduct a business.[156] Hence, authorities and courts may ultimately find that not every piece of false information violates the accuracy principle, and must be corrected (de minimis threshold). The result would then be similar to the one under personality tort law. Important false information violates the accuracy principle; trivial information possibly does not (e.g., if ChatGPT gets a birthday wrong, unless that is important in the setting).

Article 9 GDPR deals with 'processing of special categories of personal data'. Ethnic origin and sexual orientation – but not gender – are considered special category data that can only be processed under restricted circumstances (e.g., explicit consent). It is unlikely, however, that this provision would apply broadly to genAI because genAI outputs rarely feature personal identifiable information. Racist or homophobic content, for example, that is not linked to a specific person would not fall within the scope of the GDPR. To the extent that GenAI output is linked to a person, , however, Article 9 GDPR will likely be breached (as the provider will generally not be able to avail itself of the exceptions in Article 9(2) GDPR).

Lastly, Article 22 GDPR deals with 'automated individual decision-making, including profiling'. This provision limits when automated decision-making can be deployed (e.g. when based on explicit consent) and details that certain safeguards have to be in place to protect the interests of the data subject (e.g. human intervention, contestation of the decision).[157] While outputs of generative models can be classified as 'automated decisions' without any human intervention, it is questionable whether the output is a) a decision made about the data subject and b) whether the decision produces a 'legal' or similarly significant effect.

In most circumstances, receiving outputs after prompting a chat bot or diffusion model will not be a decision about an individual. Even if the output could be construed as a decision about the person, the decision probably does not produce legal or similarly significant effects. Recital 71 speaks about employment or credit decisions as automated decisions with legal or similarly significant effects, which shows that the legislator predominantly had predictive AI systems in mind when drafting this provision. However, if genAI systems make recommendations about, for example, treating patients based on summaries of their patient files, this could count as a decision producing a similarly significant effect.[158] In sum, we do not expect that the GDPR will and can play a large role in limiting the harms of genAI *discrimination*.

However, as mentioned, the complaint brought in April 2024 by Max Schrems against OpenAI is trying to use GDPR provisions to create recourse against genAI hallucinations that produce inaccurate outputs (e.g. wrong birth day of a public figure).[159] While this case does not directly

---

[155] Cf. Rec. 39 GDPR.

[156] Article 16 of the Charter of Fundamental Rights of the European Union.

[157] For detailed discussion about the limitations of this provision see Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 International Data Privacy Law 76.

[158] Case 634/21, *OQ v Land Hesse* [2023], ECR 2023:957.

[159] noyb, 'ChatGPT provides false information about people, and OpenAI can't correct it', noyb blog (April 29, 2024), https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it.

deal with discriminatory outcomes, it may shed light on the extent to which the GDPR offers remedies against *factually incorrect* genAI outputs.

3. The Digital Services Act

The EU Digital Services Act (DSA)[160] concerns, among other things, the removal of illegal content from online platforms. Roughly summarized, the DSA governs questions of liability for Internet platforms for content hosted on their platforms. If Internet platforms act as 'intermediaries' and are a 'mere conduit, providing 'caching' or 'hosting' services, they are not liable for the content hosted on their platforms.[161] This liability privilege ceases if providers of hosting services become aware of illegal activity, in which case they must take action (i.e., notice and takedown[162]).

If bias related harm rises to an equivalent of hate speech or harassment or if the created content on a platform violates Member State laws, and the platform is warned about the illegal content, the platform must take action and remove that content. However, as mentioned above, the harms that concern us rarely rise to the level of these traditional harms.[163] Yet, where pre-existing laws criminalize certain harms, platforms will need to take action.[164]

An interesting situation might arise if online platforms increasingly implement genAI in their search engines or chatbots and platforms. On the one hand, the Commission has already requested information from Bing on how they conduct their risk assessment and mitigation under Article 34 and 35 DSA in the face of hallucinations in its genAI search.[165] On the other hand, genAI use by the platform could render the liability privilege inapplicable as the platform would not only host but also create content.[166] Therefore, the platform could become liable for all content hosted on the platform regardless of the platform's actual awareness. However, such platform liability will not lead to more protection for affected parties unless Member State law already recognises our envisioned genAI discrimination harm as an illegal act. To sum up, the DSA only offers protection to the extent that Member State laws already recognise the harms outlined in this paper.

---

[160] 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 2000/31/EC (Digital Services Act) (Text with EEA relevance).

[161] Article 4-6 DSA

[162] Article 16 DSA

[163] With that said, Member States may have existing laws that already punish these or similar harms outlined in this paper. A full analysis of relevant Member State law goes beyond the scope of this paper.

[164] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Do Large Language Models Have a Legal Duty to Tell the Truth?' [2024] Royal Society Open Science

[165] https://digital-strategy.ec.europa.eu/en/news/commission-compels-microsoft-provide-information-under-digital-services-act-generative-ai-risks.

[166] Philipp Hacker, Andreas Engel and Marco Mauer, 'Regulating ChatGPT and Other Large Generative AI Models', *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2023) 10 <https://dl.acm.org/doi/10.1145/3593013.3594067> accessed 27 October 2023.

4. The AI Act

The EU Artificial Intelligence Act[167] and the forthcoming AI liability directives are unlikely to establish sufficient accountability mechanisms that would remedy the discrimination-related harms discussed in this chapter.[168]

The AI Act is predominantly a product safety law. In fact, the original drafts did not have any individual rights components.[169] Only the drafts of the European Parliament[170] foresaw that individual complaint-based mechanisms were incorporated in the draft. While the most recent version of the AI Act includes a right to lodge a complaint with a Market Surveillance Authority and right to explanation of individual decision-making,[171] those provisions are unlikely to alleviate most of the concerns we have noted above.

While bias is recognised as an AI-related issue that needs addressing,[172] it is unclear whether discrimination-related harms as exemplified in this paper can be seen as a violation of the AI Act against which complaints can be brought. For example, Article 10 AI Act does address the issue of bias in training data, but it is targeted at developers of predictive AI systems in high risk areas, not at developers of genAI. If genAI is developed for a high risk area, developers have to follow these rules. However, the viability of Article 10 in mitigating generative discrimination will crucially hinge on the interpretation of 'bias' in the AI Act. At the moment, the question remains open whether the harms envisioned in this chapter would fall under the AI Act's concept of 'bias.' Perhaps regulators will interpret 'bias' in a technical (diversity of training data) and less social and ethical (demeaning and abusive contents) way.

The right to explanation might also not be helpful for our purposes. The provision focuses on the right to have decisions explained that were rendered in high risk areas by predictive AI such as employment, criminal justice or education. The AI Act does not classify genAI as a high-risk application and thus the right to explanation does not generally apply to genAI.[173] Further,

---

[167] Based on the final EP version of the AI Act available under https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf.

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

[169] Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach' (2021) 22 Computer Law Review International 97; Hacker, Engel and Mauer (n 12); Martin Ebers, 'Standardizing AI-The Case of the European Commission's Proposal for an Artificial Intelligence Act' [2021] The Cambridge handbook of artificial intelligence: global perspectives on law and ethics; Johann Laux, Sandra Wachter and Brent Mittelstadt, 'Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act' (2024) 53 Computer Law & Security Review 105957.

[170] European Parliament, 'Compromise Amendments on the Draft Report Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD))' (2023) KMB/DA/AS <https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf>.

[171] Article 85 and 86 AI Act. Art 99 (10) also offers recourse against decisions or omission to take action of the Market Surveillance Authority.

[172] See Article 10 AI Act.

[173] Natali Helberger and Nicholas Diakopoulos, 'ChatGPT and the AI Act' (2023) 12 Internet Policy Review 3, 6 who argue that genAI should be its own high-risk category.

as we have mentioned above it is unlikely that the outputs of genAI are 'decisions' about people. So even if the outputs are rendered in a high-risk setting (e.g. employment), it is unlikely that the right to explanation will apply.

The AI Act lacks generally applicable provisions for developers of genAI systems to mitigate bias. The provisions around genAI focus on transparency, copyright protection, watermarking,[174] and reporting of energy consumption.[175] For highly capable models (trained with more than 10^25 FLOPS or those designated by the Commission) there are additional requirements because they are seen as models with systemic risks (Art 55). Providers of genAI models with systemic risks have to undertake model evaluation, risk assessments (including red teaming), have reporting duties of serious incidents and have to ensure cyber security.[176] None of these obligations explicitly address discrimination-related harms directly. While risk assessments and mitigation should, in our view, include bias and non-discrimination, only the harmonized standards (Article 40/41) and the codes of practice (Article 56) will show if these provisions on systemic risks will address discrimination issues in practice, which types of bias, and what remedies have to be taken.

To sum up, the main provisions of the AI Act are focused on non-genAI, and, unfortunately, provisions are lacking that would explicitly require developers of genAI models to create systems that are less biased.[177]

5. The AI Liability Directives

Similar issues plague the AI liability directives. At the time of writing, the European institutions are negotiating[178] an update to the current Product Liability Directive[179] and the creation of a new AI Liability Directive.[180]

The main issue with the Product Liability Directive is one of scope. The original proposal offers financial remedies against software and AI related harms that occur due to death, personal

---

[174] Article 53(1)a and 55 AI Act.

[175] Annex XI, Section II(2)(e) AI Act.

[176] Article 52a, 52d and Annex IXa AI Act. Sandra Wachter, 'Limitations and Loopholes in the E.U. AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond' (2024) 26 Yale Journal of Law and Technology.

[177] The harmonized standards might add new obligations, but this remains to be seen.

[178] For an overview of shortfalls of these proposals see Philipp Hacker, 'The European AI Liability Directives– Critique of a Half-Hearted Approach and Lessons for the Future' (2023) 51 Computer Law & Security Review 105871. For more discussion on the limitations of the PLD see Herbert Zech, Liability for AI: Public Policy Considerations, 22 ERA FORUM 147 (2021); Christpoh Schmon, Product Liability of Emerging Digital Technologies, 3 Zeitschrift für Internationales Wirtschaftsrecht 254 (2018); Martin Ebers, Liability for Artificial Intelligence and EU Consumer Law, 12 J. Intell. Prop. Info. Tech. & Elec. Com. L. 204 (2021); Christiane Wendehorst, Strict Liability for AI and Other Emerging Technologies, 11 Journal of European Tort Law 150 (2020).

[179] European Commission, Proposal for a Directive of the European Parliament and of the Council on liability for defective products 2022/0302(COD) 2022; now adopted: https://www.europarl.europa.eu/news/en/press-room/20240308IPR18990/defective-products-revamped-rules-to-better-protect-consumers-from-damages.

[180] European Commission, Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive) 2022.

injury, destruction of property, or loss of data.[181] Pure economic loss and material harms were not covered in the original draft.[182]

Recital 24 further states that 'Types of damage other than those provided for in this Directive, such as pure economic loss, privacy infringements or discrimination, should not by themselves trigger liability under this Directive'.[183] The latest draft of this proposal does introduce the idea of immaterial harm ('pain and suffering') in Recital 23, but only if it is a side-effect of one of the harms covered in the Product Liability Directive.[184]

The generative discrimination harms that we are envisioning are usually not associated with a material harm such as the destruction of property or personal injury. It is likely that this harm will exclusively occur as a stand-alone harm and therefore will not be covered by the updated Product Liability Directive.

More promising is the AI Liability Directive, which is a complimentary framework for the AI Act designed to offer remedies against AI related harms. The current version also allows recourse against fundamental rights violations if recognised by Member State laws,[185] and thus recourse against immaterial harms. The preamble even mentions 'violations of personal dignity (Articles 1 and 4 of the Charter), respect for private and family life (Article 7), the right to equality (Article 20) and non-discrimination (Article 21)'.[186]

Yet, it is again questionable whether the AILD will offer a successful route to protecting people against generative discrimination. As we have shown, the harms we are envisioning often do not fit with the fundamental rights protections against discrimination and harassment. It is unclear how novel discrimination-related harms would fit under traditional fundamental rights protection.[187]

But even if one would find an appropriate link to an existing fundamental right and Member State law does protect against violation, there are several hurdles. For instance, the evidential requirements create an almost insurmountable burden for claimants.[188] And there are several limitations of the applicability of the EU Charter. The Charter only applies in EU law matters. Moreover, the Charter's horizontal applicability is widely contested,[189] which means that the

---

[181] Art 4 (6) Product Liability Directive.

[182] Christiane Wendehorst, 'Strict Liability for AI and Other Emerging Technologies' (2020) 11 Journal of European Tort Law 150, 162; See also Hacker (n 138) 28 who criticizes this exclusion.

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

[184] European Parliament, P9_TA(2024)0132 - Liability for defective products - European Parliament legislative resolution of 12 March 2024 on the proposal for a directive of the European Parliament and of the Council on liability for defective products (COM(2022)0495 – C9-0322/2022 – 2022/0302(COD)) 2024; Wachter (n 137).

[185] Article 2 (9) Product Liability Directive.

[186] AILD Proposal, p. 10.

[187] It is not impossible that Member States already recognise these or equivalent harms. Should this be the case, then this route would open up.

[188] For details of the evidential requirements for both the PLD and AILD, see Hacker (n 138), see also, Sandra Wachter, 'Limitations and Loopholes in the E.U. AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond' (2024) 26 Yale Journal of Law and Technology. .

[189] See, e.g., CJEU, Case C-414/16 (Egenberger); C-68/17 (IR); Joint Cases C-569/16 and C-570/16 (Bauer und Willmeroth); Matteo Fornasier, 'The impact of EU fundamental rights on private relationships: direct or indirect effect?' (2015) 23 European Review of Private Law 29; Nuria Bermejo, 'Fundamental Rights and Horizontal Direct

private sector might not have to follow it. Yet industry is the predominant developer of genAI systems.[190]

To summarize, as it stands right now, both the AI Liability Directive and Product Liability Directive are not well suited to address the harms we envisioned in this paper, predominantly because pure immaterial harms are unlikely to be covered under these frameworks.

V. Technical mitigation and difficulties

With the law struggling to rein in generative discrimination: Can genAI developers technically prevent discriminatory output? Mitigating biases in technical ways is never a neutral technical fix, but always fraught with normative decisions.[191] Increasingly, "legal red teaming" frameworks are developed to navigate this landscape, also in a semi-automated way.[192] Normative decisions are especially relevant for attempts to remedy discriminatory outcomes in generative systems. This endeavor introduces unique challenges that traditional 'algorithmic fairness' (= non-discrimination) methods are ill-equipped to address.[193]

In contrast to the tasks of scoring, ranking, and classification in traditional AI, where non-discrimination might be achieved by adjusting numerical scores (e.g., by mapping from raw to fair scores[194]), ensuring that language or image output is non-discriminatory and inclusive demands a far more complex and nuanced approach. This complexity is partly because language and visual media are rich, multifaceted mediums that reflect and shape societal norms and values. Making language and image generation inclusive and non-discriminatory involves not only technical adjustments but also an understanding of linguistic and visual subtleties, cultural contexts, and the diverse ways biases can manifest in text, image, and video. Promising attempts include, for example, fine-tuning large models on small, but dedicatedly fair datasets.[195] However, even advanced safety fine-tuning remains vulnerable to people 'jailbreaking' the

---

Effect under the Charter' in Cristina Izquierdo-Sans, Carmen Martínez-Capdevila and Nogueira-Guastavino M (eds), *Fundamental Rights Challenges: Horizontal Effectiveness, Rule of Law and Margin of National Appreciation* (Springer 2021) 51; Aurelia Ciacchi, 'Egenberger and Comparative Law: A Victory of the Direct Horizontal Effect of Fundamental Rights' (2018) 5 European Journal of Comparative Law and Governance 207.

[190] For a detailed analysis of the horizontal, vertical and general application of the EU Charter in relation to genAI, see Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Do Large Language Models Have a Legal Duty to Tell the Truth?' [2024] Royal Society Open Science.

[191] See, e.g., Batya Friedman and Helen Nissenbaum 'Bias in computer systems' (1996) 14 ACM Transactions on information systems (TOIS) 330; Helen Nissenbaum, 'How computer systems embody values' (2001) 34 Computer 120; Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law' (2020) 123 W Va L Rev 735; Philipp Hacker, 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review 1143.

[192] Barclay Blair, Karley Buckley, Ashley Allen Carr, Coran Darling, Zev Eigen, Danny Tobey, Sam Tyner-Monroe, 'Legal red teaming: A systematic approach to assessing legal risk of generative AI models' DLA Piper White Paper (2024), 7.

[193] Sara Sterlie, Nina Weng and Aasa Feragen, 'Non-discrimination Criteria for Generative Language Models' (2024) arXiv preprint arXiv:240308564.

[194] See, e.g., Meike Zehlike, Ke Yang and Julia Stoyanovich, 'Fairness in Ranking, Part I: Score-based Ranking' (2022) ACM Computing Surveys (CSUR) 1; Meike Zehlike and others, 'Matching code and law: achieving algorithmic fairness with optimal transport' (2020) 34 Data Mining and Knowledge Discovery 163.

[195] Irene Solaiman and Christy Dennison, 'Process for adapting language models to society (PALMS) with values-targeted datasets' (2021) 34 Advances in Neural Information Processing Systems 5861.

model:[196] jailbreaking is, in short, a technique for making a genAI system provide harmful output which the genAI developer tried to prevent. Jailbrakers might succeed by experimenting with prompts to bypass the genAI system's safeguards, for instance.

The inherent randomness and variability of genAI models add another layer of difficulty. As mentioned, these models can produce vastly different outputs in response to slight modifications in prompts. Even the same prompt will usually, via a random vector, generate different output at different times.[197] This significantly complicates efforts to consistently audit generative systems and prevent discrimination. While unpredictability challenges the application of conventional 'fairness measures' (numerical requirements that aim to quantify discrimination), the dynamic nature of language and images, and the contextual dependency of meaning, make standardized corrections less straightforward.

Moreover, the concept of inclusive language itself is, as seen, a contested construct, requiring a delicate balance between grammatical norms, actual language use, and a fair representation of the world. The controversy surrounding Google's Gemini model, which generated historically inaccurate yet inclusively intentioned outputs, such as Black US Founding Fathers or racially diverse Nazi soldiers,[198] exemplifies the tensions inherent in attempting to balance inclusivity with historical accuracy and societal expectations. These incidents illuminate the difficulties in squaring uniform bias mitigation endeavors in AI systems with the broader social and historical considerations that inform what constitutes non-discriminatory, fair, consistent, accurate, and inclusive content.

Addressing bias in genAI thus demands not only sophisticated technical solutions but also engaging with the contested nature of inclusive language and imagery. More clearly than in regression or classification systems, the normative underpinning of bias 'correction' reemerges. Ensuring non-discrimination of genAI models in language generation is an inherently interdisciplinary challenge that extends beyond technical fixes, and can only be tackled by dialogue between diverse research fields, from computer science and law to linguistics and media studies. Exciting, and daunting, futures ahead.

VI.   Policy options

Technical strategies for bias mitigation in generative systems will only get us so far. To be rolled out in a robust and trustworthy way, they need to be accompanied with changes in regulation. In the following section, we give suggestions for policy options that could be explored by researchers, regulators, and policymakers. The options concern clarifying the law, updating the law, and shaping technology through the law.

1. Clarifying the law

Clarifying AI laws is crucial, enabling companies to innovate within legal limits confidently. Legal certainty, a goal of the AI Act, could help to ensure responsible AI development and deployment, addressing key issues like bias and data fairness, and fosters trust and compliance

---

[196] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt, 'Jailbroken: How does LLM safety training fail?' (2024) 36 Advances in Neural Information Processing Systems.

[197] David Foster, Generative Deep Learning (O'Reilly 2022), 4-5.

[198] See above.

across industries. While some clarifications may, from Spring 2026 on, be achieved by technical standards allowing companies abiding by them to avail themselves of the presumption of conformity, the following sections will offer some guidelines for interpretation with respect to non-discrimination beforehand.

a. Training data rules in the AI Act

Article 10(2)(g) AI Act mandates that providers of high-risk AI systems must mitigate bias in their training data. This provision marks a step towards addressing the foundational causes of discrimination and representational harm within AI systems and outputs. At first blush, one could be tempted to conclude that mitigating bias, in this context, means adjusting the training data until the outcomes of the AI system do not result in illegal discrimination as defined by non-discrimination law. Such an interpretation, however, will likely fail: it is practically impossible to predict the output of an AI model only on the basis of the training data.[199] Furthermore, to the extent that genAI systems are not used in non-high-risk settings, the provision does not apply at all to them.

A first and important update, therefore, would be to extend the training data rules to genAI systems. Essentially, Article 53 (listing the requirements for general-purpose AI systems) would have to refer to Article 10 AI Act.

Second, in our view, genAI providers should be compelled, under Article 10(2)(g) AI Act, to employ state-of-the-art techniques in data curation or preprocessing to foster non-discriminatory output. Such an interpretation would include efforts to counter representational harms — issues not fully addressed by existing frameworks of non-discrimination law, as we have seen. This requirement would bridge a gap highlighted by the application of the Feryn/LGBTI test in non-discrimination law. Additionally, Article 10 AI Act extends to data used for fine-tuning existing systems in high-risk sectors, to the extent that fine-tuning entities can be considered providers under Article 25(1) AI Act.

The introduction of such preemptive measures in the AI Act signifies a shift in law towards proactive steps in combating bias and promoting fairness in AI systems. By focusing on the balance and representativeness of training data, the legislation aims to tackle the root causes of bias, facilitating more equitable AI outputs. This move aligns with broader efforts to enhance the fairness of AI technologies along the entire machine learning pipeline.[200]

b. Input data rules in the AI Act

Article 26 AI Act details rules for the deployers of high-risk AI systems. Article 26(4) AI Act complements Article 10 by placing responsibility on deployers who have control over input

---

[199] Cf. Emily Black, Manish Raghavan and Solon Barocas, 'Model multiplicity: Opportunities, concerns, and solutions' (2022) Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 850.

[200] Emily Black and others, 'Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools' (2023) Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization 1.

data. Input data, in turn, are defined as 'data provided to or directly acquired by an AI system on the basis of which the system produces an output'.[201] Hence, Article 26 addresses the application data on which an AI system performs an analysis or generates content – for example, prompts in the case of genAI.

However, like Article 10, Article 26 does not apply to genAI directly, but only if it is used in non-high-risk settings. Hence, the scope of Article 26 should also be extended to deployers of genAI systems.

Article 26(4) AI Act mandates that deployers guarantee the relevance and representativeness of input data concerning the high-risk AI system's intended purpose. If deployers choose prompts unilaterally reflecting members of one protected group (e.g., men or women; white persons) when the output will refer to all protected groups may then be considered a breach of Article 26 – unless grammatical or historical contexts duly expand or justify the scope of linguistically narrow wordings.

Together, Article 10 and 26 AI Act mark an advancement in the legal framework governing AI, addressing both the development and deployment phases, aiming to prevent discrimination – but they should include genAI as well. Policymakers should consider amending both provisions, so that they apply to genAI, too.

2. Updating the law

Merely clarifying the applicability and extending the scope of existing law may not be enough, however. Rather, we need to consider updates in certain areas, particularly the applicability of non-discrimination law and the legal categories, currently focused on direct and indirect discrimination as well as harassment, in view of genAI harms.

a. Applicability of non-discrimination law

The current state of EU non-discrimination law leaves us with a problematic accountability gap for genAI. Non-discrimination law is predominantly focused on the fair allocation of resources or enabling equal access to resources (e.g. education, financial services) to different groups.[202] As a result, non-discrimination law is not well equipped to deal with the generative discrimination harms that we have identified.

Under current non-discrimination law, in order to seek a remedy claimants must demonstrate that a 'particular disadvantage' has been either suffered or is likely to occur. Non-inclusive language and biased depiction (e.g. pictures of predominantly male CEOs) do not immediately lead to a tangible 'particular disadvantage' such as being denied a loan or being fired. Harassment and hate speech laws, on the other hand, do not require a tangible disadvantage. The usage of degrading and demeaning language or acts is the harm itself. But biased genAI

---

[201] Article 3(33) AI Act.

[202] For an overview see also Frederik Zuiderveen Borgesius, Philipp Hacker, Nina Baranowska, and Alessandro Fabris, Non-discrimination law in Europe. A primer for non-lawyers (April 7, 2024), https://arxiv.org/abs/2404.08519.

language does not always rise to the level of toxicity required to warrant protection under the law.

To solve this problem inspiration can be drawn from US case law and in particular *Brown v. Board of Education of Topeka* (1954)[203] and its separate but equal doctrine. This case dealt with the question of whether racial segregation in public schools is unconstitutional under the 14th Amendment (equal protection of the laws). This doctrine stated that racial segregation is not unconstitutional so long as the public services (e.g. education) offered are equal in quality such as 'buildings, curricula, qualifications and salaries for teachers and other tangible factors'.[204] This doctrine requires a tangible difference or 'particular disadvantage' that is inflicted on one ethnic group over the other for unequal treatment to occur.

However, the court declared this doctrine unconstitutional in this case, arguing that equality goes beyond tangible differences. Referencing prior case law, the court ruled that 'those qualities which are incapable of objective measurement but which make for greatness in a law school'[205] matter and that the importance of being treated like white students and being able to engage and exchange views with other students are equally important.[206]

Despite all things being 'equal', racially separating children in grade and high school inflicts a 'feeling of inferiority as to their status in the community that may affect their hearts and minds in a way unlikely ever to be undone'[207] and that has severe knock-on effects on the learning, self-esteem, mental development and success in later life. Therefore, separate educational facilities can never be equal, said the court.[208]

With our examples of discriminatory effects of genAI there is also an intangible injury inflicted on people. Certain groups in society do not see their gender reflected in language, only see white male faces when generating pictures of CEOs, or must endure demeaning classification (e.g., Gorillas) or stereotypes (e.g., Black welfare recipients).

None of these harms necessarily lead to a tangible 'particular disadvantage'. Affected people can still use image generators or large language models. Affected people are not excluded from using the service. GenAI products and services could even be said to have the same quality for everybody. As in *Brown v. Board of Education of Topeka*, students had equal but separate educational facilities and educational support.

This is not an attempt to compare the harms generated by large language models to the harms inflicted on the Black community in the US during the era of racial segregation. We are not attempting to present these harms as being on equal footing. Rather, the logic of the landmark *Brown* case shows that it is possible to extend legal recognition to intangible discriminatory harms.

---

[203] Brown v. Board of Education of Topeka (347 U.S. 483, 1954) https://tile.loc.gov/storage-services/service/ll/usrep/usrep347/usrep347483/usrep347483.pdf [hereafter: Brown].
[204] Id., page 492.
[205] Sweatt v. Painter.
[206] McLaurin v. Oklahoma State Regents.
[207] Brown, page 494.
[208] Brown, page 383.

Discriminatory outputs of genAI largely lead to such intangible harms. The harm is found, for instance, in the constant reminder that the design of the technology was not made with a specific group in mind, that no equal consideration was given, and that providing a service that reduces the nuances of gender, ethnicity and sexuality is not a harm that society takes seriously.

In *Brown* there were no claims around harassment or hate speech, either. The harm was not done because of toxic or hateful language. The harm was done because the design of the school system signified inferiority.

Using the logic developed in *Brown* and expanded in subsequent case law, we recommend expansion of the range of harms covered by EU non-discrimination law to include the new type of biased speech harm discussed in this chapter. This applies first and foremost to genAI output, or rather: the design of these systems. Regulators could create and enforce best practice guidelines to enable developers to create systems that prevent genAI-produced discrimination.

b. Novel legal concepts for novel harms

The emergence of hard cases and partially novel types of generative harm, particularly those involving representational harm such as inadequate representation, raises the question of whether existing law needs to be expanded to include new categories of discrimination. Such practices now gain increasing relevance with genAI, but had already surfaced and been flagged before as classification models were increasingly employed in search and delivered lopsided answers to queries.[209]

While the instinct to legislate against every instance of misrepresentation is understandable, it is also crucial to acknowledge the practical limitations of such an approach. Not every slight misrepresentation of a protected group can or should be subject to legal sanction. The diversity of ethnic groups, coupled with other diversity criteria such as gender, disability, and age, makes it impractical to ensure fully equal representation for each and every protected group in every context, such as images depicting doctors or CEOs. But we can, and should, ask for more diversity than what current models often offer.[210] There are reasonable and feasible, state-of-the-art ways of reducing generative discrimination. And even more needs to be done, in research and development, to detect and mitigate such bias and raise awareness about it.

The current legal rules that require a direct link to a disadvantageous act or the establishment of a qualified toxic environment serve as filters. The rules help distinguish between instances where legal intervention is necessary and those where it might not be appropriate. This approach allows the law to focus on cases of discrimination and harassment that have tangible,

---

[209] See, e.g., Matthew Kay and others, 'Unequal representation and gender stereotypes in image search results for occupations' (2015) Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems 3819; Vivek Singh and others, 'Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms' (2020) 71 Journal of the Association for Information Science and Technology 1281; Ashleigh Rosette and others, 'The White standard: racial bias in leader categorization' (2008) 93 Journal of Applied Psychology 758;

[210] See, again, Sourojit Ghosh and Aylin Caliskan, ''Person'== Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion' (2023) arXiv preprint arXiv:231019981.

detrimental effects on individuals or groups, avoiding an overly broad application that could become unmanageable.

However, as seen, the introduction of Article 10 in the AI Act opens a potential new pathway for addressing issues of bias and representational harm. The provision's requirement to mitigate biases in training data sets doesn't necessarily align with the legal definitions of illegal discrimination (see above). Given that Article 10 and Article 26 AI Act also emphasize the importance of representativeness, this could provide a novel legal mechanism for addressing unbalanced content at its source—within the AI's training data—rather than attempting to rectify biases post-output.

3. Shaping technology

Finally, the law can also shape genAI technology more specifically, with rules on testing and auditing, randomization, inclusive content, or providing users options for inclusivity.

a. Testing and auditing for bias in generative AI

A critical first step in shaping genAI involves rigorous testing and auditing to identify and correct biases. This process would seek to ensure that the outputs of AI systems do not perpetuate stereotypes or discrimination. However, in the AI Act, such discrimination-sensitive risk management procedures are only mandatory for extremely large generative models (general-purpose AI systems with systemic risk), such as ChatGPT, or for generative systems used in high-risk sectors (see above).[211] Concerning content generation by smaller models under normal circumstances, these rules do not apply, leaving only non-discrimination law to fill the gaps. As seen, however, current law struggles to adequately address representational harms. In this respect, Article 53 AI Act, which covers all foundation models (general-purpose AI systems), should be updated to include risk assessment and mitigation concerning biased output, in the sense of incentivizing the use of state-of-the-art techniques for rendering content inclusive and representative of the overall (target) population.

b. Randomization

To ensure fair representation of protected groups, another strategy involves randomizing the selection of individuals from these groups in AI outputs. This strategy mirrors the idea of a search engine 'shuffling' search results with respect to important parameters.[212] For instance, randomly varying the depiction of genders or ethnic backgrounds in generated images or narratives over time can help achieve a more diverse and accurate representation of society. However, as mentioned, Google's Gemini model sparked debate by over-diversifying historical

---

[211] Articles 55 and 9 AI Act.

[212] Sandeep Pandey and others, 'Shuffling a Stacked Deck: The Case for Partially Randomized Ranking of Search Engine Results' (2005) Proceedings of the 31st VLDB Conference.

representations, highlighting the complexities of applying diversification through randomization.

c. Mandatory inclusive content?

Implementing post-processing filters to transform non-inclusive content into inclusive alternatives is another approach. This could include adjusting language to be more gender-sensitive. Nonetheless, this method faces challenges, such as varying opinions on what constitutes inclusive language or imagery and the specific adaptations required for different languages and legal jurisdictions. Moreover, such filters do raise significant questions with respect to free speech.[213] Such issues make specific inclusive content largely unsuitable for mandatory general law.

d. A corridor of permissible output

Training, or fine-tuning,[214] genAI consciously on a wide range of content, from standard to highly inclusive, potentially offers a more nuanced way to manage outputs. The law could mandate that genAI providers give deployers a choice: allowing deployers to select their preferred level of inclusiveness—while ensuring a base level of inclusivity and filtering out toxic or discriminatory content. Such 'displacement interpolation for language and images'[215] aims to respect deployers' preferences and societal norms while adhering to minimum legal and ethical standards.

VII. Conclusion

As genAI technologies increasingly permeate various sectors of our societies, the potential for these systems to perpetuate, and exacerbate, discrimination becomes a significant concern. This chapter has analyzed the intersection of genAI and non-discrimination law, revealing critical gaps and suggesting reforms. While we mostly rely on EU law, many conceptual discussions apply to other jurisdictions, too. Our analysis distinguishes two broad descriptive categories of discriminatory output by genAI: (i) demeaning and abusive content and (ii) inadequate representation. The first category, demeaning and abusive content, includes hate speech and harmful stereotypes. The second category involves more subtle biases like inadequate representation of protected groups in a large set of outputs unfolding, e.g., over time. For example, a genAI system may show predominantly white men when repeatedly asked to generate pictures of heart surgeons. This is often problematic.

Future research and policy-making should examine the types of content that is, and that should be, displayed for queries about, e.g., heart surgeons from various regions. Results may differ

---

[213] See, e.g., German Federal Court for Private Law, Order of July 19, 2018, Case IX ZB 10/18 (legal limits to forcing companies to make certain opinion statements, in this case a forced apology).
[214] Irene Solaiman and Chris Dennison, 'Process for adapting language models to society (PALMS) with values-targeted datasets' (2021) 34 Advances in Neural Information Processing Systems 5861.
[215] Cf. Meike Zehlike and others, 'Matching code and law: achieving algorithmic fairness with optimal transport' (2020) 34 Data Mining and Knowledge Discovery 163.

both descriptively and normatively between the US, India, and Mozambique, for example. It is an open normative question as to whether AI systems should strive to perfectly reflect statistical reality (e.g., generate pictures of white male heart surgeons 60% of the time in countries where 60% of heart surgeons are white men), or otherwise reflect a preferred normative standard (e.g., generate pictures of heart surgeons reflecting the demographic distribution in a given society or with a higher rate of historically underrepresented groups in the profession). The question is essentially whether technology should replicate existing biases or aim to represent a more equitable society.[216] While we favor the latter option, policymakers should consider and accommodate cultural and jurisdictional nuances. By critically analyzing these factors, future studies can offer insights into developing fairer and more accurate algorithms for content representation in different fields and regions of this world.

Traditional legal categories like direct and indirect discrimination or harassment are applicable to some genAI outputs that lead to disadvantageous acts or toxic communications. However, some objectionable genAI outputs do not neatly fit these legal categories because they neither involve clear disadvantageous actions nor reach a defined threshold of toxicity. The problematic outputs of genAI can be mapped on existing non-discrimination law concepts through three main types: (i) discriminatory content, (ii) harassment, and (iii) hard cases of generative harms. Discriminatory content involves outputs that lead to decisions which disadvantage protected groups, such as unfair job or credit evaluations. Inadequate representation may, but need not necessarily, amount to such content. Harassment includes speech acts or images that undermine the dignity of individuals, creating hostile environments. Discriminatory content can cross the harassment threshold.

Hard cases of generative harms, which are more complex, involve inadequate representation, propagation of harmful stereotypes, and misclassification issues. Each of these harms presents unique challenges that existing legal frameworks struggle to address effectively. In all of these contexts, we argue, if a legal violation is found, providers and deployers should be jointly and severally liable for discriminatory output by AI systems developed by providers and used by deployers.

Generative AI highlights a broader weakness of the structure, scope and implementation of non-discrimination laws. Many of the issues that generative AI brings to light have been recognized for some time, including challenges in enforcement and in gathering sufficient data to prove prima facie cases of discrimination. Additionally, there is the inherent difficulty of addressing and rectifying the underlying societal biases that are often reflected in various depictions and cultural practices. These biases, while pervasive, do not always constitute specific decisions that fall directly under the scope of non-discrimination law. Consequently, generative AI serves as a reminder of these enduring problems, and underscores the need for more effective strategies to combat discrimination in both technology and society at large.

Some of these strategies may be beyond the reach of the law. Nonetheless, we highlight some policy options that researchers and policymakers could explore. Policy options include clarifying, updating, and using law to shape technology. The AI Act mandates the mitigation

---

[216] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law' (2020) 123 W Va L Rev 735.

of biases in training and input data, establishing a proactive but vague approach to minimize genAI discrimination. More and clearer rules like this would be welcome. Additionally, policymakers should consider updating the law to recognize and address novel forms of discrimination, as genAI outputs often do not result in tangible disadvantages but can perpetuate harmful stereotypes and biases. Finally, the law can shape genAI technology through rules on testing and auditing for bias, employing randomization and mandatory inclusive content strategies, and providing users options for selecting the level of inclusiveness, thereby ensuring AI outputs do not perpetuate stereotypes or discrimination. In conclusion, generative discrimination presents difficult challenges to non-discrimination law - the proposed interpretations and adjustments mark only a first step in tackling them.