REMEDIES AGAINST DISCRIMINATORY DECISIONS TAKEN OR PREPARED BY ARTIFICIAL INTELLIGENCE SYSTEMS EUROPEAN UNION, COUNCIL OF EUROPE AND THE VENICE COMMISSION'S APPROACHES[1]

DOI: 10.62938/rdc-2025-0001

Schnutz Rudolf DÜRR
Head of Programming Department, Directorate of Programme Co-ordination at the Council of Europe; former Deputy Secretary of the Venice Commission;

Abstract

The increasing integration of AI systems into decision-making processes across both private and public sectors, driven by efficiency and cost reduction, presents significant human rights challenges, particularly concerning discrimination. As Large Language Models advance, non-technical users can easily obtain useful results, but these must be thoroughly checked for errors and AI hallucinations. Many high-risk AI applications, as defined by the EU AI Act, can directly impact human rights in areas such as education, employment, healthcare, credit assessments and even the judiciary. While human oversight is often required, it may quickly devolve into superficial plausibility checks, leading to a high risk of perpetuating bias present in AI training data.

On the basis of recommendations from the Venice Commission, this article highlights the need for individuals to have to access training data when appealing biased AI-based decisions. It suggests that specialised experts or civil society organisations, potentially using AI themselves, will likely be crucial in establishing such proof. The EU AI Act establishes a right to explanation for high-risk AI-based decisions and it mandates AI providers to maintain documentation on data type, origin, and curation methodologies of AI training. As individual decisions will typically refer to personal data, the General Data Protection Regulation (GDPR) provides additional complaints. For non-EU Member States ratifying the Council of Europe Framework Convention on Artificial Intelligence, adopting similar rules on training data documentation will be important to ensure effective complaint mechanisms.

The effectiveness of appeals against biased AI-based decisions will depend on judicial interpretation and the work of supervisory authorities in ensuring the availability of training data information. Requiring AI providers to maintain records of training data should not be viewed as burdensome "red tape" but as a critical

[1] Research for this article was partially done with the assistance of artificial intelligence. The statements made in this article are those of the author only and not of the Council of Europe

measure to enhance AI system quality and protect fundamental rights, ultimately benefiting both users and providers.

Keywords: *Artificial intelligence, discrimination / Artificial intelligence, bias, training data / Artificial intelligence, training data, retention / AI Act, remedies, AI Act, high-risk systems / GPDR, remedies / AI Convention, remedies / Venice Commission, AI bias, complaint, effective.*

Introduction

The wide availability of Large Language Models[2], leads to applications of artificial intelligence (AI) in many sectors. Progressively, more and more decisions by private and public bodies are either based on input provided by AI systems or are made directly by those systems. Eventually, even judicial decisions will be prepared on the basis of AI input. The use of AI for (the preparation of) public and private decision making can be beneficial from the viewpoint of efficiency and cost reduction. However, there is not only a general risk of error in the decisions (or judgments) prepared or taken by AI, but also a specific risk of discrimination and bias[3]. Artificial intelligence can perpetuate discrimination because of bias, which is inherent in the 'training data', which is used to teach the AI system. This risk and possible means for identification and mitigation of such vias are examined below.

In the absence of effective self-regulation[4], the risks of Artificial Intelligence have been recognised at the European level, in the European Union and the Council of Europe.

Article 5 of the Artificial Intelligence Act of the European Union (hereinafter 'the AI Act')[5] prohibits a series of AI applications that can do serious harm, e.g. subliminal manipulative or purposefully deceptive techniques distorting behaviour, untargeted scraping of facial images from the internet or CCTV footage for facial recognition or

---

[2] Rohit Patel, Understanding LLMs from Scratch Using Middle School Math, A self-contained, full explanation to inner workings of an LLM, Towards Data Science, 19/10/2024 (https://towardsdatascience.com/understanding-llms-from-scratch-using-middle-school-math-e602d27ec876).

[3] Frederik Zuiderveen Borgesius, Discrimination, artificial intelligence and algorithmic decision-making, Council of Europe, 2018 (https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73).

[4] Abigail Dubiniecki, Trustworthy AI: String Of AI Fails Show Self-Regulation Doesn't Work (https://www.forbes.com/sites/abigaildubiniecki/2024/01/25/trustworthy-ai-string-of-ai-fails-show-self-regulation-doesnt-work/), Melissa Heikkilä, How's AI self-regulation going? One year on from the White House's voluntary commitments on AI, MIT Technology Review, 23/07/2024 (https://www.technologyreview.com/ 2024/07/23/1095218/hows-ai-self-regulation-going/); Gary Drenik, The Pitfalls Of AI Self-Regulation, 22/10/2024 (https://www.forbes.com/sites/garydrenik/2024/10/22/the-pitfalls-of-ai-self-regulation/).

[5] Regulation (EU) 2024/1689 of the European Parliament and the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain regulations ('the AI Act').

the use of biometric systems categorising individuals by race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation[6].

However, a look at the wide range of AI systems, which Article 6 of the EU AI Act identifies as 'high-risk', but which are admissible, shows that such systems can thoroughly affect the lives of individuals. We can expect decisions made by AI or with the assistance of AI in areas where discrimination deriving from bias can be a significant danger for individuals affected by those decisions. According to Annex III of the AI Act, high-risk systems include *inter alia* the following fields: biometrics (remote identification systems, biometric categorisation or emotion recognition), education (school admission procedures, evaluation of exam results, detecting prohibited behaviour – cheating – during exams), employment (targeted job advertisements, recruitment, promotion, dismissal), essential public and private services (social benefits and allocations, creditworthiness, risk assessment and pricing for life and health insurance), law enforcement (assessment of risk to be a victim or an offender, evaluation of evidence), migration, asylum and border control (examination of applications for asylum, visa or residence permits), and even the judiciary (researching and interpreting both facts and the law and applying the law to a concrete set of facts – legal subsumption).

AI decision making, even if it is applied in areas below the high-risk threshold can result in erroneous or discriminatory decisions. Apart from questions of copyright[7], dangers from AI systems stem from errors, either because of incomplete or incoherent data input, or because of so-called hallucinations[8], when the system provides an answer that sounds reasonable, but that is not based on the applicable facts and legislation.

Even if a final decision is in the end taken by a human on the basis of input from AI, erroneous AI results can sound so convincing and coherent that the human will often confidently follow the recommendations made by the system[9]. Furthermore, in practice, the human in charge of making the final decision (civil servant or judge in the

---

[6] On 4 February 2025, the EU Commission published guidelines on prohibited artificial intelligence (AI) practices (https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act).

[7] Hayden Field, Anthropic will face a class-action lawsuit from US authors, The Verge, 17/07/2025 (https://www.theverge.com/anthropic/709183/anthropic-class-action-lawsuit-pirated-books-authors-downloads); Jeffrey M. Kelly, Scott N. Sherman, Adrianne Cleven, Matt Gorga, From Copyright Case to AI Data Crisis: How The New York Times v. OpenAI Reshapes Companies' Data Governance and eDiscovery Strategy, 10/07/2025 (https://www.nelsonmullins.com/insights/blogs/corporate-governance-insights/all/from-copyright-case-to-ai-data-crisis-how-the-new-york-times-v-openai-reshapes-companies-data-governance-and-ediscovery-strategy); Ignacio de Gregorio, Why This ChatGPT Court Win is Historic. A Death Sentence to Copyright Laws? 11/11/2024 (https://medium.com/@ignacio.de.gregorio.noblejas/why-this-chatgpt-court-win-is-historic-997f75927380).

[8] Vitaly Kukharenko, Why Do Neural Networks Hallucinate (And What Are Experts Doing About It)? Towards AI, 11/11/2024 (https://pub.towardsai.net/why-do-neural-networks-hallucinate-and-what-are-experts-doing-about-it-7b9342605bf7).

[9] Katie Szilagyi, Regenerating Justice: ChatGPT and the Legal Minefield of Generative AI, 01/10/2024 (http://dx.doi.org/10.2139/ssrn.4976738); Krook, Joshua and Schneiders, Eike and Seabrooke, Tina and Leesakul, Natalie and Clos, Jeremie, Large Language Models (LLMs) for Legal Advice: A Scoping Review (http://dx.doi.org/10.2139/ssrn.4976189).

public law field) may have to deal with so many individual cases that blindly following the proposals made by the AI system may be the only way to cope with the high workload imposed to achieve cost reductions[10].

Appeals against discriminatory AI based decisions face considerable hurdles. As concerns transparency, there are two parts of the problem: does the individual even know that a decision was taken or prepared by an AI system and, if the person has that knowledge, how did the AI system has come up with the results[11]?

Transparency is essential; even if only part of the preparation of the decision, for instance research, builds on information provided by an AI system, as part of good practice, this should be signalled to the person subject to the decision, but this is not a right in all cases. Article 50 of the AI Act setting out transparency obligations is not explicit on the right to be informed that a decision affecting a person was made using an AI system. At least for high-risk systems that right can be inferred from Article 86 of the AI Act because the right to an explanation of individual decision-making (see below) requires as a pre-condition that the person concerned knows that a given decision is AI based.

Further, even if the individual subject of the AI based decision is aware of this fact, in an appeal against that decision it may be very difficult for the individual to find evidence that the decision was biased and discriminatory[12].

This challenge of the effectiveness of remedies against the use of AI stems from the fact that typically AI systems are "black boxes"[13]; they are trained on vast amounts of training data and usually the providers of AI systems themselves cannot explain in detail why their system gives a specific answer[14]. AI systems make sophisticated

---

[10] An excessive workload can be a serious problem for judges even in contexts not related to AI, see for instance Venice Commission, CDL-AD(2017)004, Turkey – Opinion on the duties, competences and functioning of the criminal peace judgeships, para. 24.

[11] Gabriele Spina Alì / Ronald Yu, Artificial Intelligence between Transparency and Secrecy: From the EC Whitepaper to the AIA and Beyond, European Journal of Law and Technology (https://www.ejlt.org/index.php/ejlt/article/download/754/1044/3716); Georgie Gould, Understanding AI decision-making: Research examines model transparency, TechXplore, 19/02/2025 (https://techxplore.com/news/2025-02-ai-decision-transparency.html); European Data Protection Supervisor, TechDispatch on Explainable Artificial Intelligence, 16/11/2023 (https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en).

[12] Ivana Bartoletti / Raphaële Xenidis, Study on the impact of artificial intelligence systems, their potential for promoting equality, including gender equality, and the risks they may cause in relation to non-discrimination, Council of Europe, 2023 (https://edoc.coe.int/en/artificial-intelligence/11649-study-on-the-impact-of-artificial-intelligence-systems-their-potential-for-promoting-equality-including-gender-equality-and-the-risks-they-may-cause-in-relation-to-non-discrimination.html).

[13] Valerie du Preez, et al. From bias to black boxes: understanding and managing the risks of AI – an actuarial perspective, Published online by Cambridge University Press, 11/04/2024 (https://www.cambridge.org/core/journals/british-actuarial-journal/article/from-bias-to-black-boxes-understanding-and-managing-the-risks-of-ai-an-actuarial-perspective/9F06CF4E8EFB555202B2AA4ED78DD716).

[14] Researchers from Anthropic looked into the reasoning of the Claude 3.5 Haiku AI system and identified methods of reasoning that are very different from that of humans; they saw how incorrect 'guessing' resulted in hallucinations and how the model does calculations by way of approximation while explaining that it had come

statistical operations, which depend on a high number of parameters, which in turn depend on the data, which was used to train the system.

This means that often even the provider of the AI system cannot explain why a certain output was produced in a given context. As a consequence, for an individual subject to an AI based decision it may be very difficult to establish a link between the way that the system was trained and the decision affecting him- or herself.

Because of these risks, the scope of rules of evidence in appeals against decisions, which were taken by an AI system or where AI provided input for a decision taken by a human should be adapted in judicial practice to these challenges and the review of such decisions should allow for additional safeguards, as compared to decisions taken by humans only.

Such additional scrutiny will reduce the efficiency and cost savings achieved through the use of AI, but this is a necessary corollary of partially 'outsourcing' decision making power from humans to AI.

In the following, we will identify provisions that are relevant for appeals against decisions taken by or on the basis of AI in EU legislation and Council of Europe texts, before looking into the Venice Commission's proposal on AI training data retention.

EU legislation

General Data Protection Regulation (GPDR)

Well before the 2024 AI Act, Article 22 of the EU General Data Protection Regulation (GPDR)[15] already provided that the data subject "shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her". The term 'automated decision making' of course applies also to decisions taken by AI.

However, Article 22 (2) GPDR allows exceptions in case of a contract between the data subject and a data controller, authorisation by a law "to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests" or when it is based on explicit consent. In case of public decisions taken by an AI system, the exception of authorisation by law would probably be applicable. Elements to ensure that such a law also contains 'suitable measures' are part of the discussion in this article.

---

to the results using standard calculation methods; summary: https://www.anthropic.com/research/tracing-thoughts-language-model; papers: Jack Lindsey et. al., On the Biology of a Large Language Model, 27/03/2025 (https://transformer-circuits.pub/2025/attribution-graphs/biology.html); Emmanuel Ameisen et al., Circuit Tracing: Revealing Computational Graphs in Language Models, 27/03/2025 (https://transformer-circuits.pub/2025/attribution-graphs/methods.html).

[15] Adopted on 14 April 2016, to become effective on 25 May 2018 (https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en).

In its Opinion on the Legal Protection of Citizens, requested by the Netherlands[16], the Venice Commission discussed the exceptions to Article 22 of the GDPR and concluded that the use of (double) citizenship to identify migrants had not been covered by such an exception[17].

Even if the right not to be subject to automated data processing were excluded under these exceptions, a discriminatory decision taken by an AI system or with the help of an AI system could also violate other rights under the GPDR. These rights include transparency – Article 12, information on data collection and origin – Articles 13 and 14, right of access to data – Article 15, right to rectification – Article 16, right to be forgotten – Article 17, right to a restriction of data processing – Article 18, right to notification – Article 19, right to data portability – Article 20 or the right to object – Article 21. If any of these rights were affected, complaints under the GPDR should be possible.

The GDPR introduces a specific remedy: when a person (data subject) is of the view that his or her rights under the GPDR – Article 22 or another right – have been violated, Article 77 of the GPDR provides for a complaint to a supervisory authority in the EU Member State of the residence, place of work or the place where the infringement occurred. A decision of that authority can in turn be appealed to court (Article 78). A consistency mechanism linking supervisory authorities (Article 63) should ensure that the GDPR is applied coherently among Member States also as concerns appeals[18].

Even without such a prior complaint to a supervisory authority, a person (data subject) has a right to a direct judicial remedy against the data controller or processor.

---

[16] As a follow-up to that opinion in the field of AI, the Dutch Parliament informed the Commission that it had addressed the use of Artificial Intelligence when making administrative decisions, several new policies and guidelines were being elaborated to ensure that AI technology did not lead to discrimination and respects privacy and other fundamental rights (https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL-PL-PV(2022)003-bil).

[17] Venice Commission, Netherlands – Opinion on the Legal Protection of Citizens (CDL-AD(2021)031): "95. As concerns EU Law, Article 22 of the GDPR on the conditions for the use of algorithms for making decisions with legal effects is relevant in this case. As a main rule, a (natural or legal) person has the right not to be the subject of a decision based exclusively on automated processing significantly affecting him or her. However, there are three exceptions. Two of them are based on the explicit consent of the person or on the conclusion of a contract that requires the use of an algorithm. The third exception corresponds to cases where applicable EU or that of the Member State authorizes the exclusive use of an algorithm to base the decision in question. In such a case, appropriate measures must be provided for in order to safeguard the rights and freedoms and the legitimate interests of the person affected concerned. Furthermore, no decision adopted with the exclusive use of an algorithm may be based on "sensitive" personal data (data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs or trade union membership, as well as genetic, biometric, health data or data relating to life or sexual orientation of somebody).

96. In the Netherlands, the use of nationality as possible criteria for identifying fraud would seem to fall under these prohibitions. Discriminatory practices were systematised through algorithms".

[18] Detlev Gabel / Tim Hickman, Chapter 16: Remedies and sanctions – Unlocking the EU General Data Protection Regulation, While & Case, 28.05.2025 (https://www.whitecase.com/insight-our-thinking/chapter-16-remedies-and-sanctions-unlocking-eu-general-data-protection).

Such proceedings shall be brought in the Member State where the data controller or processor has an establishment or in the Member State where the data subject has a residence, unless the data controller or processor is a public body acting within it powers. Such appeals entitle also to compensation (Article 82).

For the topic of appeals against discriminatory decisions taken by AI or on the basis of AI input, remedies under the GPDR are relevant because in most cases a decision affecting an individual will also involve personal data of that person. Even when the specific right not to be subject to automated data processing of Article 22 GPDR does not apply in a given case, if there is a violation of another right under the GPDR, the individual can have recourse to the specific complaint under Article 77 GPDR or s/he can bring a direct claim to court against the data controller or processor.

Artificial Intelligence Act (AI Act)

In order to address the specific risks of Artificial Intelligence, the European Union adopted the AI Act[19] on 17 June 2024 and it entered into force on 1 August 2024. It will go into effect in stages[20]. The AI Act creates obligations for providers and deployers of AI systems in the single market. It imposes specific requirements on AI systems (e.g., transparency, technical documentation), which can serve as evidence in court proceedings.

After excluding some AI applications in its Article 5, the Act distinguishes between various levels of risks, notably high-risk systems (criteria listed in Annex III of the Act) and obliges the providers and deployers of AI systems *inter alia* to establish a risk management system with appropriate data governance and record keeping.

Under the AI Act providers of AI systems, both as concerns high-risk systems and general-purpose AI systems, shall keep and provide documentation including on the training data used. This is relevant in the present context of appeals based on the allegation that the data used for the training of the system was biased.

Article 11 of the AI Act provides that for high-risk AI systems a detailed technical documentation has to be drawn up. Annex IV, which lists the elements of information to be contained in that technical documentation, requires that documentation include where relevant, descriptions of "the training methodologies and techniques and the training data sets used, including a general description of these data sets, information about their provenance, scope and main characteristics". The documentation shall

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

[20] Following the recent publication of the General-Purpose AI Code of Practice, rules on general purpose AI should go into effect on 2 August 2025 (https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1787); there is resistance against this voluntary code (https://www.euronews.com/next/2025/07/10/eu-waiting-for-companies-to-sign-delayed-ai-code; Ram Iyer, Meta refuses to sign EU's AI code of practice, TechCrunch, 18/07/2025 (https://techcrunch.com/2025/07/18/meta-refuses-to-sign-eus-ai-code-of-practice/).

show how the data was obtained and selected; as well as labelling procedures and data cleaning methodologies (curation).

While this detailed documentation is required for high-risk AI systems, Article 53 sets out somewhat less stringent documentation requirements for general purpose AI systems. By reference to Annexes XI and XII of the AI Act, Article 53 obliges providers of those systems to draw up and keep up to date the technical documentation of the model, including its training and testing process and the results of its evaluation. This documentation shall be provided (a), upon request, to the AI Office and the national competent authorities and (b) to providers of AI systems who intend to integrate the general-purpose AI model into their AI systems (downstream providers). The documentation shall notably include "information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies"[21].

This means that even for general purpose AI systems, the requirements for documentation on training data are quite similar to that for high-risk systems.

Under the AI Act, any natural of legal person can lodge a complaint with a market surveillance authority according to Article 85 of the AI Act if they allege a violation of the AI Act. These authorities are set up by the Member States and investigate complaints and conduct market surveillance activities. Individuals can appeal to court against a decision of a market surveillance authority if there was a prior complaint to this authority.

A specific right to explanation of a decision which produces legal effects or similarly significantly affects an individual taken on the basis of the output of an AI system, exist only for high-risk systems (Article 86), but many human rights relevant cases will be high-risk applications. The results of such an explanation could subsequently be used in a challenge in court of the decision taken.

In addition to these specific remedies, while the AI Act focuses heavily on regulatory oversight through market surveillance authorities, individuals can bring direct legal action against providers (developers) of AI systems in national courts.

The AI Act places significant emphasis on 'high-risk' AI systems due to their potential to cause harm. Many of the specific requirements and individual rights (like the right to explanation) apply primarily to these systems. If an AI system processes personal data, both the AI Act and the GDPR will apply, and individuals have access to remedies under both regulations. A specific AI Liability Directive is still under development[22], but it seems to be facing strong resistance and might be withdrawn[23].

The obligation to provide technical documentation to AI supervising authorities is important also if it is read together with the obligation of the State under Article 6§1

---

[21] Annex XI, Section 1, paragraph 2.c and identically Annex XII, paragraph 2.c.

[22] https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en.

[23] Louise Freeman, Lisa Peets, Marty Hansen & Madelaine Harrington, The Future of the AI Liability Directive, 10 March 2025 (https://www.insideglobaltech.com/2025/03/10/the-future-of-the-ai-liability-directive/.

of the European Convention on Human Rights to disclose documents that are relevant in administrative proceedings. According to the European Court of Human Rights, administrative authorities have an obligation to give access relevant documents which are in their possession, if necessary, via a procedure for the disclosure of documents. In its case *McGinley and Egan v. the United Kingdom,* the Court found a violation of Article 6 § 1 of the Convention because the UK had prevented the applicants from gaining access to, or falsely denied the existence of, documents in its possession relevant to the proceedings[24].

Applied to appeals against discriminatory AI based decisions, this could mean that at least in proceedings against AI based administrative decisions, the State would be under an obligation to provide information it holds on the functioning of that AI system. This should normally include information on the training data which is part of the documentation submitted by the AI provider. The applicant could use this information to establish that the training data was indeed biased. This obligation to deliver the technical documentation seems straightforward in administrative cases directed against the State. It is less obvious in direct proceedings against the provider of the AI system. In such cases, the obligation under of the Council of Europe Framework Convention on Artificial Intelligence to ensure an 'effective' complaint to the authorities might possibly become a basis to obtain information on the AI training data that can then be used in direct proceedings against the AI system provider.

Council of Europe

Framework Convention on Artificial Intelligence

The 2018 study on Artificial Intelligence, and Algorithmic Decision Making published by the Council of Europe's Commission Against Racism and Intolerance (ECRI) concluded that AI decision-making brings risks and can have discriminatory effects, for instance when an AI system learns from data reflecting biased human decisions. But AI also paves the way for new types of discrimination that do not correlate with typical protected characteristics, such as racial origin, or sexual orientation[25]. As the way of 'thinking'[26] of AI is very different from that of human reasoning, such discriminations could even be based on random numerical criteria, for instance relating to personal identification numbers.

[24] ECtHR, *McGinley and Egan v. the United Kingdom*, 1998 (10/1997/794/995-996), §§ 86 and 90.

[25] Frederik Zuiderveen Borgesius, Discrimination, Artificial Intelligence, and Algorithmic Decision Making, Council of Europe, 2023 (https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/ 1680925d73).

[26] To the extent that AI does any real thinking: Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, Mehrdad Farajtabar, The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity, 06/2025 (https://machinelearning.apple.com/research/ illusion-of-thinking).

After a long preparatory phase, on 17 May 2024 the Committee of Ministers of the Council of Europe adopted the Framework Convention on Artificial Intelligence (CETS no. 225, hereinafter the 'AI Convention'), which was opened for signature on 5 September 2024[27]. The AI Convention will enter into force with the ratification by five signatories, which should include three Council of Europe Member States. So far (July 2025), it has been signed by 15 individual States and the European Union. The signature by the EU[28] commits all 27 EU Member States as the Convention covers a non-shared single market competence, which does not require a mixed agreement. The sole ratification by the European Union will bring the Convention into force for all EU Member States.

Equality and non-discrimination is one of the key principles of the AI Convention. Its Article 10 obliges the parties to "adopt or maintain measures with a view to ensuring that activities within the lifecycle of artificial intelligence systems respect equality, including gender equality, and the prohibition of discrimination, as provided under applicable international and domestic law". As to the relevant international law, the Explanatory Report to the AI Convention points out that the intention was to refer to the body of existing human rights, such as – on the global level – the International Covenant on Civil and Political Rights, the International Covenant on Economic, Social and Cultural Rights, other UN treaties, – on the regional level – Article 14 of the ECHR and its Protocol No. 12, other Council of Europe conventions, the EU Charter of Fundamental Rights, the EU Treaties or the Pact of San José (as the Convention is open to non-European States).

As concerns documentation, Article 14 of the AI Convention obliges parties to take "measures to ensure that relevant information regarding artificial intelligence systems which have the potential to significantly affect human rights and their relevant usage is documented, provided to bodies authorised to access that information and, where appropriate and applicable, made available or communicated to affected persons". There is no specific reference to AI training data, but due the very nature of artificial intelligence, such information may well be considered 'relevant' under Article 14.

The Framework Convention also requires the availability of remedies: Article 14 (2)(c) provides that each party shall adopt or maintain measures including […] an effective possibility for persons concerned to lodge a complaint to competent authorities". This requirement is somewhat qualified by the clause in paragraph 1, "to the extent remedies are required by its international obligations and consistent with its domestic legal system". However, under Article 10 above, we have seen the wide range of

[27] The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence).

[28] Andorra, Canada, European Union, Georgia, Iceland, Israel, Japan, Liechtenstein, Montenegro, Norway, Republic of Moldova, San Marino, Switzerland, Ukraine, United Kingdom, United States of America (status 20 July 2025).

international obligations relating to equality and non-discrimination, most of which do ensure remedies.

Even if it is not explicit as concerns documentation of AI training data, the AI Convention may provide a basis for remedies against AI based decisions that are biased and discriminatory and violate human rights. To ensure that the documentation of AI systems also enables an 'effective' remedy under Article 14 of the AI Convention, the views of the Venice Commission may be pertinent.

Venice Commission

The Venice Commission[29] worked on the use of digital technologies and artificial intelligence mostly in the field of elections[30]. In 2024, the Venice Commission adopted an interpretative declaration of the Code of Good Practice in Electoral Matters on digital technologies and artificial intelligence covering the areas of freedom of information, data protection and access to (social) media. This interpretative declaration calls for a possibility to challenge before an independent body both the process of adoption of digital and AI tools by the electoral management body and the concrete decisions taken by or on the basis of a recommendation of an artificial intelligence system[31]. Justified trust in any digital or AI systems used in the electoral field is of course crucial for this key area of democracy.

In view of a concrete problem of mass discrimination caused *inter alia* by the application of an AI tool in the Netherlands, the Venice Commission made a recommendation on how remedies against such decisions could be made effective. In 2021, the President of the House of Representatives of the States General of the Netherlands requested an opinion from the Venice Commission in connection with the so-called childcare allowance affair. This case had caused serious problems for a large number of parents with a migrant background, as they were faced with massive claims for reimbursement of childcare allowance, often dating back several years. Appeals were systematically rejected by the Administrative Jurisdiction Division of the Council of State (the supreme administrative court of the Netherlands).

---

[29] For a presentation of the Venice Commission, see Schnutz Dürr, The Venice Commission, in Kleinsorge, Tanja, Council of Europe, in Wouters, J., International Encylopaedia of International Laws: Intergovernmental Organizations, Alphen aan den Rijn 2023, pp. 151-163; Thomas Markert, The Venice Commission of the Council of Europe: From Advisory Body to Actor in the Defence of the Rule of Law and Democracy (1990-2022), Human Rights Law Journal, Vol. 42(2022), No.10-12, pp. 341-369; Simona Granata-Menghini, Ziya Tanyar (eds), Venice Commission: thirty-year quest for democracy through law, Lund (2020), CDL-PI(2020)013.

[30] Venice Commission, CDL-AD(2020)037, Study – Principles for a fundamental rights-compliant use of digital technologies in electoral processes; CDL-AD(2019)016; Joint Report of the Venice Commission and of the Directorate of Information Society and Action against Crime of the Directorate General of Human Rights and Rule of Law (DGI), on Digital Technologies and Elections; CDL-PI(2018)011, Compilation of Venice Commission opinions and reports concerning digital technologies in the electoral process.

[31] Venice Commission, CDL-AD(2024)044 Interpretative declaration of the Code of good practice in electoral matters as concerns digital technologies and artificial intelligence.

One aspect of the numerous recommendations made in this Opinion[32] concerned the detection of systematic bias in the artificial intelligence system that had been used to identify the recovery claims (the parents concerned had a migrant background). The opinion recommended the mandatory storage of AI training data, which can later be used in appeal proceedings to identify such bias. The Opinion stated that:

"97. In the future, more sophisticated artificial intelligence algorithms are likely to be used and it will be much harder to identify which criteria were used by these algorithms due to the very nature of learning of modern AI systems. Detecting bias in such system can be next to impossible as self-learning AI systems are fed with large amounts of training data. This data comes from the real world; it aggregates individual decisions made by humans. However, in part these past decisions made by humans may have already been made on a discriminatory basis. On the other hand, such bias could normally not be discovered without such aggregation. Therefore, AI also presents an occasion to review past practices, and this should be used to identify bias in administration.

98. Bias and discrimination in AI originate from training data, which could be indicative of real past discrimination. This discrimination can be perpetuated in the AI system. When bias is detected, the system should be re-trained. As a prudent measure, the Commission suggests that when AI systems are introduced, the anonymised training data is preserved to be able to detect such bias at a later stage. Possibly fully anonymised training data could even be made public to enable research for bias in that data".

The Venice Commission thus set out a method to enable the identification of discriminatory AI based decisions. The data used for the training of the AI system should be retained in order to be able to detect bias at a later stage, including during appeals proceedings. Ideally this data should be made available for research to detect any bias even before discriminatory decisions are taken.

Training data retention as a means to enable effective review of AI based decisions

Ability to identify biased training data

It is clear that a judge or appeal body will most likely not be able to interpret training data, but the judge or body can request an expert opinion for that purpose. Notably if the appeal is decided years after the AI system was established, more recent and more powerful AI systems could be able to easily identify parts of the training data that contained the bias.

[32] Venice Commission, CDL-AD(2021)031 Netherlands – Opinion on the Legal Protection of Citizens.

Let us imagine an example from a typical field of AI based decision-making,[33] if an AI system that was trained to decide on car traffic offences was fed tens of thousands of prior decisions, an appeal against a decision taken based on that system might allege that certain groups of persons, for instance migrants, would have received harsher penalties as compared to the rest of the population. As part of that appeal procedure, AI experts, maybe specialised NGOs, might be able to identify a part of training data, i.e. prior decisions, which indeed systematically handed down harsher penalties for that group. Such a proof would result in a favourable result for the applicant in the instant case.

What is even more important than the instant case, is the identification of bias the in the AI system as such. As a consequence of the appeal, the AI provider should be asked[34] to update their system. This can be done by removing the part of the training data that was identified as being biased and then by retraining the system with the (large) remainder of the training data. Until that update is available, decisions affecting the discriminated group should be checked manually.

In practice, a major question will of course be costs. Providers of AI systems will probably insist that training data retention will lead to significant additional costs, given the huge amount of data concerned. Current Large Language Models are indeed trained with nearly the whole of the internet. However, it may not be necessary to separately retain all training data. We have seen that the AI Act obliges AI providers to produce documentation on the source and origin of training data and the methods used to curate it. For publicly available data, this information may be enough to enable the identification of bias. Full data retention may be necessary for synthetic data[35] or data that is not available publicly, for instance the road traffic decisions in our previous example. Compared to the huge amount of public data used to train Large Language Models, it should be possible to retain this part of specialised data at a reasonable cost which would eventually become part of the price calculation for new AI systems.

Scope of training data to be retained

Referring to training data, we should point out that often large language models, such as ChatGPT, Gemini, Claude or Mistral serve as the basis for the development of

---

[33] Zhi (Aaron) Cheng , Zhanyu Dong , Min-Seok Pang, Automated Enforcement and Traffic Safety, informsPubsOnline, 15/04/2025 (https://pubsonline.informs.org/doi/10.1287/mnsc.2023.00575); Angelina Grace, Automated Traffic Violation Detection Using AI and Machine Learning 03/2025 (https://www.researchgate.net/publication/389705463_Automated_Traffic_Violation_Detection_Using_AI_and_Machine_Learning); Compte d'affectation spéciale, Mission interministérielle, Projet annuel de performances, Annexe au projet de loi de finances pour 2025, Projet annuel de performances Annexe au projet de loi de finances pour 2025, p. 19 (https://www2.assemblee-nationale.fr/static/17/Annexes-DL/PLF-2025/Controle_circulation_routiers.pdf).

[34] The fault may not be that of the AI provider; it may well have been the AI deployer (user), in our example the road traffic administration, that provided the training data, which included the biased decisions.

[35] Moez Ali, Synthetic data is the future of Artificial Intelligence – Overcoming the Limitations of Real-World Data, Medium, 18/03/2023 (https://moez-62905.medium.com/synthetic-data-is-the-future-of-artificial-intelligence-6fcfd2ce1a14); Robert Riemann, Synthetic Data, European Data Protection Supervisor (https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en).

specialised AI systems, to acquire knowledge in specific areas. Currently, two main techniques are used to teach the systems about the specific knowledge of their area of deployment: Fine-tuning and Retrieval-Augmented Generation (RAG) systems[36].

Even if the word Fine-tuning hints to minor modifications, it means adding an additional layer of training and adding weights to the system. This additional training thus adds domain specific data to the original general training data. As a consequence, for the purpose of detection of bias, that additional data will be included in wider the term "training data", including in the meaning of the term as used in the AI Act. The additional, domain specific part of the training data is of course of particular interest for the detection of bias. Coming back to the above example, the general training data teaches the AI system how to reason and to 'talk', whereas the additional data used for Fine-tuning will be prior road traffic decisions. An examination of this data may be sufficient to detect bias.

The other technique is RAG[37], which does not affect the AI system as such, but which adds data at the moment of prompting (asking) the system, by feeding the domain specific information, for instance on prior road traffic offences, into the context window of the AI system. This data is kept in a separate binary vector database. Even if preparing a domain specific RAG vector database is complex on its own, it is usually cheaper than Fine-tuning a whole AI system[38].

For RAG systems, the question arises if this domain specific data is part of the training data. In a narrow technical sense, adding data as part of the prompt would normally not be part of the training, which has a permanent impact on weights in the system itself. From a wider, purposive viewpoint, I would however argue, that this domain specific data should be considered as being covered by the obligation to document AI systems, because it determines how the system will behave in the given domain. Being present in the context window of the system, this data is at the centre of the 'reasoning'[39] of the system.

Conclusion

Since the breakthrough of Large Language Models, AI systems make steady progress and the easy prompting (asking) of such systems enables non-technicians to obtain

---

[36] Sarah Schürch, How to Make Your LLM More Accurate with RAG & Fine-Tuning. And when to use which one, 11/03/2025 (https://towardsdatascience.com/how-to-make-your-llm-more-accurate-with-rag-fine-tuning/).

[37] Patrick Lewis, et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 22/05/2020 (https://arxiv.org/abs/2005.11401).

[38] Heiko Hotz, RAG vs Finetuning – Which Is the Best Tool to Boost Your LLM Application? The definitive guide for choosing the right method for your use case, 24/08/2023 (https://towardsdatascience.com/rag-vs-finetuning-which-is-the-best-tool-to-boost-your-llm-application-94654b1eaba7/).

[39] Jim Canary, AI Tokens Explained: Understanding Context Windows and Processing, 07/03/2025, https://medium.com/@jimcanary/ai-tokens-explained-understanding-context-windows-and-processing-a99ca2dd9142.

practical and often highly useful results, even if they should be checked to detect errors and AI hallucinations.

For reasons of efficiency and cost reduction, more and more decisions both in the private and the public sectors will be prepared by or be based on information from AI systems. The long list of permissible high-risk AI systems of the EU AI Act includes many fields that affect human rights, such as school admission, evaluation of school results, job recruitment, promotion and dismissal, healthcare benefits, creditworthiness assessments and even interpreting facts and the law in the judiciary. Even if such decisions are in the end taken or approved by a human, such control will, in practice, often remain limited to a simple plausibility check. After an initial phase of close scrutiny, humans will often simply click “approve” because draft decisions will seem plausible.

In view of the wide scope of areas affecting human rights, there is a danger that part of the AI based decisions are discriminatory and that this discrimination is based on biased training data. AI systems learn from real world data that are fed to them. This data can be systematically biased and, for instance in administrative decisions, such bias may never have been revealed and corrected in appeal proceedings. Biased training data will result in a perpetuation of bias and discrimination.

Individuals who appeal against biased AI based decisions need access to the training data to be able to identify that bias. In practice, it will not the individual him- or herself establishing that proof, but they may be assisted by experts or by civil society associations specialising in detecting such bias, possibly themselves using other AI systems for that purpose.

In order to have access to such training data, it is essential that the AI providers (producers) keep records of the data fed to the system during training as had been recommended by the Venice Commission in its Opinion on the Legal Protection of Citizens for the Netherlands.

The EU AI Act does require AI providers to establish and keep up to date documentation, including information on the type and origin of data used for training and curation methodologies. It is true that scrutiny under the EU AI Act is most stringent for high-risk systems, but it is also in these areas that the danger of human rights violations is the highest.

Under the AI Act, it will be for supervisory authorities to ensure that the documentation that AI providers have to submit, contains sufficient information on the training data to enable the identification of bias.

Typically, AI systems taking or preparing decisions in individual cases will involve personal data and therefore the provisions of the General Data Protection Regulation will apply in addition to those of the AI Act. As a consequence, specific appeal procedures will be available, notably to complaint to a supervisory authority under the GPDR and the right to explanation of an AI based decision under Article 86 of the AI Act.

Article 14 of the Council of Europe Framework Convention on Artificial Intelligence provides for an effective possibility to lodge a complaint. It does not contain specific rules on the documentation of training data. Therefore, for non-EU Member States that ratify the CoE Framework Convention, its will be important to adopt similar rules in order to live up to the requirement to make such a complaint fully "effective".

In the end, the effectiveness of appeals against AI based decisions will depend on the interpretation of these rules by the courts. Of course, the preventive work of supervisory authorities, under the GDPR or the AI Act, or comparable authorities in non-EU Member States, will be essential to ensure that information on training data is available at the moment when it will be needed in appeal proceedings.

The current focus of the EU on competitiveness, removing so-called "red tape", may negatively affect rules established to protect human rights. Obliging AI providers to keep information on training data should not be seen as red tape but an important means to improve the quality of AI systems, which eventually benefits also the providers.