COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Strasbourg, 21 October 2022

GEC(2022)9
CDADI(2022)21

GENDER EQUALITY COMMISSION (GEC)

and

STEERING COMMITTEE ON ANTI-DISCRIMINATION, DIVERSITY AND INCLUSION (CDADI)

Preliminary draft Council of Europe study on the impact of artificial intelligence, its potential for promoting equality, including gender equality, and the risks to non-discrimination

Document prepared by
Ivana Bartoletti,
Global Chief Privacy Officer at Wipro, Visiting Policy Fellow at the Oxford Internet Institute, University of Oxford and co-founded the Women Leading in AI Network

and

Raphaële Xenidis,
Lecturer in EU law, University of Edinburgh,
School of Law and Marie Curie Fellow, iCourts, University of Copenhagen.

Table of Contents

Executive summary

Introduction: The context

Artificial intelligence (AI) is everywhere. Often acclaimed for its ability to reduce friction and simplify previously manual and time-consuming processes, AI research continues to hurtle down the scientific highway, crossing frontiers and changing the way people live their lives.

In healthcare, the automation of medical diagnosis could make complex services like breast cancer screening and MRI scans function as a walk-in service. This would enable dangerous diseases to be diagnosed in greater volumes and at a much earlier stage. Smart cities can support better management of traffic and allocation of resources, and large-scale data analysis can optimize resources for our environment. AI is also increasingly relied upon as an information and decision-making tool in the world of government and public policy, from housing and healthcare to education and criminal justice.

Over recent years, the potential to greatly benefit people has been somewhat eclipsed by the growing awareness of a downside: the potential for the *softwarisation*[1] of existing discrimination and inequality. For example, in what the Dutch have dubbed the "toeslagenaffaire", or the childcare benefits scandal, thousands of people have suffered the consequence of a biased self-learning algorithm that created risk profiles in an effort to spot childcare benefits fraud. The victims of this case of algorithmic profiling experienced distress and increased poverty, even leading to a case of attempted suicide.[2] A parliamentary report into the childcare benefits scandal found several grave shortcomings, including institutional biases and authorities hiding information or misleading the parliament about the facts.[3]

In 2018, Reuters reported that Amazon tried to use AI to build a resume-screening tool by using resumes that the company had collected over the previous decade.[4] As these resumes came mostly from men, and as the consequences of that fact were not seriously thought through, the new system discriminated against women and had to be discarded. In 2019, the Apple-branded credit card came under intense scrutiny because women were receiving less credit than their male spouses who had the same income and credit score.[5]

These cases are neither fringe nor extreme scenarios. Algorithmic systems are too often built and sustained by historic data and models that reproduce stereotypes and false assumptions about gender, race, sexual orientation, ability, class, geography, and other socio-cultural and demographic factors. **The bottom line is that without dedicated effort, the use of algorithmic technologies perpetuates and amplifies societal inequalities and harmful stereotypes.**

---

[1] The "softwarisation" of bias means that existing inequalities end up coded in and perpetuated in obscure and IP-protected machines, see page 10 for further explanation.

[2] Melissa Heikkila, Dutch scandal serves as a warning for Europe over risks of using algorithms, Politico, 29 March 2022, available at: https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/ (last accessed: 30 August 2022)

[3] See Tweede Kamer der Staten-Generaal, Parlementaire ondervraging kinderopvangtoeslag (2020) available at: https://zoek.officielebekendmakingen.nl/kst-35510-1.pdf.

[4] Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, 11 October 2018, available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (last accessed: 25 July 2022).

[5] Alisha Haridasani Gupta, "Are Algorithms Sexist?" *The New York Times* (15 November 2019) available at: https://www.nytimes.com/2019/11/15/us/apple-card-goldman-sachs.html (last accessed: 25 July 2022).

Awareness of the risks of algorithmic discrimination has crystallised around discussions on 'bias', which has now become a prominent public issue. A 2022 survey showed that over 36% of companies "experience[e] challenges or direct business impact due to an occurrence of AI bias in their algorithms, such as […] [l]ost revenue, [l]ost customers, [l]ost employees, [i]ncurred legal fees due to a lawsuit or legal action [and] [d]amaged brand reputation/media backlash".[6] Legislators and regulators around the world are also grappling with these risks and with the pitfalls of existing legislation to address them. Questionnaires answered by representatives of the state parties to the European Convention on Human Rights (ECHR) for the purpose of the present Study shows broad awareness of the legal issues related to algorithmic bias.[7] In almost all State Parties, policy or legislative initiatives are either ongoing or public consultations are taking place for this purpose.

The Council of Europe and in particular its Gender Equality Commission (GEC) and the Steering Committee on Anti-Discrimination, Diversity and Inclusion (CDADI) have also undertaken work in this area. The Ad Hoc Committee on Artificial Intelligence (CAHAI) was mandated in 2019-2021 to consult with stakeholders and to examine the feasibility and potential elements of a legal framework for the development, design and application of artificial intelligence, based on the Council of Europe standards on human rights, democracy and the rule of law. It published a "Feasibility Study on legal framework on AI design, development and application based on CoE standards" in 2020 as well as "Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law". Following these developments, a new Committee on Artificial Intelligence (CAI) has been set up in 2022 and mandated to draft a Framework Convention "on the development, design, and application of artificial intelligence systems based on the Council of Europe's standards on human rights, democracy and the rule of law, and conducive to innovation".[8] A regulatory instrument enacted by the Council of Europe has the valuable potential to foster a **human-rights-based approach** to the use of AI and algorithmic technologies in and beyond the international community of State Parties to the ECHR.

The aim of this study is threefold. First, it explains how bias in AI and algorithmic technologies arises and may lead to discrimination. It highlights how bias is not just related to data but to the wider human and social underpinnings of these technological artefacts. Second, the Study reviews how policy makers, legislators and companies are dealing with the discriminatory risks of algorithmic technologies and assesses which existing legal instruments could be used for this purpose in the future. It also identifies the shortcomings of existing legal tools and proposes regulatory adaptations to promote equality and prevent discrimination from arising in the development and deployment of algorithmic systems. Third, the Study explores the socio-political conditions necessary for algorithmic technologies to be used to promote equality. It sets out possibilities to leverage these technologies for equality through the legal routes of positive action and positive obligations. Finally, the Study recommends several avenues for ensuring that the use of algorithmic technologies does not automate existing

[6] See DataRobot, "DataRobot's State of AI Bias Report Reveals 81% of Technology Leaders Want Government Regulation of AI Bias" (2022), available at: https://www.datarobot.com/newsroom/press/datarobots-state-of-ai-bias-report-reveals-81-of-technology-leaders-want-government-regulation-of-ai-bias/.

[7] See section II and Annex.

[8] See Terms of reference of the Committee on Artificial Intelligence CM(2021)131 available at: https://rm.coe.int/cai-terms-of-reference/1680a7b90b.

inequalities but contributes to a better and more equitable society. All in all, this study aims to support the work of a future Expert Committee under the GEC and CDADI to draft a possible specific sectoral legal instrument on the impact of artificial intelligence systems on equality, including gender equality, and non-discrimination in 2024 and 2025.

In terms of scope, the study focuses mostly on Europe and charts the opportunities and problems that the deployment of algorithmic technologies in society poses in relation to equality and discrimination. It explores the responses that have been given and are being discussed in several countries that are members of the Council of Europe or have observer status. The Study builds on Borgesius' study on "Discrimination, Artificial Intelligence and Algorithmic Decision-Making" commissioned by the Council of Europe in 2018 as well as on the fast-developing interdisciplinary body of research on algorithmic discrimination and AI bias.[9] The Study addresses issues of algorithmic discrimination across all grounds protected under Article 14 of the European Convention on Human Rights (ECHR) but with a particular focus on the three groups of protected grounds that are gender and sex, gender identity and sex characteristics; and race, ethnic and national origin, colour, citizenship, religion, language. The study reviews the harmful consequences of AI bias in a wide range of public and private sectors, but with emphasis on employment and education.

Section 1

Unpacking 'machine bias': How can algorithmic technologies lead to discrimination?

A note on terminology: For the sake of clarity, the term "**user**" of algorithms refers to companies, public bodies or any other stakeholders who deploys an algorithm to support or automate a decision-making process. By contrast, "**end users**" are those subjected to algorithmic or algorithmically supported decisions, for instance customers, job candidates, tax payers, etc. "**Providers**" of algorithmic and AI systems are those who design and commercialize such systems without implementing them in real-life conditions. Sometimes, when algorithmic or AI systems are developed in-house, the provider and the user are the same entity.

1) What is AI?

For the purpose of this analysis, we use the **broad definition of AI** put forward by the ad hoc committee on artificial intelligence (CAHAI) of the Council of Europe, which describes AI "as a 'blanket term' for various computer applications based on different techniques, which exhibit capabilities commonly and currently associated with human intelligence".[10] The CAHAI acknowledges that "[t]hese techniques can consist of formal models (or symbolic systems) as well as data-driven models (learning-based systems) typically relying on statistical approaches, including for instance supervised learning, unsupervised learning and reinforcement learning" and that "AI systems act in the physical or digital dimension by recording their environment through data acquisition, analysing certain structured or

[9] See Frederik Borgesius, Discrimination, Artificial Intelligence and Algorithmic Decision-Making (2018) Council of Europe available at: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73.

[10] Ad hoc committee on artificial intelligence, Feasibility Study CAHAI(2020)23 (Council of Europe, 2020), [8].

unstructured data, reasoning on the knowledge or processing information derived from the data, and on that basis decide on the best course of action to reach a certain goal".[11] A further aspect of the definition is that "[these systems] can be designed to **adapt their behaviour over time based on new data** and enhance their performance towards a certain goal".[12]

The background to this broad definition of AI is that, **to date, there is no single definition of AI accepted by the scientific community**. For example, the proposed EU AI regulations define AI as "software that is developed with one or more […given…] techniques and approaches and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with".[13]

According to the EU definition, the techniques and approaches leading to software being identified as an AI system include:

- "Machine learning (including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning);
- Logic- and knowledge-based approaches (including knowledge representation, inductive (logic) programming, knowledge bases, inference/deductive engines, (symbolic) reasoning and expert systems);
- Statistical approaches, Bayesian estimation, search and optimization methods".[14]

This variety of techniques falling under the definition of AI include software powering, for example, search engines, image and speech recognition systems, machine translation websites, virtual assistants, spam filters, programmes supporting medical diagnosis, as well as machines such as self-driving cars, robots, and a myriad of objects falling under the vast category of the Internet of Things.[15] In this Study, we find it important to underline that **the regulatory subject is not AI taken in isolation but rather the broader socio-technical apparatus** constituted by the interaction of social elements with algorithmic technologies.

2) What is algorithmic bias?

Algorithms are able to process a far greater range of inputs and variables to make decisions, and can do so with speed and, arguably, reliability that far exceed human capabilities. From the ads we are served, to the products we are offered, and to the results we are presented with after searching online, algorithms play an ever-greater part in making these decisions.

However, because algorithms simply present the results of calculations **defined by humans** using data that may be provided by humans, machines, or a combination of the two (at some point during the process), they reflect and process the human biases that are incorporated

[11] Ibid.
[12] Ibid.
[13] EU AI Act, Art. 3(1).
[14] See Annex 1 of the EU AI Act: "Artificial intelligence techniques and approaches referred to in Article 3, point 1".
[15] European Parliament, "What is artificial intelligence and how is it used?" (2021) available at: https://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used.

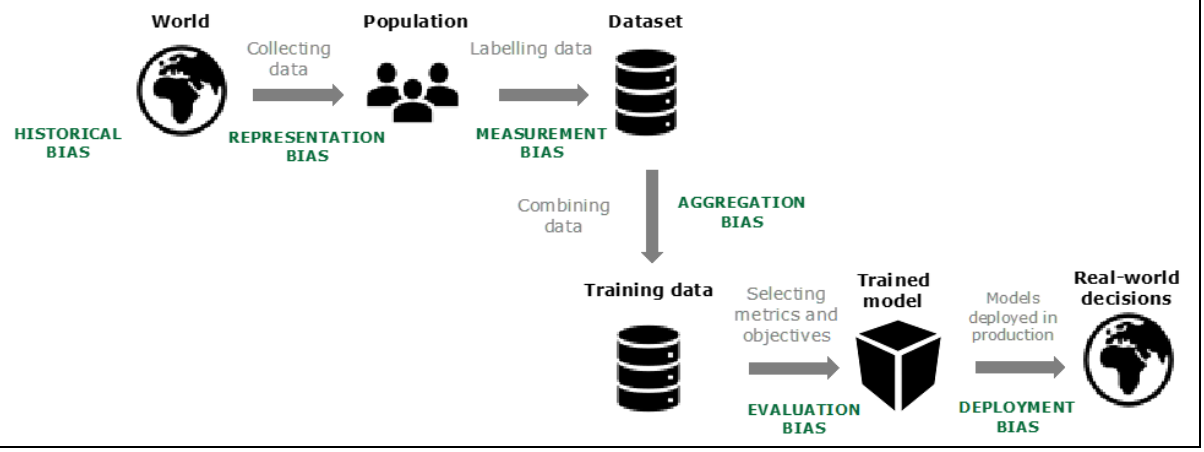
when the algorithm is programmed, when it processes data and when humans interact with it.

In a nutshell, "[algorithmic] ***[b]ias happens when seemingly innocuous programming takes on the prejudices either of its creators or the data it is fed***."[16] As a consequence, women (for example) may be denied loans and credit, and speech recognition programs may misidentify words spoken by black people at much greater rates than for white people.[17]

As Sofiya Noble's concept of "algorithmic oppression" clarifies, bias is not a "glitch" in otherwise unbiased systems but is instead **systemic and inherent in the functioning of information systems** powering search engines and other web applications.[18]

Contrary to a widespread narrative, datasets are not the only relays of bias in learning algorithms. Bias has different sources throughout the lifecycle of algorithmic applications, from their inception to their deployment and use. **The complexity of bias emergence and impact is the reason why close attention must be paid to the entire lifecycle of AI and algorithmic systems.**[19] Several taxonomies listing the sources of bias and its channelling into AI systems and outputs have been developed by researchers. For example, the diagram below by Suresh and Guttag shows the different entry points for bias, and what they entail.

Table and definitions below from: A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle[20]



The researchers distinguish five sources and types of bias in AI systems. First, what they call "**historical bias**" describes how social hierarchies and institutionalised disadvantage shape

[16] Garcia, Megan. "Racist in the Machine: The Disturbing Implications of Algorithmic Bias." *World Policy Journal* 33 (2016): 111 - 117.
[17] Allison Koenecke, et al., PNAS, March 23, 2020
[18] See Safiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: New York University Press, 2018) and Vanessa Ceia, Benji Nothwehr, and Liz Wagner, Gender and Technology: A rights-based and intersectional analysis of key trends (Oxfam Research Backgrounder, 2021), 40.
[19] Ivana Bartoletti, The Complex Issue of Algorithmic Fairness, The Yuan, September 2021, available at: https://www.the-yuan.com/129/The-Complex-Issue-of-Fairness-in-AI-Part-I.html (last accessed: 28 July 2022)
[20] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Proceedings of EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). ACM, New York, NY, USA, 9 pages. https: //doi.org/10.1145/3465416.3483305.

social data.[21] Data is therefore not neutral because it is a capture of the unequal society we live in. For example, as women have traditionally earned less than men, they may be given less credit[22] or, in the context of advertising, be served ads with lower paid job posts.[23]

In turn, "**representation bias**" arises in data collection.[24] For example, if an organisation's marketing team advertises in predominantly white neighbourhoods, the resulting customer base would not be representative of the wider population. That dataset would generate bias if used for example to train an algorithm later used to cater to broader population groups.

The researchers also shed light on "**measurement bias**", which "occurs when choosing, collecting, or computing features and labels to use in a prediction problem".[25] Many features and labels are non-problematic, such as the labelling of an image as a cat or a dog, but problems may emerge when some factors are used as a proxy. For example, postcode could be a proxy for race or sexual orientation, and occupation could be a proxy for gender. Alternatively, if proxies overly simplify the feature to be measured or the proxy reflects variations in the quality of measurements across groups, measurement bias could arise.[26]

"**Aggregation bias**" relates to how data is combined. It occurs when data groups are inappropriately combined, resulting in a model that does not perform well for any group or only performs well for the majority group.[27] The researchers mention the example of local meanings ascribed by specific communities to emoji, hashtags and sentences on social media, which differ from the meanings in the broader social media user population.[28] This could lead for instance to content moderation applying inadequate semantic filters modelled on majority groups to minority groups, with silencing effects that could unfairly restrict minority groups ability to communicate via social media.

The researchers also identify "**evaluation bias**", which occurs when evaluating a model, if the benchmark data (used to compare the model to other models that perform similar tasks) does not represent the population that the model will serve.[29] For example, the Gender Shades paper discovered that two widely used facial analysis benchmark datasets (IJB-A and

---

[21] See ibid.

[22] Apple's 'sexist' credit card investigated by US regulator, BBC, 11 November 2019, available at: https://www.bbc.com/news/business-50365609 (last accessed: 15 June 2022).

[23] Samuel Gibbs, Women less likely to be shown ads for high-paid jobs on Google, study shows, The Guardian, 8 July 2015, available at: https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study (last accessed: 15 June 2022).

[24] See Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Proceedings of EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). ACM, New York, NY, USA, 9 pages. https: //doi.org/10.1145/3465416.3483305.

[25] Ibid.

[26] Ibid.

[27] Ibid.

[28] Ibid, citing a study by Desmond U. Patton, William R. Frey, Kyle A. McGregor, Fei-Tzin Lee, Kathleen McKeown, and Emanuel Moss. 2020. Contextual Analysis of Social Media: The Promise and Challenge of Eliciting Context in Social Media Posts with Natural Language Processing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (New York, NY, USA) (AIES '20). Association for Computing Machinery, New York, NY, USA, 337–342. https://doi.org/10.1145/3375627.3375841.

[29] See Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Proceedings of EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). ACM, New York, NY, USA, 9 pages. https: //doi.org/10.1145/3465416.3483305.

Adience) were primarily composed of lighter-skinned subjects (79.6% and 86.2%, respectively).[30]

Finally, "**deployment bias**" relates to the real-world use of models, in particular if a model developed to solve a problem is used for another task.[31] This could happen for example due to a change in marketing strategy. In addition, a model is often a part of a complex socio-technical system where human and machines interact. In a 'live' environment, additional biases may therefore be introduced when humans interpret algorithmic outputs to be used as inputs further down the algorithmically supported decision-making line.[32]

So-called **automation and confirmation biases** can also strengthen these biases. Automation bias takes place when humans place greater trust in machines and technological artefacts than in their own or other humans' potentially contradictory judgment, and therefore tend to validate algorithmic outputs without questioning them. In the context of predictive machines for example, such bias can lead to biased risk assessments not being challenged by so-called humans-in-the-loop and therefore to rubberstamping behaviours. Confirmation bias happens when pre-existing beliefs influence the processing of new information, leading in particular to new information being better retained when consistent with such beliefs or being interpreted in consistency with such beliefs. In the AI context, this could lead to gender stereotypes acting as a reinforcing prism by human decision-makers when interpreting biased algorithmic outputs. In an experiment, Green and Chen also shows that human interpreters of automated risk assessments provided by an algorithm yield "**disparate interactions**", that is interpretations of similar algorithmic risk assessments are more lenient towards white than black defendants.[33]

Other taxonomies of bias have been proposed. For example, Barocas and Selbst identify key moments and situations where bias is channelled into AI systems: the **definition of "target variables"** (the feature to be measured or predicted by a model, e.g. work performance) and **"class labels"** (the possible variations in the occurrence of the target variable, for example stellar, very good, good, unsatisfactory); the use of "**training data**" (with bias occurring during labelling and data collection); "**feature selection**" (the attributes that are to be considered relevant by a model, for instance yearly income); and the use of "**proxies**" (when relevant attributes correspond to protected groups, for example yearly income and gender due to the gender pay gap).[34]

These taxonomies help debunk the myth that bias emerges from data only, and show the complex role of socio-technical interactions in the (re)production of discriminatory bias.

---

[30] Buolamwini J and Gebru T, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* (Proceedings of Machine Learning Research 2018).

[31] See ibid.

[32] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Proceedings of EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). ACM, New York, NY, USA, 9 pages. https: //doi.org/10.1145/3465416.3483305.

[33] See Green B and Chen Y, 'Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments' (2019) Proceedings of the Conference on Fairness, Accountability, and Transparency 90.

[34] Barocas S and Selbst AD, 'Big Data's Disparate Impact' (2016) 104 California law review , 677-693.

3) The discriminatory impact of AI: some concrete examples

This section illustrates how bias can give rise to discrimination across different sectors.

Recruitment: Reuters journalist Jeffrey Dastin reported in 2018 that Amazon developed a program relying on machine-learning to identify top candidates in pools of CVs. The program systematically disadvantaged women's CV because it reflected the gender gap in the workforce recruited over the past ten years. Neutralising words like "women" did not redress the discriminatory outcome as the system was able to infer gender identity from other data.[35]

Researchers based at Utrecht University partnered with a job matching platform to research how the use of gendered language in the search bar yields different results, with discriminatory allocations of information about job opportunities.[36] This not only results in strengthening stereotypes about male and female typical occupations but also results in allocative and distributive harms.

The online targeted distribution of job ads powered by optimisation services offered by social media platforms such as Facebook also proves to reinforce gender stereotypes as well as gender segregation within the workplace.[37] An experiment conducted by AlgorithmWatch in 2020 showed that when asking Facebook to distribute ads "neutrally" (without targeting a specific audience), an ad for a truck driver position was shown to a public composed of 93% men and 7% women.[38] Conversely, an ad for a position as educator was distributed to an audience composed of 96% women and 4% men.[39]

AI-powered face recognition and emotions analysis systems can also yield racial discrimination or disadvantage job candidates with disabilities.[40] This is because of lower performance rates of such devices on darker skin tones, especially for women.[41] In addition, emotions analysis software trained on neurotypical subjects might not be able to perform correctly on neurodiverse subjects. As AI-powered emotions analysis is increasingly used in

[35] See Dastin J, 'Amazon scraps secret AI recruiting tool that showed bias against women' *Reuters* (2018) available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-airecruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (last accessed 22 July 2022).

[36] See van Es K, Everts D and Muis I, 'Gendered language and employment Web sites: How search algorithms can cause allocative harm' (2021) 26 First Monday available at: https://journals.uic.edu/ojs/index.php/fm/article/view/11717/10200.

[37] See Ali M and others, 'Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes' (2019) 3 Proceedings of the ACM on Human-Computer Interaction 1.

[38] 4,864 men, but only 386 women. See Wulf J, Automated Decision-Making Systems and Discrimination: Understanding causes, recognizing cases, supporting those affected (AlgorithmWatch 2022), 7 available at: https://algorithmwatch.org/en/wp-content/uploads/2022/07/AutoCheck-Guidebook_ADM_Discrimination_EN-AlgorithmWatch_June_2022_b.pdf and Kayser-Bril N, 'Automated Discrimination: Facebook uses gross stereotypes to optimize ad delivery' AlgorithmWatch available at: https://algorithmwatch.org/en/automated-discrimination-facebook-google/ (last accessed 22 July 2022).

[39] Ibid. 6,456 women, but only 258 men.

[40] See Buolamwini J and Gebru T, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (Proceedings of Machine Learning Research 2018); Hannah Devlin, "AI systems claiming to 'read' emotions pose discrimination risks" (16 February 2020) *The Guardian* available at: https://www.theguardian.com/technology/2020/feb/16/ai-systems-claiming-to-read-emotions-pose-discrimination-risks (last accessed 22 July 2022).

[41] See Buolamwini J and Gebru T, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (Proceedings of Machine Learning Research 2018).

the recruitment sector, for instance to analyse video recordings of job candidates' presentations, this could pose accessibility and inclusion issues.

Access to goods and services, banking and insurance: In Finland the National Non-Discrimination and Equality Tribunal found direct multiple discrimination in a case where the applicant was denied a loan online. After investigating the case, the Equality Body (the Non-Discrimination Ombudsman) had found that the company used statistical models to assess credit worthiness that relied on an applicant's age, gender, language and place of residence while not taking into account an applicant's actual credit history. In that case, the applicant being male, Finnish speaker and from a rural area were treated as factors of disadvantage in the assessment performed by the financial institution.[42]

A similar story was reported in Germany, where a female customer was refused a credit while purchasing goods online. When investigating the reasons for the rejection with the credit institution, the customer learned that a combination of her age and gender seemed to have motivated the automated rejection, based on harmful intersectional stereotypes that women around 40 are often divorced and have therefore less economic power.[43]

In the insurance sector, a study conducted by the Universities of Padua, Udine, and Carnegie Mellon showed that factors such as birthplace and citizenship influence the price of car insurance policies paid by customers.[44] In a case study, they showed that indicating Ghana as an applicant's birthplace could lead to a price increase of 1000 EUR compared to an applicant indicating Italy as their birthplace.

Another study by AlgorithmWatch showed that digital discrimination extends far beyond AI.[45] Simple online forms can cause discrimination on grounds of race, ethnic origin or nationality, for example if they only allow registering patronyms containing three or more letters. Applicants with shorter names will be denied registration or unable to open an account, which is often a precondition for purchasing goods and services online.

Risk assessment in the area of security, crime prevention, policing and the justice system: In Spain the VioGén software has been used to assess risks of gender-based violence and femicide by intimate partners. Despite an overall favourable assessment, criticisms point to several cases of false negatives where low risk scores led to insufficient prevention means being deployed, with tragic consequences.[46]

---

[42] See Lorenz Matzat and Minna Ruckenstein, "Finnish Credit Score Ruling raises Questions about Discrimination and how to avoid it" (21 November 2018) *AlgorithmWatch* available at: https://algorithmwatch.org/en/finnish-credit-score-ruling-raises-questions-about-discrimination-and-how-to-avoid-it/ (last accessed 22 July 2022); Rainer Hiltunen, "Multiple discrimination in assessing creditworthiness" (1 August 2018) European network of legal experts in gender equality and non-discrimination available at: https://www.equalitylaw.eu/downloads/4658-finland-multiple-discrimination-in-assessing-creditworthiness-pdf-120-kb (last accessed 22 July 2022).

[43] See Wulf J, Automated Decision-Making Systems and Discrimination: Understanding causes, recognizing cases, supporting those affected (AlgorithmWatch 2022), 6-7

[44] The study was reported by AlgorithmWatch, see ibid.

[45] Lulamae, Josephine, "Fixing Online Forms Shouldn't Wait Until Retirement", AlgorithmWatch (13 January 2022) available at: https://algorithmwatch.org/en/unding-online-forms/ (last accessed 22 July 2022).

[46] Michele Catanzaro, "In Spain, the VioGén algorithm attempts to forecast gender violence", AlgorithmWatch (27 April 2020) available at: https://algorithmwatch.org/en/viogen-algorithm-gender-violence/ (last accessed 22 July 2022).

The Netherlands have deployed several predictive systems for crime prevention purposes, which have been harshly criticised for creating discrimination based on race, ethnicity and nationality. For instance, a 2020 investigation by Amnesty International revealed that the "Sensing Project", that aimed to prevent shoplifting and pickpocketing locally, resulted in discriminatory ethnic profiling of individuals of Eastern European origin, and in particular members of the Roma community.[47] When watching car traffic in and around the area of deployment, the system used the Eastern European origin of passengers as a predictive risk factor for crime. Other crime anticipation systems, for instance in Amsterdam, have been reported to use factors such as "number of one parent households", "number of social benefits recipients" and "number of non-Western immigrants" to identify crime "hot spots" throughout the country.[48]

At airports, security screening and border control technologies using automated gender recognition systems have been shown to discriminate against transgender, intersex, non-binary and non-conforming persons because it relies on a binary gender classification system that does not capture the real complexity of gender identity.[49]

Facial recognition is increasingly deployed for crime detection and prevention. For example, law enforcement agencies may use face recognition to compare suspects' photos to mugshots and driver's license images. While "[f]ace recognition algorithms boast high classification accuracy (over 90%)", these outcomes are not universal.[50] In 2018, the Gender Shades project revealed discrepancies in the classification accuracy of face recognition technologies for different skin tones and sexes. These algorithms consistently demonstrated the poorest accuracy for darker-skinned females and the highest for lighter-skinned males.[51] In a criminal justice setting, face recognition technologies that are inherently biased in their accuracy can potentially misidentify suspects and even lead to the incarceration of innocent people of colour as has happened in the US.[52] It is therefore concerning that, even if accurate, "face recognition empowers […] law enforcement system[s] with a long history of racist and anti-activist surveillance and can widen pre-existing inequalities".[53]

Access to public and administrative services: the use of face recognition technologies within or in association with public services can lead to excluding or denying end users public services. For instance, a photo booth at the State Office of Transportation in Hamburg,

[47] Amnesty International "We Sense Trouble: Automated Discrimination and Mass Surveillance in Predictive Policing in the Netherlands" (2020), 5 available at: https://www.amnesty.nl/content/uploads/2020/09/Report-Predictive-Policing-RM-7.0-FINAL-TEXT_CK-2.pdf (last accessed 22 July 2022).
[48] https://www.vice.com/en/article/5dpmdd/the-netherlands-is-becoming-a-predictive-policing-hot-spot
[49] See JD Shadel, "#TravelingWhileTrans: The trauma of returning to 'normal'" (The Washington Post, 2021) available at: https://www.washingtonpost.com/travel/2021/06/16/trans-travel-tsa-lgbtq/ and Quinan, C. L., and Mina Hunt. "Biometric Bordering and Automatic Gender Recognition: Challenging Binary Gender Norms in Everyday Biometric Technologies." Communication, Culture and Critique 15.2 (2022): 211-226.
[50] Alex Najibi, **Racial Discrimination in Face Recognition Technology, Harvard University, October 2020, available at: https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/#:~:text=Face%20recognition%20algorithms%20boast%20high,and%2018%2D30%20years%20old.**
[51] Gender Shades Project, available at http://gendershades.org/overview.html (last accessed: 31 August 2022)
[52] RACE AND WRONGFUL CONVICTIONS IN THE UNITED STATES, available at: https://www.law.umich.edu/special/exoneration/Documents/Race_and_Wrongful_Convictions.pdf (last accessed: 31 August 2022).
[53] Alex Najibi, Racial Discrimination in Face Recognition Technology, Harvard University, October 2020.

Germany, failed to recognise an applicant's face for the purpose of taking a biometric picture, which was needed for her administrative application. Even though the public office denied that the failure stemmed from the facial recognition software used, a local employee indicated that failures often take place in relation to applicants' skin colour.[54]

In the Netherlands, the deployment of the SyRi system (System Risk Indication), used to detect social welfare fraud, was shown to cause discrimination on grounds of income and ethnic origin before being put to halt by a court decision in 2020.[55] In 2021, a welfare scandal forced the Dutch government to resign after more than 20.000 parents were flagged by an AI system as fraudsters in relation to childcare allowance and subjected to investigation by the Dutch tax authorities.[56] The AI system treated double nationality as a high risk factor and this resulted in a disproportionate number of investigations and court proceedings being launched against families with an immigration background, whose child care benefits were suspended and some of which were requested to reimburse the benefits perceived.[57] The case also shows how the lack of accountability and transparency around the use of these systems can lead to depriving the subjects of AI decision-making from an explanation or the opportunity to appeal against the decisions.

Education: Facial recognition software have been known to be biased and lead to intersectional discrimination on grounds of race and gender.[58] When used in proctoring software in educational settings, that can negatively affect the conditions in which racialised students take exams and even their ability to do so. For example, proctoring software used by several universities in the Netherlands had trouble recognising dark-skinned students.[59] After the University did not take her complaint seriously, a student supported by the Racism and Technology Centre submitted a formal complaint to the Institute of Human Rights, the national non-discrimination authority in the country.[60] Proctoring software can also impact

---

[54] See Wulf J, Automated Decision-Making Systems and Discrimination: Understanding causes, recognizing cases, supporting those affected (AlgorithmWatch 2022), p8. This hypothesis is corroborated by studies pointing at intersectional discrimination on grounds of gender and skin colour in facial recognition software, e.g., Buolamwini J and Gebru T, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (Proceedings of Machine Learning Research 2018).

[55] Koen Vervloesem, "How Dutch activists got an invasive fraud detection algorithm banned", AlgorithmWatch (6 April 2020) available at: https://algorithmwatch.org/en/syri-netherlands-algorithm/ (last accessed 22 July 2022).

[56] Nadia Benaissa, "Het systeem doet precies wat het wordt opgedragen" (29 January 2021) *Bits of Freedom* available at: https://www.bitsoffreedom.nl/2021/01/29/het-systeem-doet-precies-wat-het-wordt-opgedragen/.

[57] Jon Henley, "Dutch government faces collapse over child benefits scandal " (14 January 2021) The Guardian available at: https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal and Björn ten Seldam & Alex Brenninkmeijer, "The Dutch benefits scandal: a cautionary tale for algorithmic enforcement" (30 April 2021) EU Law Enforcement available at: https://eulawenforcement.com/?p=7941.

[58] Buolamwini J and Gebru T, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (Proceedings of Machine Learning Research 2018).

[59] Racism and Technology Centre, "Student stapt naar College voor de Rechten van de Mens vanwege gebruik racistische software door de VU" (15 July 2022) available at: https://racismandtechnology.center/2022/07/15/student-stapt-naar-college-voor-de-rechten-van-de-mens-vanwege-gebruik-racistische-software-door-de-vu/#more-1691 (last accessed 28 July 2022).

[60] Fleur Damen, "De antispieksoftware herkende haar niet als mens omdat ze zwart is maar bij de vu vond ze geen gehoor" *De Volkskrant* (15 July 2022) available at: https://www.volkskrant.nl/nieuws-achtergrond/de-antispieksoftware-herkende-haar-niet-als-mens-omdat-ze-zwart-is-maar-bij-de-vu-vond-ze-geen-gehoor~b6810279/ (last accessed 27 July 2022).

students with disabilities negatively, for instance by generating anxiety, not allowing a carer or not letting the students take breaks away from the computer.[61] For low-income families who share rooms due to a lack of space at home, the use of a proctoring software can create disadvantage by signalling "aberrant behaviour" if family members are identified passing behind the screen.[62]

Healthcare: Criado Perez has exposed how healthcare research and industry rely on male models to assess the risks and efficacy of drugs, thus yielding less and lower quality health data for women and gender diverse persons. Such gender data gap in the healthcare sector, leads to less reliable predictive systems when it comes to diagnosing female and gender diverse patients.[63] Research shows that the data gap in health also affects other minority groups.[64]

A US study by Obermeyer at al. shows how a system used to predict health-related risks in order to allocate resources systematically disadvantaged patients with ethnic minority backgrounds. This is because the system used data about groups' previous access to healthcare, which embedded existing structural discrimination.[65]

Media and search engines: Research shows that representations of women in images returned by search engines online are biased and reflect sexist, racist and intersectionally discriminatory stereotypes. For instance, Noble shows in an experiment with the Google search engine how images of black girls and black women are sexualised.[66] Even though search engines have tried to correct these biases, a recent study surveying major search engines shows "representation bias" as well as "face-ism bias" in the way women are represented, meaning that "[w]omen are less likely to be represented in gender-neutral media content [...] and their face-to-body ratio in images is often lower" than for men".[67] Technical debiasing solutions might treat some of the symptoms of the problem, for instance re-balancing the amount of female pictures in an image search for "CEOs", but not its roots, in this case harmful stereotyping, representational and allocative harms as well as structural inequality that are deeply entrenched in our cultural and material reality. For instance, recent tests seem to show that the AI-powered art tool DALLE2, which is currently being tested, adds 'diversity prompts' to unspecific queries, for example adding the labels "black" or "female" to a prompt asking the software to generate an image of 'a CEO'.[68] This approach is

---

See the complaint at: https://racismandtechnology.center/2022/07/15/student-stapt-naar-college-voor-de-rechten-van-de-mens-vanwege-gebruik-racistische-software-door-de-vu/#more-1691

[61] Lydia X. Z. Brown, "How Automated Test Proctoring Software Discriminates Against Disabled Students" (16 November 2020) Centre for Democracy and Technology available at: https://cdt.org/insights/how-automated-test-proctoring-software-discriminates-against-disabled-students/ (last accessed 28 July 2022).

[62] Ibid.

[63] See Criado Perez C, Invisible women: Exposing data bias in a world designed for men (Random House 2019).

[64] Ibid.

[65] See Obermeyer Z and others, 'Dissecting racial bias in an algorithm used to manage the health of populations' (2019) 366 Science 447.

[66] See e.g. Safiya Noble, *Algorithms of oppression : how search engines reinforce racism* (New York University Press 2018).

[67] Ulloa R and others, 'Representativeness and face-ism: Gender bias in image search' (2022) New Media & Society.

[68] Matthew Sparkes, "AI art tool DALL-E 2 adds 'black' or 'female' to some image prompts" (22 July 2022) *New Scientist* available at: https://www.newscientist.com/article/2329690-ai-art-tool-dall-e-2-adds-black-or-female-to-some-image-prompts/ (last accessed 28 July 2022); see also OpenAI, "Reducing Bias and Improving Safety in

analogous to a form of positive action like quotas. It can be criticised for not addressing the lack of diversity in training sets, but if used at a large scale, such fixes have the merit to disseminate more diverse representations that, on the long run, can contribute to mitigating harmful stereotypes.

Online gender-based violence, hate speech, harassment: Digital discrimination also takes the form of gender-based violence, for instance when deepfake videos are used to harass women in the context of so-called "revenge porn" cases. Unconsented dissemination of sexual content, often in the form of images, has also been recognised as a form of gender-based violence that especially affect women and girls who are young or public figures such as journalists, human rights defenders, or politicians.[69] In addition, sexist and other forms of online hate speech have been highlighted as contingent on the rising use of social media platforms.[70] At the same time, content moderation particularly affects minority groups, who are at risk of being silenced[71] while at the same time subjected to hate campaigns.

Gender stereotyping across the board: A recent UN report, "I'd blush if I could: closing gender divides in digital skills through education" found that AI digital assistants with female voices can reinforce existing gender biases. This trend toward female voiced virtual assistants "seems to have less to do with sound, tone, syntax, and cadence, than an association with assistance".[72] Perhaps a female voice is chosen to seduce a user into thinking that AI is pliable and benign. But the ultimate effect is the "normalisation of this new digital servitude in our homes and daily lives through Alexa, Siri and Cortana".[73]

4) What makes algorithmic discrimination different?

Discrimination powered by algorithmic technologies presents a set of **distinct challenges** compared to human discrimination.

First, the higher performance of machines entails a **much larger effect on society**. For example, while a bank employee might unconsciously assign a higher mortgage rate to an applicant from a minority group, a software processing thousands of files per day might generalise this bias to any applicant with an African sounding name.

Secondly, human conduct is controlled by social and legal mechanisms that, although far from perfect, are meant to correct misbehaviours in the short and long term. By contrast, **the deployment of algorithmic technologies often jeopardizes accountability for, transparency in and scrutiny of decision-making processes**. For example, "[w]rong human decisions can

DALL·E 2" (18 July 2022) available at: https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/ (last accessed 28 July 2022).

[69] See Sara De Vido and Lorena Sosa, Criminalisation of gender-based violence against women in European States, including ICT-facilitated violence (European Network of Legal Experts in gender equality and non-discrimination 2021) available at: https://www.equalitylaw.eu/downloads/5535-criminalisation-of-gender-based-violence-against-women-in-european-states-including-ict-facilitated-violence-1-97-mb (last accessed 23 July 2022).

[70] See Bartoletti, Ivana. Chapter 3: Algorithms and the Rise of Populism in *An artificial revolution: On power, politics and AI*. Black Spot Books, 2020.

[71] See Rachel Griffin, 'The Sanitised Platform' (2022) 13 J Intell Prop Info Tech & Elec Com L 36.

[72] UNESCO, I'd blush if I could: closing gender divides in digital skills through education, 100 available at: https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

[73] Ivana Bartoletti, An Artificial Revolution: on Power, Politics and AI (Indigo Press).

be appealed, while both the opacity of AI technology and the unwillingness of providers to open up to public scrutiny make this very difficult to achieve in AI systems".[74]

Third, **the sources of algorithmic discrimination are difficult to identify**. Due to the complexity of these socio-technical systems, bias can affect any stage of the algorithmic pipeline. In addition, **algorithms might be proprietary, complex and difficult to understand**. Sometimes, they are effectively a sealed box, containing proceedings that may be unexplainable to a human researcher. This "softwarisation" of bias means that existing inequalities end up coded in and perpetuated in obscure and IP-protected machines. This is extremely problematic as bias becomes more difficult to identify and harder to challenge.

To sum up, at least **six challenges** arise with algorithmic and data-driven discrimination.[75] Machine-supported decisions are made at a much **greater scale** but the interaction between humans and machines make the **sources of discrimination difficult to identify and address**. The **'cleaning' of biased data is a technical challenge** and a **context-dependent** exercise, and the existence of proxies for and correlations with protected groups further complicates the task. At the same time, AI and algorithmic systems are often **non-transparent**, might not be explainable, and the **attribution of responsibility for discrimination is unclear**.

Because **the source of these biases is not ultimately technological, they cannot be resolved using technology alone.** Instead, addressing algorithmic discrimination and data-driven disadvantage requires a much greater degree of scrutiny and **a positive political decision to actively prevent the reinforcing of structural inequalities engrained in social data**. For example, to avoid "automating" gender stereotypes and the gender pay gap – the fact that women have traditionally earned less than men – employers need to make a conscious decision to target women when advertising higher paying, typically "masculine" or management jobs online. Simply entrusting their distribution to optimization algorithms instead is likely to reproduce gender stereotypes and pay inequality.[76] Understanding algorithmic bias therefore starts with recognising how algorithmic technologies escalate, entrench and perpetuate existing inequalities where no safeguards are put in place. For these reasons, **addressing algorithmic discrimination requires a multifaceted approach encompassing various disciplines** such as social science, ethics and law, and **regulatory fields** including legislation on non-discrimination, consumer protection, data protection, trade, etc.

5) Addressing algorithmic discrimination: best practices and their limits

To address the discriminatory risks of algorithmic technologies, the industry has taken initiatives ranging from **technical solutions to 'debias' and 'audit' algorithmic systems** to **voluntary codes of conduct**, **instruments for ethical AI** and other forms of **self-regulation**.

---

[74] 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, available at: https://www.ejlt.org/index.php/ejlt/article/download/754/1044/3716 (last accessed: 16 September 2022)

[75] See Gerards J and Xenidis R, *Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law* (European network of legal experts in gender equality and non-discrimination / European Commission, 2021).

[76] See Ali M and others, 'Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes' (2019) 3 Proceedings of the ACM on Human-Computer Interaction 1 and Imana B, Korolova A and Heidemann J, *Auditing for discrimination in algorithms delivering job ads* (2021).

This section exhibits some **examples of the good governance practices** adopted and assesses their **limits**.

Companies have been ramping up governance milestones in anticipation of incoming regulation especially as both ex-ante and ex-post governance measures gain popularity and significance. Large tech companies (often themselves hit by controversies around bias) have introduced ethics boards, built AI governance around existing governance structures and/or deployed debiasing techniques to address some of the issues.

For example, Microsoft has developed six AI principles to accelerate this cultural shift and to improve employees' awareness of ethical issues.[77] These include fairness, reliability and safety, privacy and security, inclusiveness, transparency and accountability. Governance is constituted by three core teams with the purposes of enacting the core principles, management of policy, governance, enablement, and sensitive use functions, and leading the implementation of responsible AI processes in the adoption of systems and tools.

IBM has developed and implemented AI Fairness 360[78], an open-source toolkit used to examine, report, and mitigate discrimination and bias in machine learning models. The main objectives of this toolkit are to help facilitate the transition of fairness research algorithms for use in an industrial setting and to provide a common framework for fairness researchers to share and evaluate algorithms.

Amazon has integrated new tools to assist in detecting discrimination in AI and ML technologies. As part of the cloud computing offering Amazon Web Services, a new test has been implemented alongside a wider suite of materials to customers seeking to develop fair, non-biased AI on the platform. The test was developed by Professor Sandra Wachter, Dr Brent Mittelstadt and Dr Chris Russell from the Oxford Internet Institute of the University of Oxford and it is called 'the Conditional Demographic Disparity (CDD)', a new test for "ensuring fairness in algorithmic modelling and data driven decisions".[79]

The developers of the image generation AI 'DALLE·2' have implemented a bias mitigation technique after evidence of representational harm in image outputs mounted. While generic prompts such as 'CEO' and 'builders' mostly generated images of men, prompts such as 'flight attendant' and 'nurse' generated images representing almost exclusively women.[80] The developers acknowledge how such stereotypes can be harmful, for instance when harming the dignity of protected groups, erasing them from socially valued situations, and enforcing mental representations of segregated social roles.[81] Stereotypical image outputs, in turn, contribute to confirming societal prejudices and feed into allocative harms, influencing the distribution of valuable social goods. The mitigation technique implemented by the

[77] Microsoft AI Principles, available at: https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6 (last accessed: 4 October 20022)
[78] IBM, introducing AI Fairness 360, available at: https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/ (last accessed: 4 October 2022).
[79] *AI modelling tool developed by Oxford academic incorporated into Amazon anti-bias software*, Oxford Internet Institute, 21 April 2021, available at: https://www.oii.ox.ac.uk/news/releases/ai-modelling-tool-developed-by-oxford-academics-incorporated-into-amazon-anti-bias-software-2/ (last accessed 29 September 2022)
[80] See OpenAI, "Reducing Bias and Improving Safety in DALL·E 2" (18 July 2022) available at: https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/.
[81] Pamela Mishkin et al, "DALL·E 2 Preview - Risks and Limitations" (2022) available at: https://github.com/openai/dalle-2-preview/blob/main/system-card.md#bias-and-representation.

developers of DALLE·2 seems to increase the diversity of population groups represented in image outputs. However, criticisms have been expressed towards the fact that diversity-related terms such as 'women' or 'black' were simply added to generic prompts to increase representativeness, thereby treating some of the symptoms of algorithmic bias without treating its root causes.[82]

While these are positive examples of existing governance efforts addressing algorithmic discrimination in the industry, it is important to highlight their limits.

The limits of technical solutions: debiasing and bias mitigation

First, **technical debiasing and bias mitigation solutions** cannot solve the problem of algorithmic discrimination in their own. As forcefully pointed out by Balayn and Gürses, "[d]ebiasing relies on conceptualisations of bias that do not capture the complexity of discrimination due to the limitations of the machine learning set-up."[83] Debiasing cannot redress algorithmic discrimination in a comprehensive or effective manner for two main reasons: On the one hand, these techniques focus exclusively on inputs and outputs of AI systems **without considering the context in which they are put to use**.[84] Debiasing techniques are algorithm-centric and **fail to consider the machine-human interaction** points that are also a source of bias.[85]

On the other hand, **debiasing techniques themselves have not yet reached a development stage that allows for deployment across the board**: "[the] use cases are limited, the proposed conceptualisations of bias can oversimplify matters of discrimination, and the effectiveness and usability of debiasing methods and auditing tools are yet to be established".[86] The practical application of debiasing techniques is also a challenge because of difficulties surrounding the access to sensitive data as well as contextual variations across use cases.[87] For instance, anti-discrimination law might require different conceptions of fairness to intervene across different use cases or at different stages of the same use case, which are difficult to translate into technical metrics as well as difficult to reconcile with each other.

This leads to the **question of what it means for an algorithm to be 'fair'?** A vast amount of research in computer science is dedicated to algorithmic 'fairness'. Fairness approaches are sometimes presented as being able to ensure the ethical and legal compliance of algorithmic systems. Yet, **'bias' and 'fairness' are technical notions that do not neatly overlap with their ethical and legal counterparts**. In discrimination law, in particular, the prohibition on bias will be limited to those targeting or otherwise negatively impacting protected groups. Removing such biases at one point of the AI lifecycle might yield fairness from a technical perspective, nevertheless that might not adequately satisfy existing legal obligations pertaining to equality throughout the AI lifecycle.

---

[82] Matthew Sparkes, "AI art tool DALL-E 2 adds 'black' or 'female' to some image prompts", New Scientist (22 July 2022) available at: https://www.newscientist.com/article/2329690-ai-art-tool-dall-e-2-adds-black-or-female-to-some-image-prompts/.

[83] Balayn A and Gürses S, Beyond Debiasing: Regulating AI and its inequalities (European Digital Rights 2021), 51 available at: https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf.

[84] See ibid, 12, 64.

[85] See ibid, 50.

[86] Ibid, 12, 50.

[87] See ibid.

In addition, computer scientists have developed a **wide range of definitions of fairness**, some of which are contradictory. Hence, **depending on the definition, an algorithm might be technically fair without necessarily complying with anti-discrimination law**.[88] For example, does fairness mean giving everyone the "same opportunity" while ignoring their wildly different starting points, or recognising the differences between people and giving some individuals a temporary advantage to counterbalance a disadvantage?[89] From a mathematical standpoint, there are several ways to achieve a fair outcome, and they all relate to different perceptions and interpretations of fairness itself. It could be argued for example, that treating a minority applicant the "same" when it comes to the provision of a loan may be fair. However, if due to historic and entrenched racism, that minority group has a higher risk of losing a job and thus being unable to repay the loan through no fault of their own, the application of fairness as simply the equalisation of outputs may lead to further entrenchment of inequality as those applicants may see their credit ratings further reduced.

Definitions of 'fairness as accuracy' and debiasing techniques aiming at acquiring *more* data and building *more* accurate algorithmic systems also present important limits. While so-called "**accuracy-affecting injustices"** stemming from issues pertaining to data representativeness, data collection and data processing practices can be resolved via changes to data policies aiming to increase accuracy in algorithmic decision-making,[90] biases resulting from past injustices require different types of solutions. So-called **"nonaccuracy-affecting injustices"** give rise to data biases that cannot be addressed via improvements in data collection practices.[91] They reflect facts that are accurate but problematic because resulting from historical discrimination and exclusion. **Only policies targeting the root causes and effects of such inequality can redress this type of bias.** For example, if an HR service wanted to automatise recruitment by predicting which candidates would be top performers, integrating more data about past recruitments will not address the causes of gender bias, which lie in gender segregation on the labour market, glass ceiling issues, the gender pay gap, gender stereotypes, etc.

Because of these limitations, solutionist narratives of debiasing should be debunked. **If at all, debiasing can only be one element of a broader anti-discrimination strategy** in relation to algorithmic systems. **Such a strategy should centre human rights and take into account the whole deployment cycle of algorithmic decision-making systems** ranging from the formulation of the problem to address, to the context of implementation of the system, its actual performance and its practical impact. In addition, as pointed out by Balayn and Gürses,

---

[88] See the discussion around differing ways of measuring bias and diverge definitions of fairness in the example of the COMPAS recidivism risk prediction system: Angwin, Julia, et al. "Machine bias." *Ethics of Data and Analytics*. Auerbach Publications, 2016. 254-264 and Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, "How We Analyzed the COMPAS Recidivism Algorithm" (2016) ProPublica available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[89] With a view to substantive and transformative equality, so-called temporary special measures or positive action provide special support or a provisional advantage to a disadvantaged group so as to transform an unequal status quo in the long-term. See the discussion in section 3 of this study.

[90] Hellman, Deborah. "Big Data and Compounding Injustice." *Journal of Moral Philosophy, forthcoming, Virginia Public Law and Legal Theory Research Paper* 2021-27 (2021).

[91] Ibid.

AI service providers should not enjoy wide discretion in choosing the strategies to prevent the discriminatory impact of their systems.[92] Rather, **democratic control and regulatory safeguards should establish a framework around accepted fairness and anti-discrimination approaches, taking full account of technical limitations and of the need to address the root causes of algorithmic discrimination**. The participation of end-users directly affected by these systems, and in particular minority groups, should also be ensured. As highlighted in our recommendations below, this should also apply to standard-setting activities.

The limits of bias audits: access to data and diverging standards

Second, **auditing biases** has been presented as another potential solution to address algorithmic discrimination. Yet, problems arise in relation to access to data and diverging standards.

Auditing is defined as "a range of approaches to review algorithmic processing systems" which "can take different forms, from checking governance documentation, to testing an algorithm's outputs, to inspecting its inner workings".[93] It has been suggested that auditing could be used as a preventive safeguard against the release of discriminatory algorithmic systems on the market.[94] However, the **lack of access to equality data**, **GDPR-related uncertainties** on the permitted processing of sensitive categories of data and **the lack of uniformly accepted standards** makes auditing algorithms for discrimination challenging.

On the one hand, legal scholars are **uncertain about whether the GDPR allows processing sensitive categories of personal data for debiasing** or more broadly for **anti-discrimination purposes**.[95] On the other hand, the **lack of equality data,** stemming from often restrictive equality data collection practices in Europe, raises issues when it comes to identifying inequality in specific domains such as access to housing, education, healthcare, work, etc. for various protected groups of population.[96] It limits access to accurate information about ground truth and the extent of structural inequality in society.

This problem of accessing sensitive data should also be considered in the broader context of **data extraction and exploitation** by big tech firms. **Access to such data for discrimination auditing and anti-discrimination purposes in general should therefore be entrusted to**

[92] See ibid, 11.

[93] Digital Regulation Cooperation Forum, "Auditing algorithms: the existing landscape, role of regulators and future outlook" (2022) available at: https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook.

[94] See Kim PT, 'Auditing Algorithms for Discrimination' (2017) 166 University of Pennsylvania Law Review Online 189.

[95] Van Bekkum, Marvin and Zuiderveen Borgesius, Frederik, Using Sensitive Data to Prevent Discrimination by AI: Does the GDPR Need a New Exception? (2022) available at: http://dx.doi.org/10.2139/ssrn.4104823 (last accessed 28 July 2022).

[96] See European Commission, *Analysis and comparative review of equality data collection practices in the European Union : legal framework and practice in the EU Member States* (Publications Office, 2017) available at: https://data.europa.eu/doi/10.2838/6934 (last accessed 28 July 2022); Lilla Farkas, *Analysis and comparative review of equality data collection practices in the European Union : data collection in the field of ethnicity* (Publications Office, 2020) available at: https://data.europa.eu/doi/10.2838/447194 (last accessed 28 July 2022); Ringelheim, Julie, "Processing Data on Racial or Ethnic Origin for Antidiscrimination Policies: How to Reconcile the Promotion of Equality with the Right to Privacy?" (2007) NYU School of Law Jean Monnet Working Paper No. 08/06, available at: http://dx.doi.org/10.2139/ssrn.983685 (last accessed 28 July 2022).

other entities, possibly including equality bodies, labour inspectorates, CSOs with a legitimate interest in the sense of Art. 11 and 12 and Art. 13 and 14 of the EU equality directives 2000/43/EC and 2000/78/EC, etc. The development of **more systematic, ethical and regulated equality data collection** throughout Europe would also be a progress for algorithmic auditing purposes. Inspiration could come from the UK, where, as part of the "Data: a new direction strategy" policy,[97] the Government has announced that a new condition will be introduced under the Data Protection Act (DPA) 2018 to allow for the processing of special category data for the monitoring and mitigation of algorithmic bias.

Furthermore, there are neither legal obligations nor uniform standards for algorithmic auditing yet. Various methodologies have been proposed.[98] Some of the toolkits developed by researchers have been adopted by major companies, for instance the 'Aequitas' instrument developed at the Oxford Internet Institute and adopted by Amazon.[99] Nonetheless, developing uniform regulatory standards for algorithmic auditing in the field of non-discrimination would substantially increase legal certainty for providers. This would also foster public trust in algorithmic systems. Finally, uniform regulatory standards for algorithmic auditing would enhance companies' take up of discrimination audits, which would in turn provide useful information for potential victims to assess the opportunity of taking (legal) action and comprehensible evidentiary material to judges.

6) Representation and participation issues: The lack of diversity and inclusion in the AI industry

The under-representation of minority groups in professional communities involved with the development of AI is an important dimension of the problem of algorithmic discrimination. The lack of diversity and inclusion in these communities means that under-represented groups do not (sufficiently) participate in the crafting of algorithmic technologies, with the **consequence** that **they cater suboptimally to the needs of these groups, disadvantages them or even erases them entirely.** A survey issued by the Council of Europe for the purpose of the present Study shows that **most responding State Parties to the ECHR are aware of the diversity issue** in the AI industry. **State Parties highlight the need to steer more women towards STEM disciplines** as this is perceived as a major factor contributing to discriminatory AI.

Some notable examples of AI bias due to lack of diversity have been exposed in a report by the AI Now Institute, founded by ex-Google executive Meredith Whittaker and principal

[97] Data: a new direction - government response to consultation, 22 June 2022, available at: https://www.gov.uk/government/consultations/data-a-new-direction/outcome/data-a-new-direction-government-response-to-consultation (last accessed: 28 July 2022)

[98] For a review, see e.g. Jack Bandy, (2021) 'Problematic Machine Behaviour: A Systematic Literature Review of Algorithm Audits.' Forthcoming, Proceedings of the ACM (PACM) Human-Computer Interaction, CSCW '21.

[99] See Saleiro, P, Kuester, B, Hinkson, L, London, J, Stevens, A, Anisfield, A, Rodolfa, KT, Ghani, R (2018) 'Aequitas: A Bias and Fairness Audit Toolkit.' Arxiv and Oxford Internet Institute (2021) 'AI modelling tool developed by Oxford Academics incorporated into Amazon anti-bias software' available at: https://www.oii.ox.ac.uk/news-events/news/ai-modelling-tool-developed-by-oxford-academics-incorporated-into-amazon-anti-bias-software-2/.

researcher at Microsoft Research Kate Crawford.[100] These include image recognition services which classified black people as gorillas and Amazon technology failing to recognize users with darker skin colours. The thesis of the report (reflecting a widely held view in the broader academic, policy and AI community) is that examples such as these occur due to "blind spots" because developers design and test models based on their own standpoint. The lack of a diverse workforce leads to a limited perspective and can result in bias that may be difficult to detect and correct before it leads to discrimination.

In addition to the widespread problem of **implicit bias**, a homogenous group is likely to have a **truncated outlook influenced by similar identities and experiences**. As an example, the Google AI Experiments programme developed a game called "Quick, Draw!" In the game, people were asked to draw pictures of everyday things like shoes to train a model.[101] All five of the game's developers at Google were men. They and early users of the game drew men's sneakers to represent a shoe. This resulted in a game which did not know that high heels were also shoes. This was not an intentional error; it was simply shaped by the perspective of the dominant representative group designing algorithms in the technology industry. **As such, any algorithm built by a majority group is at risk of failing to embed perspectives of marginalised minority groups, resulting in algorithms that only work for the majority**.

Diversity matters as it provides holistic approaches in making AI technologies more responsible. It helps address challenges faster and clearer as local knowledge and front-line experience will be embedded in the core of every decision-making or working process. Getting the right mix of minds in the room is essential to gain the necessary insight to address bias and gain competitive advantage. **Diversity should therefore be viewed as being "mission critical" when it comes to innovation.** This should translate in more diverse recruitment policies in educational and professional communities involved with the development and use of AI systems. As argued in section 3, legal obligations revolving around the notion of positive action could play a major role in this regard. In addition, diversity policies in educational and professional recruitment should be complemented by adequate training.

The *AI Now* (New York University) report[102] identified a "diversity crisis" in the AI sector, especially in the global technology industry, which is overwhelmingly white and male, and asserts that this has contributed to algorithmic gender and racial biases. A 2020 World Economic Forum report[103] painted a similarly grim picture: despite talk of greater inclusion, women's representation in tech-related jobs has declined by 32% since 1990. According to a study launched by the EU Commission in 2016, "only 24 out of every 1000 female graduates

[100] Sarah Myers West, Meredith Whittaker and Kate Crawford, *Discriminating Systems: Gender, Race, and Power in AI*, AI Now Institute NYU, April 2019, available at: https://ainowinstitute.org/discriminatingsystems.pdf (last accessed: 27 July 2022).

[101] Josh Lovejoy, *Fair Is Not the Default – Why building inclusive tech takes more than good intentions*, 15 February 2018, https://design.google/ library/fair-not-default/ (last accessed: 28 July 2022).

[102] Kari Paul, *'Disastrous' lack of diversity in AI industry perpetuates bias, study finds*, The Guardian, 17 April 2019, available at: https://www.theguardian.com/technology/2019/apr/16/artificial-intelligence-lack-diversity-new-york-university-study (last accessed: 27 July 2022).

[103] Ronit Avi and Rana El Kaliouby, *Here's why AI needs a more diverse workforce,* World Economic Forum, 21 September 2020 https://www.weforum.org/agenda/2020/09/ai-needs-diverse-workforce/ (last accessed: 27 July 2022).

had an ICT related subject in her portfolio". When it comes to employment, only 6 of those girls and women finally found a job in the digital sector. [104]

A Canadian start-up found that women make only 12% of leading machine learning researchers.[105]Another report[106] by New York University - *Discriminating Systems — Gender, Race, and Power in AI* asserts discrimination in AI systems was associated with the lack of diversity in the teams that work these technologies. **Whether the focus is on mitigation of bias in input processes, or fairness in outcomes, diversity and inclusion is one of the most powerful tool companies have at their disposal.** The blind spots created by the lack of diversity – diversity of education, perspectives, life experiences and backgrounds – make it more challenging to anticipate biases in algorithmic systems and their potential impact on different individuals and groups.

Already marginalised groups are systematically and disproportionately put more at risk of being harmed by algorithmic decision-making tools that do not represent their perspectives and interests. Beyond the moral imperative of preventing systemic racial and gender discrimination in designing new AI tools, there is also an economic one. Research has demonstrated that "companies in the top quartile for gender diversity have been 21% more likely to experience above-average profitability, while ethnic and cultural diversity correlates with a 33% increase in performance."[107]

Section 2

The legal and policy landscape in Europe: strengths and shortcomings

There is **general awareness among policy makers** that, alongside opportunities, AI brings the risks of solidifying and perpetuating existing inequalities. In a survey issued by the Council of Europe to gauge the views of State Parties to the ECHR, more than **80% of respondents viewed AI as posing risks to human rights**. **40% of respondents identified a direct risk of gender discrimination**.

Several initiatives are taking place across governments, and they encompass several issues, from female participation in STEM fields, to deepfakes and cyberbullying and algorithmic discrimination. For example, some countries, like **Finland**, have addressed the issue of the lack of transparency in algorithmic systems leading to discrimination head on, issuing

[104] Women in AI: Promoting inclusive participation across society, Aimee Van WYNSBERGH, European AI Alliance, available at: https://futurium.ec.europa.eu/en/european-ai-alliance/blog/women-ai-promoting-inclusive-participation-across-society?language=hu (last accessed: 31 August 2022).
[105] Archie de Berker, Women in Machine Learning: Negar Rostamzadeh, 20 February 2018, available at: https://medium.com/element-ai-research-lab/women-in-machine-learning-negar-rostamzadeh-dbb58dc75e81 (last accessed: 31 August 2022).
[106] Sarah Myers West, Meredith Whittaker and Kate Crawford, *Discriminating Systems: Gender, Race, and Power in AI,* AI Now Institute NYU, April 2019, available at: https://ainowinstitute.org/discriminatingsystems.pdf (last accessed: 27 July 2022).
[107] The five business benefits of a diverse team, CMI, 3 July 2019, available at: https://www.managers.org.uk/knowledge-and-insights/listicle/the-five-business-benefits-of-a-diverse-team/ (last accessed: 31 August 2022).

recommendations and guidance to raise awareness of the problem.[108] The **Netherlands** has adopted a 'Fundamental rights and algorithms Impact Assessment' that includes a 'Non-discrimination by design guideline'.[109] The Dutch Parliament has recently adopted a motion rendering human rights impact assessments compulsory for public institutions using algorithms.[110]

The **Austrian** government has published an action plan on deepfakes, including diverse measures to tackle the problem. In **Finland**, Aurora AI aims to guide citizens, especially young people, to the services they need by means of artificial intelligence. If, as a result, young people find the services they need better, this is likely to promote equality, for example in access to services or in the provision of assistance and support. The **Portuguese** Agency for Administrative Modernisation (AMA) has developed - with the help of the Commission for Citizenship and Gender Equality and other relevant stakeholders - the "Guide for the use of Artificial Intelligence in Public Administration". The guide is designed to address the concerns of non-discrimination in general and the protection of individual and collective rights in the development of algorithmic systems. It draws attention to the reliability and representativeness of the data to be collected and processed, and emphasizes the issues associated with ethics, justice, transparency, accountability and understanding of the systems.

Yet, national responses are largely **uncoordinated**. While legislators such as the European Union are in the process of adopting a uniform regulatory framework on AI, **the Council of Europe could exert wide-ranging regulatory influence in the field of human rights**. Where the EU is advocating for a 'human-centric AI', regulatory action by the Council of Europe could foster a distinct **human-rights-based approach to AI.**

This section of the Study highlights **which existing legal instruments at Council of Europe level can be used to address various dimensions of the problem of algorithmic discrimination**, ranging from **non-discrimination to data protection and privacy law to sectoral regulations**. It also briefly maps existing and forthcoming EU legal instruments and shows that both frameworks present **shortcomings and uncertainties when it comes to addressing algorithmic discrimination**. These gaps call for regulatory action at Council of Europe level, some possible contours of which are highlighted in Section 3.

I. Discrimination and equality: legal and policy instruments and their limits

This section highlights the existing legal instruments that provide a legal basis for combating algorithmic discrimination and related forms of algorithmic violence.

1) Binding legal instruments of the Council of Europe

[108] Automaattisessa päätöksenteossa on turvattava virkavastuu ja hyvän hallinnon toteutuminen, available at: https://valtioneuvosto.fi/-/10623/automaattisessa-paatoksenteossa-on-turvattava-virkavastuu-ja-hyvan-hallinnon-toteutuminen (last accessed: 28 July 2022).

[109] Ministry of the Interior and Kingdom Relations, 'Fundamental rights and algorithms Impact Assessment' (March 2022) available at: https://www.government.nl/binaries/government/documenten/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms/Fundamental+Rights+and+Algorithms+Impact+Assessment.pdf.

[110] See European Centre for Not-for-Profit Law, "Netherlands sets precedent for human rights safeguards in use of AI" (2022) available at: https://ecnl.org/news/netherlands-sets-precedent-human-rights-safeguards-use-ai.

The European Convention on Human Rights

Article 14 ECHR and Art. 1 of Protocol No. 12 lay out a prohibition on discrimination that provides a **legal basis for banning algorithmic discrimination**.

Article 14 ECHR prohibits discrimination based on an open-ended list of protected characteristics:

> *"The enjoyment of the rights and freedoms set forth in [the] Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status."*

Its application depends on the existence of a violation of a substantive right protected by the ECHR.

Entered into force in 2005, **Protocol No. 12 to the Convention** has been ratified by 20 out of 46 State parties to the ECHR so far. Article 1 lays out a free-standing general prohibition of discrimination:

> *"1. The enjoyment of any right set forth by law shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.*
>
> *2. No one shall be discriminated against by any public authority on any ground such as those mentioned in paragraph 1."*

The Istanbul Convention

The Council of Europe Convention on preventing and combating violence against women and domestic violence provides a **basis for prohibiting algorithmic violence against women, including algorithmic stereotyping, and online violence such as cyber-harassment, bullying and online sexist hate speech**.

The Istanbul Convention was adopted in 2011, entered into force in 2014 and has been ratified by 37 state parties. It recognises gender-based violence (GBV) as a form of discrimination. Its provisions focus on prevention, protection, prosecution and the development of integrated policies in relation to combating violence against women.

Particularly relevant to issues of online gender-based violence is **Art. 17** on "participation of the private sector and the media" which states that:

> *"Parties shall encourage the private sector, the information and communication technology sector and the media, with due respect for freedom of expression and their independence, to participate in the elaboration and implementation of policies and to set guidelines and selfregulatory standards to prevent violence against women and to enhance respect for their dignity".*

The Framework Convention for the Protection of National Minorities

The Framework Convention for the Protection of National Minorities provides a **legal basis for combatting algorithmic discrimination on grounds of national minority status as well as online violence such as hate speech**.

Entered into force in 1998, the Convention counts 39 State parties. In its **Art. 4**, the Convention states that:

> *"1. The Parties undertake to guarantee to persons belonging to national minorities the right of equality before the law and of equal protection of the law. In this respect, any discrimination based on belonging to a national minority shall be prohibited.*
>
> *2. The Parties undertake to adopt, where necessary, adequate measures in order to promote, in all areas of economic, social, political and cultural life, full and effective equality between persons belonging to a national minority and those belonging to the majority. In this respect, they shall take due account of the specific conditions of the persons belonging to national minorities."*

Art 6(2) lays out that "*The Parties undertake to take appropriate measures to protect persons who may be subject to threats or acts of discrimination, hostility or violence as a result of their ethnic, cultural, linguistic or religious identity."*

Art. 9 relating to freedom of expression, which states that *"The Parties shall ensure, within the framework of their legal systems, that persons belonging to a national minority are not discriminated against in their access to the media"*, could become particularly relevant for issues of discrimination on social media platforms, cyberharassment and hate speech.

The European Charter for Regional or Minority Languages

Entered into force in 1998, the Charter has been ratified by 25 countries so far. **Art 7(2)** of the Charter lays out that *"The Parties undertake to eliminate, if they have not yet done so, any unjustified distinction, exclusion, restriction or preference relating to the use of a regional or minority language and intended to discourage or endanger the maintenance or development of it".* Again, in principle **this provision extends to the algorithmic and online realms, where it can be relied on to address digital discrimination in its many forms.**

2) Relevant policy instruments of the Council of Europe

A number of **non-binding standards and policy instruments** complement the binding legal provisions and are **relevant when it comes to addressing the discriminatory effects of AI and algorithmic decision-making**.

In March 2019, the "**Recommendation on Preventing and Combating Sexism**" drafted by the Gender Equality Commission was adopted by the Council of Ministers.[111] It recognises that "[t]he internet has provided a new dimension for the expression and transmission of sexism, especially of sexist hate speech, to a large audience, even though the roots of sexism do not lie in technology but in persistent gender inequalities".[112] It enjoins Member States to "integrate a gender equality perspective in all policies, programmes and research in relation to artificial intelligence to avoid the potential risks of technology perpetuating sexism and gender

[111] Council of Europe, "Preventing and combating sexism", Recommendation CM/Rec(2019)1 adopted by the Committee of Ministers of the Council of Europe (27 March 2019), available at https://rm.coe.int/prems-055519-gbr-2573-cmrec-2019-1-web-a5/168093e08c.
[112] Ibid.

stereotypes".[113] The recommendation also foresees a positive role for AI as it requires State Parties to "examine how artificial intelligence could help to close gender gaps and eliminate sexism".[114] It lists key aspects such as women's and girls' participation in IT education and industries, the mainstreaming of gender equality in the design of data-driven instruments, awareness-raising as regards gender bias in big data, transparency and accountability. In turn, the recent recommendation **"On combating hate speech"** co-drafted by the Steering Committee on Anti-Discrimination, Diversity and Inclusion (CDADI) and the Steering Committee on Media and Information Society (CDMSI) indicates that "internet intermediaries should identify expressions of hate speech that are disseminated through their systems and act upon them in the framework of their corporate responsibility".[115]

The **Council of Europe Gender Equality Strategy 2018-2023** also recognises that "sexism and discrimination against women includ[e] **sexist hate speech online**" as well as online gender-based violence.[116] In addition, GREVIO, which monitors the implementation of the Istanbul Convention, also published a General Recommendation No.1 on the **digital dimension of violence against women** in 2021 which highlights legal issues around online sexual harassment, stalking and the digital dimension of psychological violence.[117]

In May 2022, the Committee of Ministers adopted a new **Recommendation on combating hate speech** jointly drafted by the Steering Committees on Anti-Discrimination, Diversity and Inclusion (CDADI) and on Media and Information Society (CDMSI).[118] It recognises the existence of a "**power asymmetry between some digital platforms and their users**" and makes recommendations for tackling **online hate speech in relation to policies pertaining to content moderation, micro-targeting and online advertising, content amplification, recommender systems and underlying data collection strategies**.

In May 2022, the Committee of Ministers adopted a "**Recommendation on protecting the rights of migrant, refugee and asylum seeking women and girls**" which **demands that human rights impact assessments are conducted before AI and automated decision making systems are introduced in the field of migration** and that **the design, development and application of such systems are non-discriminatory**.[119] It also calls for involving refugee, asylum-seeking and migrant women and representative CSOs "in discussions on the development and deployment of new technologies affecting them".

Other instruments such as the **Guidelines of the Committee of Ministers of the Council of Europe on upholding equality and protecting against discrimination and hate during the**

[113] Recommendation II.B.7, ibid, p. 19.
[114] Ibid.
[115] Council of Europe, Recommendation CM/Rec(2022)16[1] of the Committee of Ministers to member States on combating hate speech (20 May 2022), [30].
[116] Council of Europe Gender Equality Strategy 2018-2023 adopted by the Committee of Ministers (March 2018), p. 10, 16, 18, available at https://rm.coe.int/prems-093618-gbr-gender-equality-strategy-2023-web-a5/16808b47e1.
[117] Group of Experts on Action against Violence against Women and Domestic Violence, General Recommendation No. 1 on the digital dimension of violence against women (20 October 2021) available at: https://rm.coe.int/grevio-rec-no-on-digital-violence-against-women/1680a49147 (last accessed 22 July 2022).
[118] Council of Europe, Recommendation CM/Rec(2022)16 on combating hate speech, available at https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680a67955#_ftn1.
[119] Council of Europe, Recommendation CM/Rec(2022)17 of the Committee of Ministers to member States on protecting the rights of migrant, refugee and asylum-seeking women and girls, [22]-[25] available at https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680a69407.

Covid-19 pandemic and similar crises in the future mention the need to ensure that "digital tools for dealing with the crisis and the resulting risks" "are not discriminatory against persons belonging to vulnerable groups or otherwise violate their rights".[120]

Together, **these recommendations address** a number of **issues contributing to algorithmic discrimination** as pointed out earlier in this Study: **the lack of diversity, equal representation and equal participation in educational and professional fields related to the AI industry, the lack of binding obligation to mainstream equality-related concerns in the development of algorithmic systems and the lack of clearly defined accountability mechanisms**.

3) Comparative insights: other relevant European and international provisions

The European Union also has a very developed legal framework on discrimination and equality. **Art 21(1)** of the **EU Charter of Fundamental Rights** prohibits discrimination "*on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation*" and **Art. 21(2)** states that "[w]ithin the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited". **Art. 23** indicates that "[e]quality between women and men must be ensured in all areas, including employment, work and pay" and allows positive action. In secondary law, **Directive 2000/43/EC** guarantees equality on grounds of race or ethnic origin at work, in the access to goods and services and in education. **Directive 2000/78/EC** prohibits discrimination on grounds of disability, sexual orientation, religion or belief and age in the workplace and vocational training. **Directive 20004/113/EC** guarantees gender equality in the access to goods and services and so does **Directive 2006/54/EC** in relation to work.

In 2022, the European Commission published a "**European Declaration on Digital Rights and Principles for the Digital Decade**" that reflects the Commission's wish to develop a "**human-centred AI**" and exposes the EU's approach to digital transformation. The Commission's rationale is that digital rights should ensure that EU citizens have access to digital technologies and are protected from their harmful consequences. Chapter III of the Declaration includes a commitment to "ensuring that algorithmic systems are based on suitable datasets to avoid unlawful discrimination and enable human supervision of outcomes affecting people".[121]

At United Nations level, a number of instruments protect against discrimination beyond existing general human rights instruments: in particular the **International Convention on the Elimination of All Forms of Racial Discrimination** (CERD), the **Convention on the Elimination of All Forms of Discrimination against Women** (CEDAW), and the **Convention on the Rights of Persons With Disabilities** (CRPD). More specifically, the CERD Committee issued a **General recommendation No. 36 on preventing and combating racial profiling by law enforcement officials** in 2020. This document recognises how the use of artificial

[120] Steering Committee on Anti-Discrimination, Diversity And Inclusion (CDADI), Guidelines of the Committee of Ministers of the Council of Europe on upholding equality and protecting against discrimination and hate during the Covid-19 pandemic and similar crises in the future (2020), [27] available at: https://rm.coe.int/prems-066521-gbr-2530-cdadi-guidelines-web-a5-corrige/1680a3d50c.

[121] European Commission, "European Declaration on Digital Rights and Principles for the Digital Decade" COM(2022) 28 final (Brussels 2022).

intelligence leads to entrenching racial inequalities and makes recommendations to prevent and redress racial bias and discrimination.

Although these legal and policy instruments do not stop at the borders of the digital world, their applicability to the various forms of algorithmic discrimination suffers a number of shortcomings.

4) Limits and uncertainties: where does algorithmic discrimination fall into the cracks?

This legal and policy patchwork addresses some of the discriminatory risks of AI and automated decision-making. Yet, **many uncertainties remain concerning the extent to which existing legal provisions can be used to promote equality and counter discrimination arising from the use of these technologies.** Hence, the aim of this subsection is to explore existing gaps in the equality and non-discrimination framework described above when it comes to algorithmic discrimination. **Three main issues** arise: (1) the **lack of neat overlap between existing concepts of direct and indirect discrimination and forms of algorithmic discrimination**; (2) **procedural issues** linked to **evidence** and **responsibility**; (3) **challenges linked to the protection of specific characteristics by the law**. As explained in Section 3, addressing those gaps calls for enforcing existing positive obligations to promote equality and mainstreaming preventive approaches to algorithmic discrimination under Art. 14 ECHR.

Qualification issues: direct vs indirect algorithmic discrimination

Although Article 14 ECHR does not distinguish between direct and indirect discrimination, the European Court of Human Rights (the Court) carved out the distinction in its case law.[122] **Direct discrimination** arises from "a **difference in the treatment of persons in analogous, or relevantly similar, situations**" and where this difference is "**based on an identifiable characteristic" or "status"**.[123] For example, where two workers are similarly qualified for a promotion but one is preferred over the other "because of" their sex, this would give rise to direct sex discrimination.

At the beginning of the 2000s, the **Court recognised the existence of indirect discrimination** where states "**fail to treat differently persons whose situations are significantly different**".[124] It ruled in *DH* that "a difference in treatment may take the form of disproportionately prejudicial effects of a general policy or measure which, though couched

[122] This has been done by reference to EU equality law and the case law of the European Court of Justice, see *D.H. and Others v. The Czech Republic* Application no. 57325/00 (European Court of Human Rights, Grand Chamber, 13 November 2007), [184].

[123] See e.g. *Kjeldsen, Busk Madsen and Pedersen v. Denmark* Application no. 5095/71, 5920/72, 5926/72 (European Court of Human Rights, 7 December 1976), [56]; *Burden v. the United Kingdom* Application 13378/05 (European Court of Human Rights, Grand Chamber, 29 April 2008), [60]; *Carson and Others v United Kingdom* Application no. 42184/05 (European Court of Human Rights, Grand Chamber, 16 March 2010), [61], and more recently *Biao v. Denmark* Application no. 38590/10 (European Court of Human Rights, Grand Chamber, 24 May 2016), [89]. See also European Union Agency for Fundamental Rights and Council of Europe, *Handbook on European non-discrimination law* (Publications Office of the European Union 2018), 43 and European Court of Human Rights, *Guide on Article 14 of the European Convention on Human Rights and on Article 1 of Protocol No. 12 to the Convention* (Council of Europe 2020), 11.

[124] *Thlimmenos v. Greece* Application no. 34369/97 (European Court of Human Rights, 2 April 2000), [44].

in neutral terms, discriminates against a group".[125] For instance, a neutrally formulated policy that would make the recruitment of candidates conditional on a minimum height might have indirectly discriminatory effects on women, who are on average smaller than men.

Once a *prima facie* finding of direct or indirect discrimination has been established, an **open justification system** applies whereby **discrimination** can only be found **where there is "no objective and reasonable justification"**.[126] In other terms, both direct and indirect discrimination can be justified if it pursues a **legitimate aim** and if there is a "relationship of **proportionality between the means employed and the aim sought** to be realized".[127] Because the same justification regime applies in principle under both frameworks, qualifying algorithmic discrimination as direct or indirect has less significant repercussions on available means of redress than under EU law, where this qualification conditions the applicability of a closed or an open regime of justifications[128] Nonetheless, it is important to understand how courts, including the ECtHR, will qualify algorithmic discrimination.

So far, it has been argued that algorithmic discrimination mainly falls within the framework of indirect discrimination, in particular because developers are unlikely to input protected characteristics in the datasets used to train ADM systems.[129] According to Hacker, for example, "in machine learning contexts, indirect discrimination is the most relevant type of discrimination" while "[d]irect discrimination will be rare in algorithmic decision making, and largely limited to cases of implicit bias in labelling".[130] Borgesius and Kelly-Lyth also respectively argue that "non-discrimination law prohibits many discriminatory effects of algorithmic decision-making, in particular through the concept of indirect discrimination"[131] and that "most biased algorithms will fall under the indirect discrimination framework".[132]

At least three arguments support this view: (1) Indirect discrimination captures situations where formally neutral measures produce disadvantage because they intervene in, and embed, an unequal social context.[133] This resonates with the ways in which data-driven

[125] *D.H. and Others v. The Czech Republic* Application no. 57325/00 (European Court of Human Rights, Grand Chamber, 13 November 2007), [184].
[126] *Case "relating to certain aspects of the laws on the use of languages in education in Belgium" v. Belgium* Application no 1474/62; 1677/62; 1691/62; 1769/63; 1994/63; 2126/64 (European Court of Human Rights, 23 July 1968), [10] at 34.
[127] Ibid, see also *Marckx v. Belgium* Application no. 6833/74 (European Court of Human Rights, 13 June 1979), [33].
[128] Under EU law, direct discrimination cannot, in principle, be justified (safe closed exceptions), while indirect discrimination gives rise to a proportionality test with an open-ended regime of justifications.
[129] This argument builds on an analogy with the US anti-discrimination framework, see e.g., Solon Barocas and Andrew D. Selbst, 'Big Data's Disparate Impact' (2016) 104 California law review 671. Yet, the distinction between direct and indirect discrimination in ECHR law differs from the US distinction between notions of "disparate treatment" and "disparate impact".
[130] Hacker, 'Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law', 1152-1153.
[131] Zuiderveen Borgesius, 'Strengthening legal protection against discrimination by algorithms and artificial intelligence', 1578. He nevertheless acknowledges a range of enforcement issues.
[132] Aislinn Kelly-Lyth, 'Challenging Biased Hiring Algorithms' (2021) 41 Oxford Journal of Legal Studies 899, 906.
[133] See Tobler, *Limits and potential of the concept of indirect discrimination*, 85. On the perpetrator's vs. the victim's perspective, see Alan David Freeman, 'Legitimizing Racial Discrimination Through Antidiscrimination Law: A Critical Review of Supreme Court Doctrine' (1978) 62 Minnesota Law Review 1049.

technologies incorporate and perpetuate society's unequal *status quo*.[134] (2) Indirect discrimination focuses on the structural dimension of discrimination.[135] This focus resonates with the fact that machine learning (ML) algorithms derive rules from group patterns. Third, the concept of indirect discrimination allows addressing distinctions not based on legally protected grounds that in effect impact protected groups.[136] Since such proxy discrimination is one of the prevailing forms of algorithmic discrimination, as will be explained below, the framework of indirect discrimination presents a further advantage.

Despite the consensus on classifying algorithmic discrimination as indirect, **such a qualification "by default" raises a number of doctrinal and procedural issues**.[137] As recent research shows, the notion of **direct discrimination could capture some cases of algorithmic discrimination where a whole group is consistently impacted**, no matter the criterion used for decision-making.[138] Going further, **fitting the discriminatory effects of algorithmic bias within one or the other notion raises crucial normative questions** about key concepts of non-discrimination law.[139] In this sense, CAHAI recognised in its 2020 Feasibility Study that **"[t]he increased prominence of proxy discrimination in the context of machine learning may raise interpretive questions about the distinction between direct and indirect discrimination or, indeed, the adequacy of this distinction as it is traditionally understood".**[140] For example what can be considered a "neutral" criterion for decision-making in light of existing feedback loops and redundant encoding issues? Is algorithmic discrimination, which feeds structural inequality into individual decision-making, a collective or individual form of unfair treatment? Should the user of an algorithm be considered a perpetrator when a machine autonomously "learns" to discriminate? Answers to these questions will determine, in theory, whether the notion of direct or indirect discrimination can be used to capture algorithmic discrimination.[141]

---

[134] See Anna Lauren Hoffmann, 'Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse' (2019) 22 Information, Communication & Society 900.

[135] See Hugh Collins and Tarunabh Khaitan, 'Indirect Discrimination Law: Controversies and Critical Questions' in Hugh Collins and Tarunabh Khaitan (eds), *Foundations of Indirect Discrimination Law* (1 edn, Hart Publishing 2018), 19.

[136] For example, part-time work is a matter of gender equality where most part-time workers are women. See Tobler, *Limits and potential of the concept of indirect discrimination*, 24 and Janneke Gerards, 'Discrimination grounds', in: Dagmar Schiek, Lisa Waddington and Mark Bell (eds), Cases, Materials and Text on National, Supranational and International Non-Discrimination Law, Oxford and Portland, Oregon: Hart Publishing 2007, 33-184.

[137] Gerards J and Xenidis R, Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law (European network of legal experts in gender equality and non-discrimination / European Commission, 2021).

[138] See Adams-Prassl, Binns and Kelly-Lyth, "Directly discriminatory algorithms", Modern Law Review (forthcoming).

[139] Gerards J and Xenidis R, Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law (European network of legal experts in gender equality and non-discrimination / European Commission, 2021).

[140] CAHAI, "Feasibility Study on legal framework on AI design, development and application based on CoE standards" (2020), [13], p. 5.

[141] See Gerards J and Xenidis R, Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law (European network of legal experts in gender equality and non-discrimination / European Commission, 2021) and Xenidis R, 'Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience' (2021) 27 Maastricht Journal of European and Comparative Law 736.

Procedural issues: proof, proportionality, responsibility and liability

In practice, however, the opacity of algorithmic decision-making systems means that the evidence necessary to characterize direct discrimination will often be lacking. The information might only become available *ex post* and might remain partial, so that one might only be able to observe the effects of an algorithmic system after it has been used. For example, if a credit scoring algorithm systematically denies credit to people living with a disability, one might not have access to the criteria used for such a decision but might only be able to observe a pattern of rejection in relation to applicants with a disability. Similarly, one might not be able to access information regarding the entire pool of applicants, so that there might not be any certainty regarding potential applicants with a disability who have been granted a credit or other applicants who received a rejection.

Proof issues: For potential applicants, the opacity of algorithmic decisions amounts to substantial barriers to redressing discrimination. Information asymmetries between users and subjects of algorithmic decision-making or decision-support systems mean that isolated end users will not have the capacity to monitor the impact of algorithmic decisions on groups of other end users. They will not be able to access information about the decision-making criteria either. Even in potential cases of indirect algorithmic discrimination, the absence of transparent and meaningful information on relevant decision criteria and victims' lack of birds-eye view on decisions taken could prevent awareness that discrimination has occurred. This can eventually preclude any legal action from even being started.

When bringing cases to court, however, **applicants will be aided by existing rules on the burden of proof**: once a *prima facie* case of discrimination has been established by the applicant, in principle the burden of proof shifts to the defendant, who is responsible for showing that the difference in treatment is justified. **Yet, legal issues still arise: How to provide enough elements, and which type of information to adduce, to make a *prima facie* case** of discrimination so as to trigger the shift of the burden of proof onto the defendant? In the algorithmic context, information asymmetries might defeat even the possibility to show discrimination *prima facie.*

Proportionality test: Once a differential treatment between similarly situated persons or the absence thereof between differently situated persons has been established, judges must conduct a proportionality test to assess whether it can be objectively justified. This two-step test aims to find whether the practice fulfils a legitimate aim, and whether the means employed are reasonably proportionate to the aim pursued.[142] Answering these questions lead to considerable legal uncertainty because of the necessity for judges to assess technical trade-offs that might not be accessible or intelligible to them (e.g. which fairness metrics were to be used? How to balance trade-offs between various definitions of equity?[143] How to balance

[142] Registry of the European Court of Human Rights, Guide on Article 14 of the European Convention on Human Rights and on Article 1 of Protocol No. 12 to the Convention (30 April 2022) available at: https://www.echr.coe.int/Documents/Guide_Art_14_Art_1_Protocol_12_ENG.pdf (last accessed 22 July 2022).

[143] Equity is a philosophical and statistical term used to describe whether an algorithmic system treats different groups fairly. There are different definitions of equity (e.g. all groups get similar rates of false positives and negatives vs the performance of an algorithm is calibrated to be similar for all groups) that can be incompatible with each other. There is no neat overlap between the statistical term 'equity' and the legal term 'equal treatment'.

accuracy vs fairness? etc.).[144] The technical barriers arising here could contribute to shielding ADMS from judicial review. In these conditions, recent research points towards a permissive application of the proportionality test in the context of algorithmic opacity.[145]

Responsibility and liability: **The question of responsibility and liability for algorithmic discrimination is thorny. Some commentators argue that the law should allow for "an extension of the grounds for defence of respondents [which] could allow them to establish that biases were autonomously developed by an algorithm".[146] However, such an argument raises the difficult question of who should be held liable for algorithmic discrimination in the absence of legal personhood of AI systems. Moreover, the distribution of liability between AI providers and users (those deploying them) is another difficulty as both could bear responsibility for a discriminatory system. In light of the many sources of algorithmic bias, from data to model features and implementation, it is nearly impossible to identify a single and precise cause of algorithmic discrimination.**

Issues relating to the personal scope of non-discrimination law: the mismatch between algorithmic systems and protected grounds of discrimination

The last set of challenges that arises concerns the lack of overlap between the personal scope of non-discrimination legal provisions and the idiosyncratic forms of algorithmic subjectivity.

Proxy discrimination and the indirect discrimination route: Research shows that algorithmic discrimination takes place even when protected characteristics are removed from a given dataset. This is because algorithmic profiling relies on data points which, combined, can lead to clustering that overlaps with protected groups. For instance, commuting time between home and workplace or postcode could lead to inferences about socio-economic status and ethnicity given the existing spatialization of socio-economic and racial inequalities.[147] In particular, **redundant encoding** issues arise when variables in a dataset correlate with a protected category, for instance commuting time and ethnic background, which can be inferred by machine learning algorithms. This combines with issues of **feedback loops**, which describe situations where a system relies on data arising from past discrimination as a basis for predictions. Algorithmic discrimination is therefore very likely to take the form of **proxy discrimination**.

Article 14 ECHR bans discrimination "on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status". Proxy discrimination, for example based on **behavioural data such as screen time, wifi usage, geolocalisation data**, etc., could therefore be captured under Art. 14 ECHR **via the indirect discrimination route**, by showing a strong disadvantageous effect based on one of the grounds explicitly listed.[148] **The problem is that**

---

[144] See Binns R, 'Algorithmic Decision-making: A Guide For Lawyers' (2020) 25 Judicial Review 2.

[145] Pablo Martínez-Ramil, "Discriminatory algorithms. A proportionate means of achieving a legitimate aim?" (2022) Journal of Ethics and Legal Technologies 4(1).

[146] Grozdanovski L, 'In search of effectiveness and fairness in proving algorithmic discrimination in EU law' (2021) 58 Common Market Law Review, 99.

[147] See Williams BA, Brooks CF and Shmargad Y, 'How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications' (2018) 8 Journal of Information Policy 78.

[148] Proxy discrimination could in certain cases be treated as direct discrimination, depending on how the scope and boundaries of protected groups are delineated. For a discussion of this problem within the notion of direct

such proxy discrimination might escape the legal protection against discrimination because of the procedural difficulties exposed in the above section. [149]

"New" algorithmic groups and the notion of "other status": **In addition, algorithms can generate new categorizations based on seemingly innocuous characteristics, such as web browser preferences or apartment number, or more complicated categories combining many data points. For example, an online store may find that most consumers using a certain web browser pay less attention to prices; the store can charge those consumers extra. Despite not corresponding to criteria protected under non-discrimination law, some of these algorithmic groups might deserve legal protection, for example if patterns of algorithmic differentiation expose them to systematic socio-economic disadvantage.**

Where discrimination against algorithmic groups does not overlap with categories explicitly protected by Art. 14 ECHR, **the open-ended list of protected grounds in Art. 14 and the flexible approach of the European Court of Human Rights (the Court) towards protecting "new grounds" arguably provides an avenue for protection**.[150] It has been argued that "semi-open" anti-discrimination clauses such as Art. 14 ECHR provide better solutions for redressing algorithmic discrimination than fully closed discrimination provisions such as in EU secondary law.[151] For instance, the Court has protected groups on the basis of their professional status or place of residence.[152] This open-ended approach, based on the notion of "**other status**", could facilitate extending the coverage of new algorithmic groups under Art. 14 ECHR. Yet, this poses the question of the **normative limits of anti-discrimination law**: what are the contours of its mandate? What kinds of injustices is it meant to address?

Furthermore, **some algorithmic clusters lack social salience and are therefore difficult to depict as groups deserving protection from discrimination law**.[153] The "new" algorithmic groups emerging from intangible algorithmic clustering are subject to distinctions that have very tangible socio-economic effects and could consolidate into **"emergent" structural discrimination**.[154] By contrast to socially salient algorithmic groups, such distinctions will

discrimination in the EU context, see Xenidis R, 'Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience' (2021) 27 Maastricht Journal of European and Comparative Law 736.

[149] See e.g., Anton Vedder & Laurens Naudts (2017) Accountability for the use of algorithms in a big data environment, International Review of Law, Computers & Technology, 31:2, 206-224 and Naudts, L. (2019). How Machine Learning Generates Unfair Inequalities and How Data Protection Instruments May Help in Mitigating Them. In R. Leenes, R. van Brakel, S. Gutwirth & P. De Hert (Editors), Data Protection and Privacy: The Internet of Bodies (Computers, Privacy and Data Protection).

[150] See Gerards, Janneke, and Frederik Zuiderveen Borgesius. "Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence." *Colorado Technology Law Journal, forthcoming* (2020).

[151] Ibid.

[152] See Van der Mussele v. Belgium Application no. 8919/80 (European Court of Human Rights, 23 November 1983) and Carson and Others v United Kingdom (2010), [70]-[71].

[153] See Matthias Leese, The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union, 45 SECURITY DIALOGUE 494–511, 501 (2014); Monique Mann & Tobias Matzner, Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination, 6 BIG DATA & SOCIETY, 5–6 (2019).

[154] Ibid.

systematically escape ECHR equality law. Recent scholarship has proposed extending the scope of anti-discrimination law to cover such harmful algorithmic distinctions.[155]

A last problem pertaining to the personal scope of ECHR equality law arises when **algorithmic decision-making blurs the lines between the individual and the group**. In particular, group-based patterns are used to make decisions about individuals. This presupposes that membership into given algorithmic groups might be ascribed to individuals even when this is not factually correct. For instance, if a user displays the typical web traffic patterns of a woman between 25 and 30 residing in an urban environment, that gender and age identity might be assigned to them to serve as a basis for further decision-making. If that ascribed algorithmic cluster does not match the real user's identity, the user will not have any opportunity to correct the results of algorithmic profiling and ensuing treatment. However, if that user experienced gender-based discrimination, for example higher health insurance prices, they could claim "**discrimination by association",** a notion recognised by the Court in 2008.[156]

Intersectional discrimination: Finally, algorithmic discrimination is likely to be intersectional in nature, that is to involve several discrimination grounds or vectors of disadvantage.[157] Because of the granularity of algorithmic profiling, AI systems are able to infer several protected social memberships and potentially **cluster users according to different problematic classifications**. For example, algorithmic profiles might contain information regarding gender, age, ethnic background, religious beliefs, sexual orientation or gender identity based on the analysis of online behaviours, consumer preferences, etc. Identifying and redressing intersectional cases of algorithmic discrimination proves even more challenging than single-axis cases because of the lack of disaggregated equality data, which does not allow comparing potential disparities between algorithmic outputs and the actual situation of intersectionally marginalised groups.[158] Debiasing approaches also show limits when it comes to redressing the discriminatory consequences of biases affecting intersectional minorities.[159] Against this background, intersectional discrimination has often fallen into the cracks of judicial redress. Although the ECtHR has successfully (albeit implicitly) grappled with intersectional discrimination in a case like *BS v Spain*,[160] it has failed to recognise it explicitly and to redress it in others like *SAS v France* or *Garib v The Netherlands*.[161] This **lack of**

---

[155] See Wachter S, 'The Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law' (2022)Tulane Law Review (forthcoming) .

[156] Molla Sali v. Greece Application no. 20452/14 (European Court of Human Rights, 19 December 2018), [141].

[157] The explanatory memorandum to ECRI's General Policy Recommendation No. 14, [1] defines intersectional discrimination as "a situation where several grounds interact with each other at the same time in such a way that they become inseparable, and their combination creates a new ground".
See also Gerards J and Xenidis R, Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law (European network of legal experts in gender equality and non-discrimination / European Commission, 2021).

[158] Data categorisation might also be problematic and lack representativeness, with consequences on attempts to fix algorithmic discrimination. See Ruberg, B. and Ruelos, S.,'Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics' (2020) *Big Data & Society*.

[159] Balayn A and Gürses S, Beyond Debiasing: Regulating AI and its inequalities (European Digital Rights 2021), 62-63.

[160] B.S. v. Spain Application no. 47159/08 (European Court of Human Rights, 24 July 2012).

[161] See e.g., S.A.S. v. France Application no. 43835/11 (European Court of Human Rights, 1 July 2014) or Garib v. The Netherlands Application no. 43494/09 (European Court of Human Rights, Grand Chamber, 6 November 2017).

robust legal framework against intersectional discrimination, often due to formalistic comparison-based conceptions of equality, will prove particularly problematic in the context of algorithmic discrimination.

II. Privacy and data protection law: Fairness and accuracy

In addition to legal instruments pertaining to equality and discrimination, **privacy and data protection law can also be leveraged to tackle algorithmic discrimination**. The concept of fairness in privacy law relates to an organisation's intent to use personal information in good faith, with the intention of balancing the interests of data controllers and data subjects (the individuals). There is general agreement for example that the processing of personal information which is beyond an individual's knowledge/expectation would lead to an unfair situation in the eyes of privacy regulators. **However, the idea of fairness can have many possible nuances: non-discrimination, fair balancing, procedural fairness, bona fide, etc**.

The relation between discrimination and (un)fairness can be found in many legislative acts, proposals, and policy documents across the globe. Convention 108+, alongside the GDPR and many other privacy laws around the world, states that, in order to ensure fair and transparent processing in respect of the data subject, the controller should use appropriate mathematical or statistical procedures for profiling, implement technical and organizational measures appropriate to prevent potential risks for the interests and rights of the data subject. Risks may include discrimination on the grounds of racial or ethnic origin, political opinion, trade union membership, genetic status or sexual orientation.

Fairness is an overarching principle which requires that personal data shall not be processed in a way that is detrimental, discriminatory, unexpected or misleading to the data subject. It can be argued that fairness in privacy law relates to the **need to address the power imbalance between data subjects (individuals) and the digital ecosystem** and, for this reason, in recent times, privacy law has been leveraged quite extensively to deal with the harms of AI and algorithmic decision making, as outlined in a report issued by the Future Privacy Forum.[162] The report highlights actions taken by Data Protection Authorities including detailed transparency obligations about the parameters that led to an individual automated decision, a broad reading of the fairness principle to avoid situations of discrimination, and strict conditions for valid consent in cases of profiling and automated decision making.

For the purpose of this study, we are looking into two elements of fairness from a privacy standpoint:

- ***Fairness as procedures:*** transparency and fairness are inextricably linked because it is arguable that opening the source code to external scrutiny or providing a meaningful explanation on the processing of personal information by the AI system could lead to identification of bias and its root causes, and thus a positive increase in public accountability. For example, The Italian *Corte di Cassazione* issued a sentence in 2021 stating that a data subject's consent cannot be deemed valid if the algorithm is not transparent as the data subject would not be able to understand what they are

[162] AUTOMATED DECISION-MAKING UNDER THE GDPR – A COMPREHENSIVE CASE-LAW ANALYSIS, Future Privacy Forum, available at: https://fpf.org/blog/fpf-report-automated-decision-making-under-the-gdpr-a-comprehensive-case-law-analysis/

consenting to.[163] This case was welcomed by the Italian privacy regulator, Garante, as a demonstration of how the privacy law (and the GDPR in this case) is fit for upholding individuals' rights in the age of AI.

- ***Fairness as the protection of individual vulnerabilities***: in privacy law, fairness is often conceived as a corrective tool for rebalancing asymmetric or unbalanced relationships between organisations and individuals. Take for example the case of algorithmic platforms where the French Conseil d'Etat (as rephrased by CNIL) affirms that "fairness consists of ensuring, in good faith, the search engine optimisation (SEO) or ranking service, without seeking to alter or manipulate it for purposes that are not in the users' interest".[164] On a more general level, in the algorithmic environment, "fairness could well represent a solution to the problem of *unbalanced relations* between controllers of algorithms and users".[165]

For many countries both within and outside of Europe, the modernization of Convention 108 – with the introduction of **new rights for data subjects in algorithmic decision-making contexts,** particularly in connection with artificial intelligence – represents a common ground, as the treaty serves as a borderline standard for how countries should go about protecting the privacy rights of their citizens in the age of AI. The GDPR, which has many similarities with **Convention 108 + (**although the Council of Europe has a much wider reach and territoriality than the EU) also contains provisions to support individual rights in the context of AI and algorithms, including the renowned Article 22, which safeguards individuals from automated decision making.

There are several other safeguards that apply to such data processing activities, notably the ones stemming from the general data processing principles in Article 5, the legal grounds for processing in Article 6, the rules on processing special categories of data (such as biometric data) under Article 9, specific transparency and access requirements regarding algorithmic decision-making (ADM) under Articles 13 to 15, and the duty to carry out data protection impact assessments in certain cases under Article 35.

However, there are **limitations** in current privacy instruments when it comes to AI and algorithmic decision making, including:

- Exercising data subjects' rights in the context of AI and algorithmic decision making is rather complex. For example, even with the guidance of Data Protection Working Party 29 on automated individual decision-making and profiling, the assertion of GDPR Article 22 ("**solely" automated**, and "**legal or similarly significant effects**") presents practical challenges.

  Transparency of algorithmic management is the first step towards genuine accountability. However, **transparency and explicability** requirements in relation to bias mitigation raise questions around the **intersection of privacy and trade secret laws**. For an algorithm to be explainable it needs to have a degree of accessibility,

[163] *Corte di Cassazione, Civile Ord. Sez. 1 Num. 14381*, ItalgiureWeb, 25 May 2021 available at: http://www.italgiure.giustizia.it/xway/application/nif/clean/hc.dll?verbo=attach&db=snciv&id=./20210525/snciv@s10@a2021@n14381@tO.clean.pdf (last accessed: 26 May 2021)

[164] Conseil d'État, "Le Numérique et les droits fondamentaux", 2014, pp. 273 and 278-281.

[165] Understanding algorithmic decision-making: Opportunities and challenges, available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf

whether by internal or external auditors, a regulator, or a tribunal. However, a company's own algorithm may also be covered by trade secrets legislation. There are interesting developments in this sense thanks to the emergence of Secure Multi Party Computation that may enable an AI to be interrogated without having access to the actual code. Nevertheless, that is still a long way off.

III. AI sectoral regulations: strengths and limits for promoting equality and addressing discrimination

In addition to discrimination, privacy and data protection laws, sectoral regulations will also be relevant for addressing algorithmic discrimination.

The Council of Europe is currently developing regulation that would address algorithmic discrimination as part of an effort to promote human rights, democracy and the rule of law. This would take the form of a legally binding transversal instrument addressing issues in the public sector as well as binding and non-binding sectoral regulations.[166] In 2020 CAHAI prepared a "**Feasibility Study on a legal framework on AI design, development and application based on Council of Europe standards**", which recognises that "AI systems [can] be used in a way that perpetuates or amplifies unjust bias, also based on new discrimination grounds in case of so called 'proxy discrimination'".[167] At the same time, CAHAI considers that "AI systems can foster and strengthen human rights more generally, and contribute to the effective application and enforcement of human rights standards", for instance "by detecting biased (human or automated) decisions, monitoring representation patterns of different people or groups (for example women in the media) or analysing discriminatory structures in organisations".[168]

In its 2021 document "**Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law**", CAHAI recommends including "a provision on respect of *equal treatment and non-discrimination* of individuals in relation to the development, design, and application of AI systems to avoid unjustified bias being built into AI systems and the use of AI systems leading to discriminatory effects" in the legally binding transversal Framework Convention on AI regulation which is currently under preparation.[169]

CAHAI also proposes complementary regulation for the public sector, where it recommends that "documentation and logging processes" pertaining to the development of the system "should be meticulously kept to ensure transparency and traceability". It recommends that "[a]dequate test and validation processes, as well as data governance mechanisms should be put in place" to assess risks "of unequal access or treatment, various forms of bias and discrimination, as well as the impact on gender equality".[170]

---

[166] See CAHAI, "Feasibility Study on legal framework on AI design, development and application based on CoE standards" (2020), [54].
[167] Committee on Artificial Intelligence, "Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law", *Council of Europe* (2022), [13]
[168] Ibid, [20].
[169] Ibid, [27]
[170] Ibid, [60]

As other sectoral regulations are envisaged in Europe, it is important to flesh out the **added value of regulating AI at Council of Europe level.** Arguably, regulation by the Council of Europe can have **strong influence worldwide** due to the broad membership of the Council of Europe, its distinctive human rights-based approach and the fact that the instrument would be open for ratification to non-state parties as well. The CAHAI's "Possible Elements" document point towards minimum standards and an approach focused on the public sector, in line with the European Convention on Human Rights mechanism, which differs from the "market approach" taken by the EU in its draft EU AI Act.[171] A commonality between the two regulations would be the risk-based approach they both adopt to AI systems.[172] Yet the Council of Europe has the potential to foster a distinct **human-rights-based approach to AI and algorithmic technologies.**

Sectoral regulation of AI is also currently underway in the EU. The draft **EU AI Act** follows a risk-based approach and classifies AI systems as "**high-risk**" if they are deployed in the following areas: biometric identification and categorisation of natural persons, management and operation of critical infrastructure (road traffic, water, has, heating and electricity supply), education and vocational training, employment, workers management and access to self-employment, access to and enjoyment of essential private services and public services and benefits, law enforcement, migration, asylum and border control management, administration of justice and democratic processes. AI systems that present an "**unacceptable risk**", are prohibited for example "practices that have a significant potential to manipulate persons through subliminal techniques beyond their consciousness or exploit vulnerabilities of specific vulnerable groups such as children or persons with disabilities in order to materially distort their behaviour in a manner that is likely to cause them or another person psychological or physical harm". AI systems that present a **limited risk** are subjected to specific transparency obligations and those with **low or minimal risk** to codes of conduct.

Although the EU AI Act foresees promising transparency obligations with a view to bias mitigation, in particular in relation to training data and decision criteria,[173] several **criticisms** have been put forward regarding the way in which the EU AI Act proposes to ensure that fundamental rights are respected. For example, it approaches AI systems from a product liability perspective and thus **does not foresee complaint mechanisms** that would enable **victims of algorithmic discrimination or NGOs with a legitimate interest** to **request that changes are made to these systems after their deployment** in compliance with anti-discrimination law.[174] Moreover, commentators have criticised the **absence of legal obligations** for providers and users of AI systems **to conduct *ex ante* human rights impact assessments**.[175] The absence of any equality mainstreaming clause or positive obligation

[171] See Marten Breuer, "The Council of Europe as an AI Standard Setter" *Verfassungsblog* (4 April 2022) available at: https://verfassungsblog.de/the-council-of-europe-as-an-ai-standard-setter/.
[172] See Committee on Artificial Intelligence, "Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law", *Council of Europe* (2022), [19].
[173] See in particular Art. 10 on Data and data governance of the EU AI Act.
[174] See Joan Lopez Solano, Aaron Martin, Siddharth de Souza and Linnet Taylor, "Governing data and artificial intelligence for all Models for sustainable and just data governance" (Panel for the Future of Science and Technology, European Parliamentary Research Service 2022), 52 available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729533/EPRS_STU(2022)729533_EN.pdf.
[175] See ibid.

requiring AI and algorithmic systems to promote equality is also regrettable. **These are aspects on which the Council of Europe instrument should focus in order to create complementarity with the EU AI sectoral regulations and to ensure that its human rights mandate is at the core of the new legal provisions.**

As explained in Section 3 below, future AI sectoral regulations at Council of Europe level should also include a **legal obligation for AI and algorithmic systems to promote equality**. Norwegian equality legislation could offer a useful yardstick in this context as it foresees equality promotion as a legal obligation.[176]

Section 3

Promoting equality in and through the use of AI: the role of positive action and positive obligations

While the previous section highlighted relevant legal and policy instruments and Council of Europe level and at EU and international level, it has also pointed at problematic gaps, shortcomings and uncertainties in the applicability of these instruments to the problem of algorithmic discrimination. As shown in this section, avenues for fixing these issues should promote a **paradigm shift**. First, it could be recommended that **existing rules should be revisited in light of the new power and information asymmetries inherent in algorithmic technologies**. Second, we recommend that **positive action and positive obligations be used as an avenue for crafting a legal obligation to prevent discrimination and promote equality in and through the use of algorithmic systems**. Taking these two steps would **elevate 'equality by design' as a prominent feature of the Council of Europe's human-rights-based approach** to algorithmic discrimination,

I. Revisiting existing rules in light of new power asymmetries

This section aims to outline avenues for responding to the issues highlighted in Section 2 in relation to the applicability of existing legal provisions.

First, in light of existing research has shown that in the absence of safeguards, algorithmic bias systematically pervades algorithmic decisions, a **presumption of algorithmic bias** could be posited where no preventive measures have been taken by users of algorithmic systems. This is justified by the pervasiveness of bias in the design process of AI systems, ranging from biases in data collection and datasets to biases in problem design, algorithmic models and implementation of AI recommendations.[177] As argued by Eubanks, "when automated decision-making tools are not built to explicitly dismantle structural inequalities, their increased speed and vast scale intensify them dramatically".[178] In other terms, algorithmic discrimination is very likely to arise where no safeguards have been put in place. When it perpetuates inequality, the use of biased AI systems should be equated with actively enacting structural disadvantage and amplifying the unfair distribution of valuable social goods. The

[176] See Chapter 4 of the Norwegian Act relating to equality and a prohibition against discrimination (Equality and Anti-Discrimination Act), available at: https://lovdata.no/dokument/NLE/lov/2017-06-16-51#KAPITTEL_4.

[177] Grozdanovski suggests that it is possible to read the existence of such a presumption in the EU White paper on Artificial Intelligence, see Grozdanovski L, 'In search of effectiveness and fairness in proving algorithmic discrimination in EU law' (2021) 58 Common Market Law Review.

[178] Eubanks V, Automating inequality: how high-tech tools profile, police, and punish the poor (First edition. edn, St. Martin's Press 2018).

foreseeability of discriminatory harms arising from algorithmic bias thus justifies conceptualizing algorithmic discrimination as a form of **negligence**. Drawing from Moreau's work on discrimination and tort-based theories of discrimination law,[179] it is possible to derive a **social responsibility for users of algorithmic systems to take reasonable action to prevent the aggravation of discrimination** in society. This approach resonates with the discussions currently taking place in the EU context and in particular the Commission's proposal for a "rebuttable presumption for AI-related damages".[180]

Second, the pervasive use of AI systems establishes **new power and information asymmetries**. It becomes very **difficult for subjects of algorithmic decisions to identify discrimination** due to a combination of personalization, automation and opacity of decision-making processes. Comparison with similarly placed individuals and social interactions are important heuristic devices when it comes to acquiring presumptions of discrimination. Yet, reading social cues or comparing oneself to other loan applicants in the context of an online credit service becomes impossible. This information asymmetry makes it difficult to suspect discrimination in the first place. Even when suspicion arises, **collecting evidence is a further challenge** because decisions or the algorithmic recommendations supporting them are not readily available to consult and often not disclosed by users of ADM systems. Hence, **presenting proof to establish a presumption of discrimination in courts is a key legal challenge**. Even though the shift of the burden of proof can help mitigate the power asymmetries created by opaque algorithmic systems,[181] the threshold to trigger this shift should reflect end users' position and limited access to *prima facie* evidence.

Bringing together the foreseeability of algorithmic bias and existing information asymmetries reveals how the pervasive deployment of AI systems in decision-making processes **upsets the balance between the position of possible victims of discrimination and that of the providers and users of these systems**. While victims are subjected to more pervasive discrimination which they are contemporaneously less able to identify and prove, profit-makers enjoy increased power thanks to AI systems that enhance economic profits while possibly shielding them from liability for their discriminatory consequences due to the legal obstacles listed above. Hence, **the legal framework needs to be adjusted to reflect and integrate the power shifts and imbalances** that derive from the use of AI systems in a vast array of decisions that open or close life opportunities and therefore intensely affect inequality in society.

Revisiting existing rules on the burden of proof can help restore the effectiveness of non-discrimination law in light of new power and information asymmetries between users and subjects of algorithmic decision-making systems. Positing a presumption of algorithmic bias as suggested above would allow **shifting the burden of proof onto the defendant as soon as no preventive measures have been taken**. Such preventive measures could take the form, for instance, of an impact assessment, an audit or a certification of the algorithmic system used,

---

[179] See Sophia Moreau, 'Discrimination as negligence' (2010) 40 Canadian Journal of Philosophy 123; Oppenheimer DB, 'Negligent Discrimination' (1993) 141 University of Pennsylvania law review 899.

[180] See in this sense Luca Bertuzzi, "LEAK: Commission to propose rebuttable presumption for AI-related damages" (Euractiv, 2022) available at: https://www.euractiv.com/section/digital/news/leak-commission-to-propose-rebuttable-presumption-for-ai-related-damages/.

[181] See C-109/88 Handels- og Kontorfunktionærernes Forbund I Danmark v Dansk Arbejdsgiverforening, acting on behalf of Danfoss EU:C:1989:383.

as exposed in the recommendations section. Failure to take adequate preventive measures could then amount to negligence. This mechanism would support potential victims in adducing accessible *prima facie* evidence with a view to shifting the burden of proof onto users. Such an adaptation of the legal framework would also **mainstream positive action** and **preventive obligations** against algorithmic bias, as further outlined below.

Third, the adaptation of existing rules suggested above should be combined with a public supervisory approach.[182] **Empowering equality bodies, discrimination ombudspersons and national human rights institutions to monitor the discriminatory impact of algorithmic decision-making and support systems** should be made a priority. This involves providing these institutions with necessary legal rights and investigative powers (e.g., to access datasets and decision criteria), the right resources, but also with capacity to prevent discrimination by cooperating with users of ADM systems – for instance companies using ADMS to support recruitment procedures – to collect relevant data on the impact of their decisions, and to assist potential victims in relation to obtaining redress. Monitoring could take the form of **situation testing** where these authorities test the outcomes of a given system by comparing results for different groups. For instance, they could submit test CVs or credit applications from majority and minority groups to try and reveal algorithmic discrimination in contexts where companies use ADM systems. They could also conduct **audits** to detect potential bias if granted access to relevant systems. Such **public enforcement** methods could support victims by mitigating existing obstacles to establishing *prima facie* discrimination.

The monitoring function of equality bodies should be supported by **legal obligations around transparency.** Users of algorithmic systems should be required to **provide meaningful and intelligible information on the criteria used for decision making**. At the moment, the GDPR does not offer a right to explanation.[183] In the area of goods and services, consumer protection should also be explored as a tool to request information about algorithmic decisions for consumers who have been potentially discriminated against. This could help address the power asymmetries created by the opacity of ADMS between the subjects of algorithmic decisions and their authors.

Fourth, it is necessary to ensure the reviewability of algorithmic systems in light of non-discrimination obligations. Where applicants, lawyers or judges are presented with technical information concerning a specific system, such information is not likely to be intelligible in terms of the system's discriminatory or non-discriminatory nature. **Technical discussions about the adequacy of given fairness metrics and appropriate thresholds for trade-offs between accuracy and equity are difficult to assess from the perspective of legal obligations arising from anti-discrimination law**. In this context, how to ensure that ADMS undergo a **proportionality test** that guarantees the effectiveness of non-discrimination law? Here again, several solutions can be envisaged, as further articulated in the recommendations section of this Study. On the one hand, **transparency obligations** weighing on the users of ADMS could guarantee access to an intelligible account of the technical and fairness choices made by

[182] See Xenidis R and Senden L, 'EU Non-discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination' in Bernitz U and others (eds), *General Principles of EU Law and the EU Digital Order* (Wolters Kluwer 2019).

[183] See Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a right to explanation of automated decision-making does not exist in the general data protection regulation." *International Data Privacy Law* 7.2 (2017): 76-99.

developers and users. On the other, **mainstreaming positive action** could lead to a **positive obligation to prevent algorithmic bias** that would **displace the proportionality assessment from the technical to the legal terrain**. What judges would consider, then, would rather be the appropriateness of the preventive measures taken to ward off bias, rather than technical fairness and equity choices.

Finally, we suggest that **liability, as a judicial construct approximating responsibility, should be allocated strategically so as to facilitate access to justice and remedies** in cases of algorithmic discrimination. In the context of the ECHR and other legal instruments at Council of Europe level, where obligations weigh on public authorities, we suggest that **state parties should hold users of AI systems liable for algorithmic discrimination arising from the deployment of their system**. As explained in the Section on recommendations, **this can be complemented with legal obligations for providers to conduct human rights impact assessments *ex ante* to prevent discriminatory harms.** This will also allow encourage the **documenting of any preventive measures** taken by the provider so as to **ensure that meaningful information can be provided to the user and end-users** of the system in case of legal proceedings.

The approach proposed here, which revolves around a presumption of algorithmic bias, negligence and prevention, could contribute to legal certainty and the effectiveness of the ECHR anti-discrimination provisions by alleviating victims' burden of proof, fostering preventive safeguards, clarifying the allocation of liability and helping better define available justifications for defendants. All in all, we suggest that **a more substantive approach to equality should drive the interpretation of anti-discrimination provisions** to safeguard their effectiveness in the context algorithmic discrimination.

II. An obligation to promote equality in and through the use of algorithmic systems: the role of positive action and positive obligations

This report has shown how AI systems, without the right guardrails and controls, can lead to further exclusion of vulnerable groups. Notwithstanding the discriminatory potential of AI, researchers and developers have explored the opportunities offered by AI for identifying and redressing inequality. This requires a **paradigm shift where baselines for software design and deployment are systematically called into question and checked in relation to their inclusionary or exclusionary impact.** In other terms, the deployment of a new AI system should be "purposeful and intentional in its inclusivity" and "must empower communities and present a benefit to all of society".[184] This requires a set of obligations on companies to having to do so, and a set of pre-market and post-release controls. Below, we argue that such a paradigm shift requires the vast array of available positive action measures including awareness-raising, promotion-based measures, temporary special measures and quota to be utilised for equality, diversity and inclusion purposes across the board.

[184] Renee Cummings, "This is how AI can support diversity, equity and inclusion", World Economic Forum, available at: https://www.weforum.org/agenda/2022/03/ai-support-diversity-equity-inclusion/. See also Equality Now, A Call For An Intersectional Feminist Informed Universal Declaration On Digital Rights, available at: https://www.equalitynow.org/news_and_insights/universal-declaration-on-digital-rights/.

Rooting out bias and inequality requires a conscious, arguably political and social choice. In the first instance, it should be recognized that AI systems are not neutral but reproduce and amplify structural inequality and the systems of exclusion and disadvantage that are institutionalized in society. This necessitates steeping away from a perpetrator's perspective on discrimination and instead acknowledging that majority norms and unquestioned assumptions underlying software development and deployment lead to the needs of minority groups not being accommodated.[185] Assuming that a system will equally cater for various groups will *de facto* prevent minority groups from benefitting from AI applications and related opportunities to the same extent as other groups. Therefore, **substantive equality and anti-discrimination 'by design' should be placed at the centre of the legal regulation of AI development and deployment.**

1) What is positive action?

Positive action, also called temporary special measures or positive measures in the European context, is a range of policies that can be adopted with a view to reaching full or *de facto* equality. It builds up on a critique of formal equality or equality of opportunity that denounces these frameworks' blindness towards the different starting positions of different social groups. For example, giving the same job opportunity to a worker with disability and an able-bodied worker might lead to a higher dropout rate in the first case because no accommodation measure has been taken to ensure that the worker living with a disability is actually able to perform their tasks. Instead, anchoring policies in theories of substantive equality dictates the adoption of special accommodation measures that create conditions where historically disadvantaged groups can participate in society and reap the benefits of that participation to the same extent as privileged groups. Concretely, that would mean ensuring that a worker living with a disability can access a safe and adapted physical and psychological working environment, for instance through special equipment, flexible working hours, etc. So-called transformative equality theories point in the same direction but place more conceptual emphasis on transforming the unequal status quo in the long-term, for example through granting specific and temporal advantages to structurally disadvantaged groups. An example of such equality policies is flexible quota schemes whereby, for example, an employer faced with equally qualified male and female candidates in a recruitment process would give preference to the female candidate where women are under-represented in the professional community at stake.

In the context of the Council of Europe, positive action is not a legal obligation but has for example been encouraged by the European Commission against Racism and Intolerance (ECRI) "as an effective tool for achieving a fair and even playing field in society for members of disadvantaged groups".[186] In the EU, non-discrimination law allows for special measures in the framework of positive action within certain limits such as the prohibition on strict quota that would give an automatic preference to under-represented groups and the need for special

[185] For a powerful account of the perpetrator's perspective on discrimination vs understanding discrimination as a structural phenomenon, see e.g., Freeman AD, 'Legitimizing Racial Discrimination Through Antidiscrimination Law: A Critical Review of Supreme Court Doctrine' (1978) 62 Minnesota Law Review. This has been recognized in law though the concept of indirect discrimination.

[186] European Commission against Racism and Intolerance, Seminar with national specialised bodies to combat racism and racial discrimination on positive action: explanatory note (2007), available at: https://rm.coe.int/seminar-with-national-specialised-bodies-to-combat-racism-and-racial-d/16808b54b0.

measures to aim to transform the status quo in the long run.[187] The definition of positive action in the context of the Council of Europe and the European Convention on Human Rights is similar. The concept of "temporary special measures" is often used. ECRI's General Policy Recommendation no. 7 for example indicates that "[t]he law should provide that the prohibition of racial discrimination does not prevent the maintenance or adoption of temporary special measures designed either to prevent or compensate for disadvantages suffered by persons [from protected groups] or to facilitate their full participation in all fields of life".[188] It also states that "[t]hese measures should not be continued once the intended objectives have been achieved".[189]

2) Positive obligations under the ECHR

To approach the question of how to promote equality in and through the use of AI, the legal basis exposed above, which authorizes positive action, can be considered together with another important specific feature of the ECHR, namely the notion of **positive obligations**. Positive obligations entail that states have, in certain circumstances, the duty to actively take measures to achieve equality and prevent discrimination.[190] This goes further than limited passive or negative duties not to discriminate because it implies taking preventive action against discrimination or positive action measures to promote equality as a means to comply with Article 14 ECHR.

In its General Policy Recommendations No. 7, ECRI specifically endorses positive obligations to promote equality and prevent discrimination in the form of constitutional provisions, duties for public authorities, as well as obligations for public bodies to condition "the awarding of contracts, loans, grants or other benefits" to the respect of the positive obligation to promote equality and prevent discrimination.[191] This can be used as a legal basis to create an equality mainstreaming obligation in the context of AI use by public authorities.

Positive obligations and positive action provide an interesting legal basis for utilising AI to promote equality in two regards. On the one hand, it can be argued that positive obligations to prevent discrimination require states to use positive action in order to create safeguards to prevent unlawful algorithmic bias from emerging at any level of the AI lifecycle. On the other hand, positive obligations to promote equality could be interpreted as a requirement for states to invest in using the new opportunities created by AI to better serve disadvantaged communities so that they can fully enjoy the rights guaranteed by the ECHR. The next paragraphs lay out strategies for doing so.

3) Centring positive action

---

[187] For a detailed account, see Raphaële Xenidis and Hélène Masse-Dessen, 'Positive action in practice: some dos and don'ts in the field of EU gender equality law' (2018) 2 European equality law review 36.

[188] ECRI General Policy Recommendation No. 7 on National Legislation to Combat Racism and Racial Discrimination (2002), [5].

[189] Ibid.

[190] See European Court of Human Rights, Guide on Article 14 of the Convention (prohibition of discrimination) and on Article 1 of Protocol No. 12 (general prohibition of discrimination) (2022), [42-43] available at: https://www.echr.coe.int/Documents/Guide_Art_14_Art_1_Protocol_12_ENG.pdf. See also e.g., European Court of Human Rights, Application no. 34369/97 *Thlimmenos v. Greece* (2 April 2000) and European Court of Human Rights, Application no. 11146/11 *Horváth and Kiss v. Hungary* (29 January 2013).

[191] Ibid, [2], [8] and [9].

A *sine qua non* condition for using AI for good is positive action. Positive action can take many forms ranging from support measures such as information dissemination among targeted communities, dedicated training and funding programmes, to temporary special measures and flexible quota systems.[192] For example, key priorities should include **diversifying** educational and professional communities involved with all phases of the development and deployment of AI applications through financial support and awareness-raising efforts. This can be part of a broader effort to attract and retain more women and people from marginalized communities to STEM fields.

Where necessary, temporary special measures and flexible quota schemes should be used to ensure parity and inclusion in educational and professional communities. Positive action measures in the form of e.g., special accommodation and anti-stereotyping measures should aim to render these environments more inclusive so as to retain minority groups in the long-term and reduce drop-out rates.

Training should be provided to these communities via a transformation of educational curricula, with ethical issues, legal requirements and social science approaches to discrimination and inequality being part and parcel of higher and professional education. Complementary training should also be provided regularly to experts, stakeholders and professional communities in the AI industry on an *ad hoc* basis or as continuous education. Such training should address structural inequality, gender mainstreaming, and stereotyping.

An approach centred on substantive equality and positive action might also require adapting existing legal arrangements. Indeed, **as the emergence of new technologies shifts power dynamics between users and subjects of AI systems, the justice arrangements and normative dispositions underpinning legal rules become unsettled. Re-balancing such power asymmetries therefore entails adapting the legal architecture.** As explained below, rules around the shift of the burden pf proof might be eased for victims of algorithmic discrimination via the positing of a presumption of algorithmic bias.[193] Such a presumption could arise where users of an AI system have not put antidiscrimination safeguards in place, i.e., where they have assumed AI systems to be neutral towards protected groups. As described below, valid safeguards could take several forms such as audits, certifications, equality impact assessments. Further details on this proposed legal adaptation are provided in section 4.

4) Using data analytics to detect discrimination

A second possibility for AI to be used for promoting equality is through deploying the power of data analytics to detect discriminatory patterns in the allocation of resources, the dissemination of information, the representation of groups or the performance of given

[192] See Christopher McCrudden, Resurrecting positive action (2020) 18(2) *International Journal of Constitutional Law*, 429.

[193] Not to be confounded with a presumption of algorithmic discrimination because such bias might or might not be discriminatory. For other suggestions on easing the burden of proof in relation to algorithmic discrimination, see Janneke Gerards and Raphaële Xenidis, *Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law* (European network of legal experts in gender equality and non-discrimination / European Commission, 2021) and AlgorithmAudit, White Paper: Reversing the burden of proof in the context of (semi-)automated decision-making (2022) available at: https://drive.google.com/file/d/1RHdqoGVgwv-FTv8qC9fAlsVl8eUTcR7s/preview.

systems. Several examples show that data analytics can also be utilized to unpack bad models and end practices that replicate bias. For instance, AI image recognition technologies could be used to analyse large amounts of data and assess representations of women and minorities across different media sectors ranging from TV programmes to movies, online and physical advertising, etc. In content moderation, AI has been used to detect hate speech in order to report and remove offensive content.[194] At the same time, it is crucial to prevent that such deployment of AI silences minority groups.[195] Detecting discriminatory language in job ads automatically could also be a way to put AI to the service of the promotion of equality. Going even further, recommender systems could be used to recommend alternative inclusive language to substitute discriminatory content in job ads.

5) AI as a means to serve underserved communities and improve accessibility

Beyond detection, AI systems can also be purposively developed to serve marginalized, at-risk or underserved communities. For instance, AI can be used to improve accessibility to information or existing goods and services. Training automated translation systems on regional or minority languages that are spoken only by a small number of persons would improve access to key services. AI could also serve the promotion of equality in the criminal and policing sector, for instance when put to use to prevent risks of gender-based violence as in Spain with the VioGen software. In the health sector, AI could be used to enhance access to healthcare in disenfranchised areas and to improve diagnosing capacities for traditionally under-represented groups.

The condition for such positive usages of AI is however to invest resources into diversifying and training the professional communities involved in developing and using AI and to take positive action measures to ensure that these systems serve marginalised groups. At the same time, "technosolutionism" should be avoided and AI should not be perceived as a panacea to solve discrimination. It is crucial to remember that social issues require a social approach – not a purely technological one. While AI can certainly be developed and used for the promotion of equality, it is important to view it as a complementary tool in the framework of well-funded and carefully thought-through equality policies. This requires a conscious shift of approach.

[194] European Commission against Racism and Intolerance, General Policy Recommendation No. 15 On Combating Hate Speech CRI(2016)15, [140] available at: https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01.

[195] See for example the sexist and racist effects of automate content moderation: Gerrard Y and Thornham H, 'Content moderation: Social media's sexist assemblages' (2020) 22 1266.

Recommendations