Parliamentary Assembly
Assemblée parlementaire
http://assembly.coe.int

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Provisional version

Doc. …
… September 2020

Preventing discrimination caused by the use of artificial intelligence

Report[1]
Committee on Equality and Non-Discrimination
Rapporteur: Mr Christophe Lacroix, Belgium, Socialists, Democrats and Greens Group

Summary

Artificial intelligence (AI), by allowing massive upscaling of automated decision-making processes, creates opportunities for efficiency gains – but in parallel, it can perpetuate and exacerbate discrimination. Public and private sector uses of AI have already been shown to have a discriminatory impact, while information flows tend to highlight extremes and foster hate. The use of biased datasets, design that fails to integrate the need to protect human rights, the lack of transparency of algorithms and of accountability for their impact, as well as a lack of diversity in AI teams, all contribute to this phenomenon.

States must act now to prevent AI from having a discriminatory impact in our societies, and should work together to develop international standards in this field.

Parliaments must moreover play an active role in overseeing the use of AI-based technologies and ensuring it is subject to public scrutiny. Domestic antidiscrimination legislation should be reviewed and amended to ensure that victims of discrimination caused by the use of AI have access to an effective remedy, and national equality bodies should be effectively equipped to deal with the impact of AI-based technologies.

Respect for equality and non-discrimination must be integrated from the outset in the design of AI-based systems, and tested before their deployment. The public and private sectors should actively promote diversity and interdisciplinary approaches in technology studies and professions.

[1] Reference to Committee: Doc 14808, Ref. 4434 of 12 April 2019.

A. Draft resolution[2]

1. Artificial intelligence (AI) is transforming the way we live. It enables massive upscaling of processes and is already used by a wide range of private and public entities, across fields as diverse as selection procedures that determine access to employment and education, the evaluation of individual entitlements to welfare payments or to credit, or the targeting of advertising and news.

2. Many uses of AI can have a direct impact on equality of access to fundamental rights, including the right to private life and the protection of personal data; access to justice and the right to a fair trial, in particular as regards the presumption of innocence and the burden of proof; access to employment, education, housing and health; and access to public services and welfare. The use of AI has been found to cause or exacerbate discrimination in these fields, leading to denials of access to rights that disproportionately affect certain groups – often women, minorities, and those who are already the most vulnerable and marginalised. Its use in information flows has also been linked to the spread of online hate that spills over into all other social interactions.

3. Machine learning used to build AI-based systems relies on vast datasets (big data), much of which is personal data. Effective guarantees of the protection of personal data remain essential in this context. At the same time, data are by nature biased, as they reflect discrimination already present in society as well as the bias of those who collect and analyse them. Choices about which data to use and which to ignore in AI-based systems, as well as a lack of data on key issues, the use of proxies and the difficulties inherent in quantifying abstract concepts, can also lead to discriminatory results. Biased datasets are at the heart of many cases of discrimination caused by the use of AI, and remain a major issue to be resolved in this field.

4. The design and purpose of AI-based systems are also crucial. Algorithms optimised for efficiency, profitability or other objectives, without taking due account of the need to guarantee equality and non-discrimination, may cause direct or indirect discrimination, including discrimination by association, on a wide variety of grounds, including sex, gender, age, national or ethnic origin, colour, language, religious convictions, sexual orientation, gender identity, sex characteristics, social origin, civil status, disability, or health status. This makes it especially important, wherever their use may have an impact on access to fundamental rights, for AI-based systems to incorporate full respect for equality and non-discrimination in their design, from the outset, and to be rigorously tested before they are deployed, as well as regularly after their deployment, in order to ensure that these rights are guaranteed.

5. The complexity of AI systems, and the fact that they are frequently developed by private companies and treated as their intellectual property, can lead to serious issues of transparency and accountability regarding decisions made using these systems. This can make discrimination extremely difficult to prove and can hinder access to justice, in particular where the burden of proof is placed on the victim and/or where the machine is assumed by default to have made the correct decision, violating the presumption of innocence.

6. The lack of diversity in many tech companies and professions heightens the risk that AI systems will be developed without due regard to their potentially discriminatory impacts on some individuals and groups in society. Women and minorities' access to science, technology, engineering and mathematics (STEM) professions needs to be improved, and a true culture of respect for diversity urgently needs to be developed within these professional milieus. Using interdisciplinary and intercultural approaches throughout all stages of the design of AI systems would also contribute to strengthening them from the standpoint of equality and non-discrimination.

7. Finally, strong, clear and universally accepted and applicable ethical principles must underpin the development and deployment of all AI-based systems. The Assembly considers that these principles can be grouped under the following broad headings: transparency, including accessibility and explicability; justice and fairness, including non-discrimination; human responsibility for decisions, including liability and the availability of remedies, safety and security; and privacy and the protection of personal data.

8. The Assembly welcomes the fact that both public and private actors have begun to examine and develop ethical and human rights standards applicable to the use of AI. It welcomes in particular the Committee of Ministers' Recommendation Rec/CM(2020)1 on the human rights impact of algorithmic systems, along with its accompanying guidelines on addressing the human rights impacts of algorithmic systems, and the recommendation of the Council of Europe Commissioner for Human Rights on 'Unboxing Artificial Intelligence: 10 steps to protect Human Rights'. It endorses the general proposals made in these texts for application also in the area of equality and non-discrimination.

[2] Draft resolution adopted unanimously by the Committee on 11 September 2020.

9. The Assembly stresses that legislators must not hide behind the complexities of AI to prevent them from introducing regulations designed to protect and promote equality and non-discrimination in this field: the human rights issues at stake are clear and require action. In addition to ethical principles, procedures, tools and methods for regulating and auditing AI-based systems in order to ensure their compliance with international human rights standards, and in particular with the rights to equality and non-discrimination, are needed. Given the strong transnational and international dimensions of AI-based technologies, international standards also appear to be needed in this field.

10. In the light of these considerations, the Assembly calls on member States to:

10.1. review their antidiscrimination legislation, and amend it as necessary, so as to ensure that it covers all cases where direct or indirect discrimination, including discrimination by association, may be caused by the use of AI, and that complainants have full access to justice; in the latter respect, pay particular attention to guaranteeing the presumption of innocence and ensuring that victims of discrimination do not face a disproportionate burden of proof;

10.2. draw up clear national legislation, standards and procedures to ensure that AI-based systems comply with the rights to equality and non-discrimination wherever the enjoyment of these rights may be affected by the use of such systems; and

10.3. ensure that equality bodies are fully empowered to address issues of equality and non-discrimination that arise due to the use of AI, and to support individuals bringing cases in this field, and that they have all the necessary resources to carry out these tasks.

11. In order to ensure that the use of AI-based technologies by public authorities is subject to adequate parliamentary oversight and public scrutiny, the Assembly calls on national parliaments to:

11.1. make the use of such technologies a part of regular parliamentary debates, and ensure that an adequate structure for such debates exists;

11.2. require the government to notify the parliament before such technology is deployed;

11.3. require the use of such technologies by the authorities to be systematically recorded in a public register.

12. In order to address underlying issues of diversity and inclusion in the field of AI, the Assembly further calls on member States to:

12.1. promote the inclusion of women, girls and minorities in STEM education paths, from the earliest ages and to the highest levels, and work together with industry to ensure that diversity and inclusion are fostered throughout career paths;

12.2. support research into data bias and the means by which its impact can be effectively countered in AI-based systems;

12.3. promote digital literacy and access to digital tools by all members of society.

13. The Assembly invites all entities, both public and private, working on and with AI-based systems, to ensure that respect for equality and non-discrimination is integrated from the outset in the design of such systems, and adequately tested before their deployment, wherever these systems may have an impact on the exercise of or access to fundamental rights. To this end, it invites these entities to consider building capacity for a Human Rights Impact Assessment framework for the development and deployment of AI-systems by both private and public entities. In addition, it encourages the use of interdisciplinary and diverse teams at all stages in the development and deployment of AI-based systems.

14. Finally, the Assembly calls on national parliaments to support work being carried out at international level, in particular through the Council of Europe's Ad hoc Committee on artificial intelligence (CAHAI), to ensure that human rights standards are effectively applied in the field of AI, and that respect for the principles of equality and non-discrimination is guaranteed in this field.

B. Draft recommendation[3]

1. The Assembly refers to its Resolution … (2020) entitled “Preventing discrimination caused by the use of artificial intelligence”. It notes that this resolution was adopted as work was ongoing within the Council of Europe by the Ad hoc Committee on artificial intelligence (CAHAI).

2. The Assembly recalls that equality and non-discrimination are fundamental rights and that all Council of Europe member States are required to respect these rights, in accordance with the European Convention on Human Rights, as interpreted by the caselaw of the European Court of Human Rights, and with the European Social Charter, as interpreted by the European Committee on Social Rights.

3. The Assembly therefore calls on the Committee of Ministers to take into account the particularly serious potential impact of the use of artificial intelligence on the enjoyment of the rights to equality and non-discrimination when assessing the necessity and feasibility of an international legal framework for artificial intelligence.

[3] Draft recommendation adopted unanimously by the Committee on 11 September 2020.

C. Explanatory memorandum by Mr Lacroix, rapporteur

1. Introduction

1. Artificial intelligence is transforming the way we live. Automated decision-making processes are deployed in selection procedures for access to jobs or higher education; they are used to evaluate a person's creditworthiness, or to determine their entitlement to welfare benefits; they determine the information that is made available to internet users in their personalised newsfeeds or search engine results; they define who is targeted by political and other advertising, and by what messages.

2. It is frequently assumed that when decisions are made by machines, they will be objective and free of bias. Yet this ignores the role necessarily played by human beings in designing the algorithms at play, as well as the bias already existing in the data used to feed them. Today, there is ample evidence that the use of AI can not only reproduce known discriminatory outcomes, but also produce new ones.

3. The emergence of artificial intelligence (AI) that is not subject to regulation under a sovereign and independent democratic process thus risks leading to increasing human rights violations, and notably causing, perpetuating or even exacerbating discrimination and exclusion, whether or not this is its express aim.

4. The challenges are multiple and will affect individuals and our societies as a whole. Moreover, the technology and algorithms used know no borders. This means that national measures to prevent discrimination in this field – while essential – cannot provide a sufficient answer in themselves and makes international regulation especially important. The surest way to ensure that the human rights issues at stake are effectively addressed is to take a strong, common, multilateral approach.

5. My report seeks to define and propose a basic international framework for human-oriented AI based on ethical principles, respect for human rights and non-discrimination, equality and solidarity. The overall aim is to ensure that everyone's rights are guaranteed, in particular the rights of those people most exposed to the potentially discriminatory effects of the use of AI such as women, ethnic, linguistic and sexual minorities, workers, consumers, children, the elderly, people with disabilities or other people at risk of exclusion.

2. Scope of this report

6. The wide-scale deployment of AI affects more and more areas of citizens' daily lives and can have a considerable impact on their access to rights, including by causing discrimination. As politicians, we therefore have a particular responsibility to reflect on the possible regulation of such systems, in order, inter alia, to prevent such discrimination.

7. It is important to note from the outset that my report has been drafted in parallel to the preparation by a number of Assembly committees of several other reports dealing with AI. These reports concern: "Justice by algorithm – the role of artificial intelligence in policing and criminal justice systems"; "The brain-computer interface: new rights or new threats to fundamental freedoms?"; "Legal aspects of "autonomous" vehicles"; "Need for democratic governance of artificial intelligence"; "Artificial intelligence and labour markets: friend or foe?"; "Artificial intelligence in health care: medical, legal and ethical challenges ahead". Many questions related to equality and non-discrimination arise in the context of those reports. While I have briefly referred to some of these questions in my own report, for the sake of efficiency, I have deliberately focused my analysis on other issues linked to the prevention of discrimination caused by the use of AI.

8. For the purposes of this report, a description of the concept of AI is provided in the attached appendix. I would add here that in general usage, the term "artificial intelligence" is nebulous, and its scope imprecise. Its essence can however be understood by analogy with human intelligence. If the latter can be considered for each individual to be formed by the sum of their experience and of what they have learned, AI can be understood as a combination of "big data" (vast sets of pre-existing data, replacing individual experience) and the use of such data in machine learning.[4] The latter involves defining a mathematical model, based on algorithms and carried out using techniques such as artificial neuronal networks, that allows a machine to learn from a given dataset in order to be able to make (accurate) predictions when faced with unknown situations.[5]

[4] In other words, if "experience + learning = human intelligence", then "big data + machine learning = artificial intelligence". Evgeniou T., "AI Regulatory Challenges", presentation made at the 1st meeting of the OECD Parliamentary Group on Artificial Intelligence (AI), Paris, 26 February 2020.

[5] Frankle J., "Artificial intelligence for Policymakers", presentation made at the 1st meeting of the OECD Parliamentary Group on Artificial Intelligence (AI), Paris, 26 February 2020.

9. This helps to conceptualise the mechanisms at stake. However, to enable a coherent framework of regulation to be developed, it is important to determine the threshold beyond which a system ought to be regulated. On the one hand, defining AI in such a way as to cover all computer coding would mean that every word-processing programme or even every internet site could be considered as constituting AI, which would risk being too broad. (It would be difficult to devise a coherent regulatory system applicable to all forms of AI if the definition of it were too broad.) On the other hand, if the legal definition used is based only on techniques and applications that are already in use, it would run the risk of being unsuited to covering future developments, as legislative processes are often slow, while the AI sector is developing at an extraordinary speed.[6]

10. I would stress that, when it comes to preventing discrimination caused by AI, what matters is not so much what AI is or how it works; our work should focus on AI-based systems as a function of what they do. To put it another way, the policies and regulations that we establish with respect to AI, and in particular with respect to preventing discrimination caused by the use of AI, should cover automated decision-making processes, in particular where they are based on machine learning.[7]

11. The objective of the automated decision-making processes that concern us in this report is in general to make choices and/or to order things in a certain way. Which candidates will be invited to a job interview, and what level of pay will be offered to the persons selected? Who will be accepted to study in which university? What amount of welfare benefits are you entitled to? What news items will appear in your newsfeed, and in what order will they appear?

12. Of course, any selection process, whether automated or not, requires choices to be made. What is of interest in the present report is to identify measures that national authorities, and other relevant actors, should take in order to ensure that the results obtained when AI is used in automated decision-making processes are fair and, in particular, that they neither produce nor perpetuate discrimination in our societies.

3. The use of AI already produces discriminatory results

"[The attribution, through AI, of lower credit limits to women] matters for the woman struggling to start a business in a world that still seems to think women can't be as successful or creditworthy as men. It matters to the wife trying to get out of an abusive relationship. It matters to minorities harmed by institutional biases. It matters to so many. And so it matters to me."
Jamie Heinemeier Hansson[8]

13. As mentioned above, automated decision-making processes are already widely used in daily life, in fields as diverse as the administration of justice, higher education selection processes, recruitment, "optimisation" of staff working hours and evaluations of creditworthiness or of entitlement to welfare benefits.

14. As discussed below, there is broad evidence that such processes often produce unfair results, discriminating on grounds (for example) of gender, ethnic origins, social status or mental health. As legislators, it is our duty to address such human rights violations.

15. In the next two sections, in which I highlight some known cases of discrimination already caused by the use of AI, I have distinguished between private sector and public sector uses of AI. Different rights may indeed be at stake in these fields and when discrimination occurs, the avenues of redress potentially available to victims also vary. I then examine, in a third section, information flows managed through AI. These raise distinct issues and, in particular, can exacerbate discrimination by reinforcing stereotypes and prejudice.

3.1. The private sector

16. AI can provide powerful tools for streamlining services provided to customers and improving a company's business performance, as it can significantly speed up processes that would take humans longer to complete. However, the use of AI can also have highly negative effects for some groups of people.

17. One of the world's largest private employers, Amazon, invested years in developing an AI-based recruiting tool. It was trained to vet applicants using resumes submitted to the company over a 10-year period. In 2018, however, Amazon decided to abandon the tool, because it displayed gender bias.[9] Large companies

[6] By way of illustration, the number of scientific articles published in this field increases by 30% each year. Evgeniou T., "AI Regulatory Challenges", presentation made at the 1st meeting of the OECD Parliamentary Group on Artificial Intelligence (AI), Paris, 26 February 2020.
[7] Frankle J., "Artificial intelligence for Policymakers", presentation made at the 1st meeting of the OECD Parliamentary Group on Artificial Intelligence (AI), Paris, 26 February 2020.
[8] Heinemeier Hansson J., "About the Apple Card", blogpost, 11 November 2019.
[9] Dastin J., "Amazon scraps secret AI recruiting tool that showed bias against women", Reuters, 10 October 2018.

are also increasingly using automated personality tests during hiring processes, as a means of filtering out applicants. But these have been found to discriminate against candidates with mental health disorders, excluding them based on evaluations that are poor predictors of job performance.[10]

18. Targeted online advertising, based on machine learning, is another well-known source of discrimination in the field of employment. Independent research has for example found that significantly fewer women than men are shown online advertisements from companies that provide assistance in finding highly paid jobs.[11] This kind of discrimination is difficult to detect, except through extensive research, and almost impossible for individuals to contest as they cannot know what they are not seeing.

19. Beyond employment, New York's Department of Financial Services has also been invited to investigate allegations that several major tax-return companies used Google's advertising features to hide additional tax filing options from low-income individuals who would have been eligible to file their tax return for free.[12] Amnesty International has highlighted practices of Facebook that allowed advertisers in the field of housing either to target or to exclude certain groups of people based on their ethnicity or age.[13] The US Department of Housing and Urban Development moreover charged Facebook in March 2019 with encouraging, enabling and causing discrimination based on race, colour, religion, sex, familial status, national origin and disability, through its advertising platform. The platform allowed advertisers to hide their advertisements from users in certain neighbourhoods, or to choose not to advertise housing to users having certain interests, such as "hijab fashion" or "Hispanic culture".[14]

20. Cases such as this show how online behaviour, such as an internet user's choice of search topics, may be used to infer sensitive private information such as their ethnic origin, religious beliefs, sexual orientation or gender identity, or their affinity with certain topics, and to target advertising in ways that may be discriminatory. This also raises serious questions about the protection of privacy.[15]

21. To align itself with antidiscrimination law requirements, Facebook agreed to introduce changes to its algorithms to prevent such targeting in future.[16] However, machine-learning based algorithms have been shown to cause discrimination even when advertisers are not deliberately targeting advertising based on criteria that correspond to characteristics protected under anti-discrimination legislation. Thus, research has found that job postings for lower-paid work (janitors, taxi-drivers) are shown to a higher proportion of minorities, and jobs for preschool teachers and secretaries are shown to a higher proportion of women.[17]

22. Similar issues have been shown to arise when AI-based systems are used to assess creditworthiness. Shortly after the Apple Card was launched in 2019, for example, hundreds of customers began complaining of sexist results. Men were granted credit limits many times higher than those granted to women in an identical situation, who moreover had no means of contesting the decision. The only explanation Apple staff were able to give customers was, "It's just the algorithm."[18]

23. In my country, Belgium, some health insurance companies are currently seeking to use smartphone apps in order to gather information about the health status of the persons covered by their insurance policies. This raises serious privacy issues and could also be a source of discrimination based on health status.

24. These examples are more than mere anecdotes: they show clearly that the use of AI not only has the potential to produce discriminatory effects, whatever the ground of discrimination, but that in many cases, it is already doing so.

---

[10] O'Neil C., "Ineligible to Serve: Getting A Job", *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, Broadway Books, 2016.
[11] Spice B., "Questioning the fairness of targeting ads online", Carnegie Mellon University, 7 July 2015.
[12] NYDFS, "Governor Cuomo calls on DFS to investigate claims that advertisers use Facebook platform to engage in discrimination", Press Release, 1 July 2019.
[13] Amnesty International, "Surveillance Giants: How the business model of Google and Facebook threatens human rights", November 2019, p. 37-38.
[14] Jee C., "Facebook has been charged over housing ads that discriminate on race, colour, and religion", MIT Technology Review, 29 March 2019.
[15] Wachter S., "Affinity profiling and discrimination by association in online behavioural advertising", Berkeley Technology Law Journal, 2020, 35(2) (forthcoming).
[16] ACLU, "Facebook agrees to sweeping reforms to curb discriminatory ad targeting practices", 19 March 2019.
[17] Hao K., "Facebook's ad-serving algorithm discriminates by gender and race", MIT Technology Review, 5 April 2019.
[18] See inter alia Apple Newsroom, Apple Card launches today for all US customers, 20 August 2019; the series of tweets on this subject by David Heinemeier Hansson, https://twitter.com/dhh/status/1192540900393705474, and the many replies he received, including from a cofounder of Apple https://twitter.com/stevewoz/status/1193330241478901760; Natarajan S. and Nasiripour S., "Viral Tweet About Apple Card Leads to Goldman Sachs Probe", bloomberg.com, 9 November 2019.

25. Whether direct or indirect, such discrimination is a breach of fundamental rights. If it is allowed to occur or continue unchecked, the use of AI will in effect be perpetuating discrimination, and in some cases exacerbating it. Women, LGBTI people, people belonging to ethnic or religious minorities, people with disabilities and others will remain locked into lower-paying jobs on discriminatory grounds, with fewer opportunities to access credit or goods and services, and persons belonging to groups considered undesirable by housing providers will remain excluded from certain residential areas, not only discriminating against them individually but also aggravating spatial and social segregation in society.

26. Although the examples mentioned above are mostly American, it should be noted that the companies involved lay claim to millions, sometimes billions of customers across all continents, including Europe, and that their algorithms may produce discriminatory effects in every Council of Europe member State. Parliaments have a duty to address these issues, ensuring that antidiscrimination laws are robust enough to protect individuals and combat systematic discrimination, that companies using discriminatory AI can be held to account, and that effective remedies are in place.

3.2. The public sector[19]

27. The private sector is not alone in having recourse to artificial intelligence: it is often used by public authorities, notably in the context of the welfare state, where social services use digital technologies to assess individuals' eligibility for welfare, calculate the amount of their entitlement or investigate possible fraud or error. Numerous examples raise serious concerns about digital technologies using personal data in ways that breach privacy and/or wrongly deprive individuals of welfare benefits.

28. Thus, in the Netherlands, the SyRI system risk indicator, used to detect a risk of fraud, compiles and compares data from several government databases. In one of the pilot projects launched in this context, the data of 63 000 people who received low-income welfare benefits, and who were not suspected of any wrongdoing, were matched with data on water usage held by public companies supplying water, to identify automatically whether people were living alone or together. 42 cases of fraud were detected, meaning the success rate was a mere 0,07%. This raises serious questions regarding the respect of the right to privacy as well as the presumption of innocence. A Dutch court recently ruled that the legislation governing SyRI contained insufficient protection against interference in private life, as the measures taken to prevent and combat fraud in the interest of economic wellbeing had been disproportionate. Moreover, the system lacked transparency and its targeting of poor neighbourhoods could amount to discrimination on the grounds of socioeconomic or migrant status.[20]

29. In other European cases, the Polish Supreme Court has quashed a system set up in 2014 that sorted the unemployed into three categories based on data collected at the moment of registration for benefits and on a computer-based interview; a similar system was however introduced in Austria in 2018. In Sweden, an automated system for collecting activity reports from jobseekers was abandoned in 2018 because 10 to 15% of the automated decisions taken on the basis of the information collected were found to have been incorrect.

30. In the United Kingdom, the universal credit system, which combines six welfare benefits into one, is the first government service to have become digital by default, with an algorithm being used to calculate benefits each month based on information received in real time from employers, tax authorities and government departments. Many people have lost benefits because they simply lack the skills to fill in the new online forms. Moreover, benefits may be automatically reduced, without explanation or notification, on the basis of the results produced by the algorithm. The latter, and not the beneficiary of the welfare payments, is given the benefit of the doubt. Yet the authorities admit that each month, roughly 2% of the millions of transactions carried out (i.e. tens of thousands of cases) produce incorrect results. As the COVID-19 pandemic leaves increasing numbers of people jobless and reliant on welfare, there is a real risk that more individuals will be unjustly deprived of access to social welfare.[21]

31. A similar system put in place in Australia (commonly referred to as "Robodebt") has produced particularly harmful results. Automated data-matching (replacing human examination previously carried out manually by public servants) was introduced in 2016 to find discrepancies between the income data of welfare recipients held by the tax authorities and the social services, in order to detect possible overpayments of benefits or fraud. From

---

[19] The examples set out in this chapter were presented during the Committee's hearing on 4 December 2019 by Mr Christiaan van Veen, Director of the Digital Welfare State and Human Rights Project, Center for Human Rights and Global Justice, New York University School of Law; Special Advisor on new technologies and human rights to the UN Special Rapporteur on Extreme Poverty and Human Rights.

[20] Henley J. and Booth R., "Welfare surveillance system violates human rights, Dutch court rules", *The Guardian*, 5 February 2020.

[21] Lavelle D., "Coronavirus is shining a light on the wretched universal credit system", *The Guardian*, 3 April 2020.

this moment, anyone for whom a discrepancy was evaluated by the algorithm as suspect was required to provide evidence to the contrary via an online form, without which their allocations would be reduced or cut out altogether. The algorithm, however, took tax authority data (which are based on a full year) and compared it with fortnightly income, ignoring the fact that the income of welfare recipients is often very irregular, due for example to short-term contracts or seasonal work. As a result, thousands of people were wrongly deprived of welfare payments, and many of them were unable to challenge these decisions (automated notifications were sent to an old address; they did not have access to the portal via which they could have forwarded the required evidence). In many cases, people suddenly found themselves in serious debt, and some cases of suicide were reported. Some sources calculate that the authorities have attempted to claim back almost 600 million AUD (360 million EUR) from citizens based on this system, which often generates errors but under which the burden of proof is shifted to the individual and the results are very difficult to challenge.

32. These examples are just the tip of the iceberg; the number of issues will grow as governments seek to increase their use of technology in the name of greater efficiency. The recent scandal surrounding the automated adjustment of A-level results in the United Kingdom in the context of the COVID-19 pandemic, which particularly affected pupils from disadvantaged areas, and similar issues around the results of the international baccalaureat, are just two such examples.[22]

33. There are three key concerns as regards non-discrimination. First, there tends to be a lack of prior scrutiny, democratic oversight and public debate about these issues. Thus, there was very little parliamentary debate about the introduction of SyRI in the Netherlands in 2006, despite warnings from the data protection authority and other parties. Furthermore, freedom of information requests are often frustrated due to broad exceptions or the authorities' own lack of understanding of the technology used. The unequal impact of such systems on the poor and marginalised therefore often goes unnoticed. Second, AI tends to be perceived as necessarily fairer and more accurate than humans. However, while this may be true for very specific tasks, it is much less certain wherever a broader context needs to be taken into account. Where technology enables massive upscaling of processes but at the same time leads to wide-scale errors, those who are least able to be able to challenge the system (for example the poor or elderly, migrants, people with a disability) will again be disproportionately affected. This is especially serious when AI is used in the context of the welfare state. Finally, digital technologies are often deliberately targeted at poor and marginalised people, expanding the possibilities for constant State surveillance of these persons over time.

34. Faced with these problems, parliaments must realise that the issues at stake are not merely technical but highly political. AI-based systems are expensive to put in place, are used to implement particular political objectives that (in the examples given above) are caught up in politicised debates about the welfare State, and often come at a high human cost. Oversight and discussion of the use of these technologies therefore need to be made part of regular parliamentary debates. This could be done through setting rules, for example requiring governments to notify parliaments in advance of the use of such technologies, requiring their use to be systematically recorded in a public register, and ensuring that a structure for such discussions exists. Parliamentarians do not need to be experts in AI in order to understand the underlying political and societal issues. For example, in Australia and the UK, where automated calculations are used as a basis for the authorities demanding that citizens reimburse welfare payments, in many cases erroneously, the burden of proof has in effect been reversed and at the same time, the decisions are difficult or impossible to challenge. Due process is not followed and the right to a remedy is stymied. In the Netherlands, the presumption of innocence and the right to privacy of tens of thousands of people was violated yet only a few cases of welfare fraud were identified. The right to social assistance is also infringed when automated processes make it harder for people to claim benefits to which they are entitled. The least well-off people in society are hardest hit by this.

[22] See inter alia Hill A. and Davies C., "A-level results day 2020 live: 39.1% of pupils' grades in England downgraded – as it happened", *The Guardian*, 13 August 2020; Evgeniou T. et al., "What happens when AI is used to set grades?", *Harvard Business Review*, 13 August 2020.

3.3. The particular case of information flows

"The trouble with the internet…is that it rewards extremes. Say you're driving down the road and see a car crash. Of course you look. Everyone looks. The internet interprets behaviour like this to mean everyone is asking for car crashes, so it tries to supply them."
Evan Williams, Twitter founder[23]

35. I wish finally to draw attention to another field in which the use of AI may exacerbate discrimination, by fostering extremism and hate. It has been shown that the algorithms of certain websites, especially social media sites, automatically recommend increasingly radicalised viewpoints to their users.

36. The US presidential elections in 2016 have frequently been cited as an instance where the online dissemination of radical or extreme views (as well as fake news, an issue which however falls outside the scope of this report) may have had a decisive impact on the outcome of the elections. A former Google employee, Guillaume Chaslot, built an algorithm to explore whether bias existed in the videos automatically recommended by YouTube's algorithm during these elections to viewers having watched videos containing the words "Trump" or "Clinton". The results not only revealed a high number of recommendations of videos expressing extreme views, but also, that regardless of which candidate's name the viewer had initially searched for, YouTube's algorithm was ultimately far more likely to recommend videos that were more favourable to Trump than to Clinton (often anti-Clinton conspiracy videos).[24]

37. The tendency for the YouTube algorithm to recommend extreme videos is moreover not limited to the political domain: experiments in other spheres have led to similar results. For example, watching videos on vegetarianism leads to videos on veganism; videos about jogging lead to others about running ultramarathons; videos on the flu vaccine lead to anti-vaccination conspiracy videos.[25] Similar concerns have been raised about Facebook's newsfeeds.[26]

38. The algorithms used on social networks and newspapers' websites tend to be optimised for profitability. They therefore promote by default elements likely to attract high number of clicks or "engagements". Readers are thus encouraged to spend longer on a site (and therefore to be exposed to more advertisements) by stimulating their natural tendency to be attracted to titles that titillate.

39. I wish to underline that this spiral is not inevitable, just as the pattern of users of information sites and social media being caught up in "bubbles" of fellow users all sharing the same points of view is not inexorable. Much depends on the objectives that have been defined for the algorithm being used: in other words, what has it been designed to optimise? Objectives other than direct profitability can be fixed, such as presenting online readers with the widest possible range of points of view. This choice has been made by some Scandinavian newspapers in developing the algorithms used for their online editions.[27]

4. Data

"Garbage in, garbage out"[28]

40. Data play a crucial role in the field of artificial intelligence. On the one hand, vast datasets (so-called "big data") are needed to train and refine machine learning in order to develop complex AI systems. Such data generally concern people and are often generated by them (for example, when they choose to click on a link in a newsfeed or when they fill in an online form).

41. Despite the protection introduced, at least in EU countries, by the General Data Protection Regulation (GDPR) [29], users of information systems are however not always aware that such data are being collected, nor of how they may subsequently be used. This raises many issues as regards respect for private life – a common thread running through many uses of AI. For the purposes of the present report, I will simply underline that

[23] Lewis P., "'Fiction is outperforming reality': how YouTube's algorithm distorts truth", *The Guardian*, 2 February 2018.
[24] Cited in Streitfeld D., "'The Internet is broken': @ev is trying to salvage it", *New York Times*, 20 May 2017.
[25] Tufekci Z., "YouTube, the Great Radicalizer", *New York Times*, 10 March 2018.
[26] Evans J., "Facebook isn't free speech, it's algorithmic amplification optimised for outrage", *Techcrunch.com*, 20 October 2019.
[27] Bucher T., *If…Then: Algorithmic Power and Politics*, New York: Oxford University Press, 2018, p. 142.
[28] Aphorism commonly used by computer scientists, cited from at least as early as 1957: Mellin W.D., "Work with new electronic 'brains' opens field for army math experts", *The Hammond Times* 10 (1957), p.66.
[29] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

users of information systems do not all have the same degree of knowledge in this field. Thus, people are not all on an equal footing as regards online data collection and the protection of their personal data.

42. On the other hand, data are never unbiased: they are always a function of the time and place where they were collected.[30] Bias is inherent in existing human data, and both leads to and springs from stereotyping and prejudice. The biases prevalent at that time and in that place, as well as in the minds of those designing and conducting data collection exercises, are reflected in data collected.

43. The use of biased datasets, or datasets that reflect historical bias, prejudice or discrimination, is a major cause of discrimination in AI. Where, historically, fewer women and/or fewer people belonging to ethnic minorities have been employed in certain fields, or they have been employed on lower salaries, or credit has been refused to people belonging to certain groups, or minorities have tended to engage with advertisements for home rentals rather than home-buying, AI that bases its optimisation decisions on recognising and reproducing historical patterns will simply serve to entrench discrimination.[31] Correcting this flaw requires not only awareness of the historical patterns but deliberate design decisions, an area I explore further below.

44. In some cases, bias may be easy to fix – for example, when it comes to facial recognition software that performs less accurately based on skin colour, using a broader range of photographs to train a machine may rectify some problems. Enlarging the dataset in such a case may be relatively simple, as a vast range of photos of people of different ethnic origins is freely available on the internet. Nonetheless, how easy it is to resolve such an issue also depends on the use being made of the data. In the notorious Google Photos case, where the application as initially released automatically labelled photos of African Americans as "gorillas",[32] the racist labelling could be relatively easily fixed. Where AI-based facial recognition techniques are used not simply to categorise images as belonging to certain groups but in the criminal justice system, for example to identify specific individuals in situations where their individual liberty may be at stake, and where such systems perform significantly less accurately for people with dark skin, far more sophisticated solutions may be required in order to address the far-reaching human rights consequences of such flaws.

45. Some data have historically not been measured at all, leading to the invisibility of certain groups of people in available datasets. People who do not identify as either male or female (notably some intersex people) not only experience as hostile online forms where they are required to tick either a "male" or "female" box, but their specific situation cannot be measured and discrimination against them can be neither identified nor prevented. In the field of medical testing, women have also been historically excluded from medical trials, meaning both their health and their bodies' responses to medicine are less well understood than men's. Problems such as these predate the use of AI-based systems – but they mean that the use of such systems, trained using historical data, will tend to reproduce and entrench existing discrimination.[33]

46. Moreover, data provide a certain representation of reality, but often oversimplify it. As a result, they can only provide a more or less rough approximation of the reality that they are intended to represent.[34] Many States for example refuse to allow the collection of ethnic data (often reasoning that past misuse of such data shows that they should never be collected again). Instead, proxies such as the country of birth of individuals, their parents or their grandparents are used. These proxies capture many people belonging to ethnic minorities but miss many others, whose families have lived for generations in a country and who may still face discrimination based on skin colour or language.[35]

47. In other cases, information relevant to an algorithm may be extremely difficult to calculate. This is especially the case where it comes to abstract concepts (for example, fairness – a crucial question in the judicial field), as opposed to quantifiable elements such as the number of knife crimes recorded in a given area in a specified time-period. The question of justice by algorithm is the subject of ongoing work by another committee, and I shall therefore not examine it in detail in this report. However, I wish to highlight the serious

---

[30] Evgeniou T., "AI Regulatory Challenges", presentation made at the 1st meeting of the OECD Parliamentary Group on Artificial Intelligence (AI), Paris, 26 February 2020.
[31] Hao K., "Facebook's ad-serving algorithm discriminates by gender and race", MIT Technology Review, 5 April 2019.
[32] CNIL, Comment permettre à l'homme de garder la main? Les enjeux éthiques des algorithmes et de l'intelligence artificielle, December 2017, p. 31-32.
[33] Wachter-Boettcher S., *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*, New York, London, W. W. Norton, 2017; Jackson G., "Centuries of exclusion has meant women's diseases are often missed, misdiagnosed or remain a total mystery", *The Guardian*, 14 November 2019. The use of AI in the medical field is currently the subject of ongoing work in another report of the Assembly.
[34] O'Neil C., "*Civilian casualties: Justice in the Age of Big Data*", Weapons of Math Destruction, New York, Broadway Books, 2016, p. 95; Bucher T., *If…Then: Algorithmic Power and Politics*, New York: Oxford University Press, 2018, p. 8-12.
[35] Issues around ethnic data collection are regularly dealt with by ECRI; for a recent example see Cahuc P. and Valfort M.-A., "La pandémie de COVID-19 va renforcer les disparités raciales et ethniques", *Le Monde*, 19 August 2020.

discrimination that can arise when algorithms prioritise "efficiency" (relying on elements that are easy to count) over fairness (taking into account the broader implications of an algorithm's outcomes for society as a whole). The infamous COMPAS programme used in certain American states to analyse the risk of recidivism (in order to assist judges in deciding whether or not to impose a prison sentence) provides a regrettable case in point.[36]

48. Overall, then, the capacity of AI-based systems to prevent discrimination is highly dependent on the data used to train them. Historical bias, the absence of data on key issues, the use of proxies and the difficulties inherent in quantifying abstract concepts all need to be addressed and effectively resolved when AI-based systems are deployed, wherever their use may have an impact on human rights. As legislators, we must be aware of the crucial human rights issues at stake in this area and devise ways to ensure that citizens are protected from such discrimination.

5. Design and purpose

> *"King Midas said, 'I want everything I touch to turn to gold,' and he got exactly what he asked for. That was the purpose that he put into the machine, so to speak, and then his food and his drink and his relatives turned to gold, and he died in misery and starvation."*
> Stuart Russell, AI researcher[37]

49. Algorithms are designed to work single-mindedly towards a specific aim that has been identified by their programmers. A mathematical model is developed to allow a machine to learn from a given dataset and optimised in order to be able to make accurate predictions when faced with unknown situations. Choosing the appropriate objective for an algorithm is crucial, as all future design decisions, and the results ultimately produced through the use of an AI-based system, will depend on that choice. Ill-conceived objectives or policy choices will lead to undesirable, unfair results.

50. Automated recruitment processes used by large companies, for example, are often optimised for "efficiency", meaning that they serve to reduce as far as possible the numbers of candidates that have to be interviewed or to have their applications screened by human beings. If the algorithm is designed to select candidates that appear to fit existing company culture, however, it is unlikely to assist in increasing diversity within that company or changing its culture.[38] "Efficiency" concerns are also at the heart of many problematic public-sector uses of AI, in particular in the context of the welfare state, which I described earlier.

51. If the objective of a machine-learning model is misaligned with the need to avoid discrimination, then the results it produces will again perpetuate or exacerbate discrimination. Facebook's advertising tool, referred to earlier, offered advertisers a range of optimisation objectives from which to choose: the number of times an advertisement was viewed, the amount of engagement it generated, or the amount of sales to which it led. If showing a higher proportion of white users homes for purchase led to more engagement, then the algorithm would do so, thereby discriminating against black users. This happened because the algorithm was optimised for business goals, without taking into account the need to ensure respect for human rights, such as equal access to housing.[39]

52. In the case of information flows, too, the fact that algorithms are designed to optimise for engagement remains a major issue. "Free" platforms such as Facebook and YouTube, but also online media, financed through advertising, have an interest in increasing the time that users spend on their sites, and thus exposing them to more advertisements. If data show that users tend to be most drawn to extreme content, algorithms will promote that content. While such platforms tend to argue that they are "merely serving viewers what they want", the algorithms' objective is to maximise advertising revenues for the companies concerned.[40] In the process, the question of what is genuinely informative or newsworthy is ignored, while sexist, racist, antisemitic, Islamophobic, anti-Gypsyist, homophobic, transphobic and other hate-mongering speech is often promoted.

53. Even before the specific objective for which an algorithm is to be optimised is determined, fundamental questions also need to be asked about the purpose of an AI-based system. To take an example from the "real" world, stop-and-search programmes will amplify bias in the criminal justice system when they only target poor neighbourhoods looking for certain types of criminal behaviour, while leaving alone wealthier neighbourhoods (where white-collar crimes frequently occur, and domestic violence is as likely to happen as elsewhere). Data

---

[36] O'Neil C., "Civilian casualties: Justice in the Age of Big Data", *Weapons of Math Destruction*, New York, Broadway Books, 2016, p. 146-150 and 164-165.
[37] Russell S., "Three principles for creating safer AI", TED Talk, 2017.
[38] Lauret J., "Amazon's sexist AI recruiting tool: how did it go so wrong?", becominghuman.ai, 16 August 2019.
[39] Hao K., "Facebook's ad-serving algorithm discriminates by gender and race", MIT Technology Review, 5 April 2019.
[40] Tufekci Z., "YouTube, the Great Radicalizer", *New York Times*, 10 March 2018; Evans J., "Facebook isn't free speech, it's algorithmic amplification optimised for outrage", *Techcrunch.com*, 20 October 2019.

gathered from biased policing strategies and fed into AI-based systems will automatically show higher levels of criminal activity in the areas targeted, which then leads to an increase in policing in those areas, creating a deeply discriminatory "feedback loop" that is harmful to the inhabitants of poorer neighbourhoods (often ethnic minorities) while failing to capture activity in other areas that is equally worthy of police attention. Such "feedback loops" occur because algorithmic outcomes based on skewed datasets tend to confirm discriminatory practices – and because no evaluation of what is missing from those datasets is conducted.[41]

54. It should be underlined that – just like any selection procedure – no automated decision-making process will ever be entirely neutral, as it will always be the result of design choices, i.e. of specific conceptions of how things should be ordered. Because of their capacity to make large numbers of decisions at very high speed, however, the consequences of implementing discriminatory AI-based systems can be dramatic. As legislators, we need to find ways to ensure that design choices made in the development and implementation of automated decision-making processes systematically incorporate the need to protect human rights and prevent discrimination.

6. Diversity

"'Black girls need to learn how to code' is an excuse for not addressing the persistent marginalisation of Black women in Silicon Valley"
Safiya Umoja Noble[42]

55. Algorithms reflect the values, beliefs and convictions of those who design them. The capacity of automated decision-making processes to include and reflect the diversity present in our societies also depends on these factors.[43]

56. The lack of diversity that typically exists within tech companies entrusted with designing algorithms raises serious issues here. The underrepresentation of women in science, technology, engineering and mathematics (STEM) studies and professions has already been recognised by the Assembly, which has invited States to take measures to encourage women and girls to follow school and tertiary-level education in these fields.[44] However, the issues extend beyond gender inequalities. Researcher Kate Crawford has highlighted the social, racial and gender endogamy that characterises the environments in which those who train artificial intelligence are recruited today.[45]

57. Diversity in the workforce is not merely a "pipeline" problem, however. Discrimination in the workplace also leads to high turnover rates among the women and minorities who have managed to join it. "Bro" and "Fratboy" cultures, sexual harassment and gender bias in the workplace contribute to women in tech companies leaving their careers at the mid-point twice as often as men. Persisting, racist perceptions of people of colour as "non-technical" lead them to be marginalised in the tech field too.[46]

58. The lack of diversity in tech companies is not merely discriminatory in itself: it translates directly into discriminatory AI. Safiya Umoja Noble, in her work on search engines, has explored in depth how dominant groups are able to classify and organise the representations of others, in ways that reproduce and exacerbate racist prejudice.[47] The design of online forms also tends to correspond only to the simplest of life histories, excluding anyone who does not fit the norm.[48] The overwhelming use of female characters for chatbots, who are often cast in a subservient role, also perpetuates harmful gender stereotypes.[49]

---

[41] O'Neil C., *Weapons of Math Destruction*, New York, Broadway Books, 2016, p. 146-150 and 164-165.
[42] Noble S. U., *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, New York University Press, 2018, p. 66.
[43] Bucher T., *If…Then: Algorithmic Power and Politics*, New York: Oxford University Press, 2018, p. 158; Wachter-Boettcher S., Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech, New York, Londres, W. W. Norton, 2017.
[44] Resolution 2235 (2018) on Empowering women in the economy.
[45] Crawford K., "Artificial intelligence's white guy problem", *New York Times*, 25 June 2016.
[46] Presentation during the Committee's hearing on 12 September 2019 by Alice Coucke, Snips AI; Noble S. U., *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, New York University Press, 2018.
[47] Noble S. U., *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, New York University Press, 2018, p. 86.
[48] Wachter-Boettcher S., *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*, New York, London, W. W. Norton, 2017.
[49] EQUALS and UNESCO, "Think Piece 2: The rise of gendered AI and its troubling repercussions" in *I'd Blush If I Could': Closing gender divides in digital skills through education*, EQUALS and UNESCO, 2019.

59. To address these issues, it is vital – but not enough – to increase the diversity of the workforces engaged in developing AI, by taking decisive measures to improve women and minorities' access to the STEM professions. It is also crucial to ensure that all students of these subjects, who are designing technology for people, are trained and educated in the histories of marginalised people. Taking an interdisciplinary approach to designing AI – involving from the outset not only tech experts but experts from the social sciences, humanities and the law – would also go a long way towards preventing discrimination caused by the use of AI.

7. The "Black Box" syndrome: Transparency, explicability and accountability

"I was given no explanation. No way to make my case."
Jamie Heinemeier Hansson[50]

60. A common experience of victims of discrimination caused by the use of AI is that, even if they are able to show that discrimination occurred, they cannot obtain any explanation as to why the discriminatory outcome was produced. French students, for example, faced such situations when university applications began to be dealt with in 2018 under the new, online Parcoursup system.[51] Often, as in the Apple Card case referred to earlier, the designers of an algorithm are themselves unable to explain exactly why it has produced a given result or set of results. The specific elements that are taken into account by a particular algorithm, and the weight given to each element, are rarely known – and private companies that have invested in developing such algorithms are usually loath to reveal them, as they view them as their (highly valuable) intellectual property.

61. This lack of transparency and explicability is often referred to as the "black box" syndrome. It can make decisions reached through the use of AI extremely difficult for individuals to challenge. It also creates obstacles for national equality bodies seeking to support complainants in bringing cases in this field. Yet the use of AI can have seriously harmful consequences for some groups or individuals, discriminating directly or indirectly against them, and can exacerbate prejudice and marginalisation. It is crucial to ensure that effective and accessible remedies are in place to deal with discrimination when it occurs, and that the authors of such discrimination can be held accountable for it.

62. In the next chapter, I outline some elements that could be incorporated in legal norms in order to ensure that they provide a robust framework both for preventing discrimination caused by the use of AI and for dealing with cases when they arise. Here, I wish to emphasise that when companies or public authorities are confronted with evidence of discrimination, they do find ways to resolve the problem – whether due to public pressure or in order to give effect to a court decision. It is up to us to ensure that such measures are taken sooner, rather than later.

8. The need for a common set of legal standards, based on agreed ethical and human rights principles

63. Algorithmic bias is not a purely technical problem for researchers and tech practitioners but raises human rights issues that concern us all and that go beyond the (non-binding) ethical charters that have already been drawn up by major companies. We cannot afford to hide behind the complexities of AI in a form of learned helplessness that prevents our societies, and in particular us as legislators, from introducing regulations designed to protect and promote equality and non-discrimination in this field.[52] We urgently need procedures, tools and methods to regulate and audit these systems in order to ensure their compliance with international human rights standards. Effective domestic legislation is certainly needed – but so too are international standards, given the strong transnational and international dimensions of AI-based technologies.

64. Without seeking to be exhaustive, I would like to draw attention here to some key texts that have already been adopted as well as to some work that is ongoing at international level, and which are helping to prepare the way forward. In this respect, I note with interest the launching of the work of the Ad hoc Committee on Artificial Intelligence (CAHAI) on 18 November 2019. This intergovernmental committee of the Council of Europe is entrusted in particular with examining the feasibility and potential elements on the basis of broad multi-stakeholder consultations, of a legal framework for the development, design and application of artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law. [53] I also note with interest the recommendation of the Council of Europe's Commissioner for Human Rights, "Unboxing artificial intelligence: 10 steps to protect human rights", which clearly notes the importance of

[50] Heinemeier Hansson J., "About the Apple Card", blogpost, 11 November 2019.
[51] Défenseur des droits (France), Décision 2019-021 du 18 janvier 2019 relative au fonctionnement de la plateforme nationale de préinscription en première année de l'enseignement supérieur (Parcoursup).
[52] Zimmermann A. et al., "Technology can't fix algorithmic injustice", Boston Review, 9 January 2020.
[53] Terms of reference of the CAHAI approved by the Committee of Ministers on 11 September 2019, at its 1353th meeting.

equality and non-discrimination issues in this field. I also note the study entitled "Discrimination, artificial intelligence and algorithmic decision-making", which throws valuable light on the risks of discrimination caused by algorithmic decision-making processes and other types of AI.[54] Finally, I also welcome the contribution of Recommendation Rec/CM(2020)1 of the Committee of Ministers on the human rights impacts of algorithmic systems, accompanied by guidelines on addressing the human rights impacts of algorithmic systems.[55]

65. Beyond the work being carried out by the Council of Europe, many important initiatives at European and international level are worthy of mention, including the Ethics guidelines for trustworthy AI published by the European Commission in April 2019,[56] and a study issued by its Fundamental Rights Agency on Data quality and artificial intelligence – mitigating bias and error;[57] the OECD Principles on AI, adopted in May 2019;[58] the G20's declaration on human-centred artificial intelligence, adopted in June 2019;[59] and the work of several United Nations, including UNESCO, which was still under way at the time of drafting this report.

66. Based on an analysis of such key work at international level, the appendix to this report outlines five core ethical principles that must underpin all regulatory work in the field of AI: transparency, justice and fairness, responsibility (or accountability), safety and security, and privacy. It underlines that the extent to which respect for these core principles needs to built into particular AI systems depends on the intended and foreseeable uses to which those systems may be put. In essence, the greater the potential harm that may be caused by the use of a given AI system, the more stringent the requirements that should be observed. Here I would underline that the right to equality and non-discrimination is fundamental and intrinsic to human-rights- and law-based democracies. It must be strictly observed and protected at all times, and the use of AI must never be allowed to undermine these principles, whatever the context.

67. In the particular context of equality and non-discrimination, there are some points on which we must be especially vigilant. First, all grounds of discrimination need to be effectively covered by antidiscrimination legislation – including but not limited to real or perceived sex, gender, age, national or ethnic origin, colour, language, religious convictions, sexual orientation, gender identity, sex characteristics, social origins, civil status, disability, health status and so on. The list of protected grounds set out in law should be open – that is, while it should be as complete as possible, it should not be defined as an exhaustive list.

68. Second, both direct and indirect discrimination must be effectively covered. This is especially important where some countries do not allow the collection of certain data (for example, ethnic data) but algorithms infer such characteristics from proxies (country of birth of a person or their parents, postcode, search interests etc). For similar reasons, the law should also prohibit discrimination based on a person's real or supposed association with a certain group.[60]

69. Third, given the particular difficulties inherent in proving discrimination caused by the use of AI, it is crucial to ensure that victims of discrimination are not faced with an excessive burden of proof. The human rights consequences of requiring individuals to demonstrate their "innocence" in the face of automated decision-making can be immense, as I set out earlier in this report. The shared burden of proof set up under the EU Equality Directives provides a useful model here[61].

70. Fourth, effective enforcement mechanisms are essential, and the support of national equality bodies can be crucial here. Many have already begun thinking critically about the role they can play to guarantee and promote equality in a world where the use of AI is constantly expanding.[62] We should support that work and ensure that equality bodies have adequate resources to carry it out effectively

71. Finally, transparent business models are especially important where AI-based systems are affecting choices offered to individuals online, and where it is difficult for individuals to compare algorithmic outcomes. AI should also be subject to rigorous bias testing before it is deployed, and to equally rigorous, systematic and

---

[54] Zuideveen Borgesius F., Discrimination, artificial intelligence and algorithmic decision-making, study published by the Directorate General of Democracy, Council of Europe, 2018.
[55] Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems (adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies.
[56] High-Level Expert Group on AI, Ethics Guidelines for trustworthy AI, Brussels, European Commission, 8 April 2019.
[57] Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights, FRA, June 2019.
[58] Recommendation of the Council on Artificial Intelligence, OECD, 22 May 2019.
[59] https://www.mofa.go.jp/files/000486596.pdf.
[60] Wachter S., "Affinity profiling and discrimination by association in online behavioural advertising", Berkeley Technology Law Journal, 2020, 35(2) (forthcoming).
[61] Directives 2000/43/EC and 2000/78/EC.
[62] Allen R., Q.C., and Masters D., "Regulating for an equal AI: a new role for equality bodies", Equinet, Brussels, 2020.

periodic bias testing afterwards.[63] Periodical testing is even more important where machine learning is used to enable the algorithm to evolve after it is has been launched, and it may come to behave in ways that were not anticipated by developers from the outset.

72. The above considerations are specifically tied to issues of discrimination caused by the use of AI. More broadly speaking, I wish to stress the importance of involving civil society and citizens from the outset in any reflexions on these issues, to ensure critical participation and joint acceptance regarding algorithms. In addition, the regulation of algorithms could go hand in hand with building algorithmic alternatives proposed by the public authorities (through a genuinely proactive policy in this field), in order to counterbalance the purely commercial logic of the major tech companies; this technology policy based on a public and ethical initiative would be implemented in addition to regulatory policies.

73. As regards the recognition of human rights, the case-law of the European Court of Human Rights must, as always, guide States in the drafting in domestic legal standards and provide red lines for their action. Last, bearing in mind that AI is a rapidly developing field, it would seem useful to examine, in advance, the possibility of recognising new human rights such as the right to human autonomy, the right to transparency and justification, the right of overview of AI, or the right to moral integrity.

9. Conclusions

74. As is the case with most technological developments, artificial intelligence is neither inherently good nor inherently bad: like a sharp knife which can be used for both benign and very harmful purposes, it is not necessarily the tool that is the problem, but the way it is designed and (some of) the ways in which it can be used. My report is directed firmly towards ensuring that AI is not merely smart, but beneficial to humans and to our societies.

75. Using AI can have an impact on a number of rights (protection of personal data, private life, access to work, access to public services and welfare, cost of insurance, access to justice, the right to a fair trial, the burden of proof…), and may affect some groups more than others. While it does not necessarily create inequalities, it risks exacerbating those already existing in society and leading to unfair results in many fields of life.

76. Machine learning relies on the use of enormous datasets; but data are always biased, as they reflect the discrimination already present in society as well as the bias of those who collect and analyse the data.

77. Algorithms are created by people, and very often by homogeneous teams. The absence of diversity in the world of artificial intelligence creates a context that fosters the reproduction of damaging stereotypes. Diversifying the teams working in this field is a major concern, which requires work not only upstream, in the field of education, but also in ensuring that the working environment is inclusive in the longer term.

78. The lack of transparency of artificial intelligence systems often hides their discriminatory effects, which are observed only after the system has been deployed and many people have been subjected to its negative effects. To avoid such consequences, designers must integrate a pluralistic, multidisciplinary and inclusive perspective from the outset: who will use (or be affected?) by this sytem? Will it produce fair results for everybody? If not, what do I need to fix to correct the unfair results? How can the objectives of the tool (the results for which it should be optimised) be defined in order to ensure that they do not produce discriminatory results?

79. Evaluating the extent to which the principles of equality and non-discrimination are respected before an artificial intelligence tool is deployed is all the more important given that many human rights may be at stake. In addition, it can be extremely difficult to contest the outcomes of an algorithm, due to the “black box” effect of the latter. Yet often, it is those who are the most marginalised and the least able to contest the results who are most affected, sometimes with devastating consequences.

80. States must take up these questions and come to grips with the challenges they pose to our societies, now and for the future. The issues transcend borders and therefore also require transnational responses. Private actors in this field, who are the main designers of tools based on artificial intelligence, most also be directly involved in seeking solutions. Parliaments too must take a keen interest in the implications of the growing use of AI and its impact on citizens. Together, we need to regulate these systems in order to limit their discriminatory effects and provide an effective remedy when discrimination occurs.

[63] Spinks S., “Algorithmic bias within online behavioural advertising means public could be missing out, says Associate Professor Sandra Wachter”, Oxford Internet Institute blogpost, 26 November 2019.

81. Artificial intelligence is often linked in people's minds to innovation, but it must also, and more importantly, respond to another imperative: inclusion. The aim is not to hinder innovation but to regulate it in a manner that is proportionate to the issues at stake, so that artificial intelligence systems fully integrate respect for equality and the prohibition of all forms of discrimination.

Appendix

Artificial Intelligence – description and ethical principles

There have been many attempts to define the term "artificial intelligence" since it was first used in 1955. These efforts are intensifying as standard-setting bodies, including the Council of Europe, respond to the increasing power and ubiquity of AI by working towards its legal regulation. Nevertheless, there is still no single, universally accepted 'technical' or 'legal' definition.[64] *For the purposes of this report, however, it will be necessary to describe the concept.*

The term "artificial intelligence" is generally used nowadays to describe computer-based systems that can perceive and derive data from their environment, and then use statistical algorithms to process that data in order to produce results intended to achieve pre-determined goals. The algorithms consist of rules that may be established by human input, or set by the computer itself, which "trains" the algorithm by analysing massive datasets and continues to refine the rules as new data is received. The latter approach is known as "machine learning" (or "statistical learning") and is currently the technique most widely used for complex applications, having only become possible in recent years thanks to increases in computer processing power and the availability of sufficient data. "Deep learning" is a particularly advanced form of machine learning, using multiple layers of "artificial neural networks" to process data. The algorithms developed by these systems may not be entirely susceptible to human analysis or comprehension, which is why they are sometimes described as "black boxes" (a term that is also, but for a different reason, sometimes used to describe proprietary AI systems protected by intellectual property rights).

All current forms of AI are "narrow", meaning they are dedicated to a single, defined task. "Narrow" AI is also sometimes described as "weak", even if modern facial recognition, natural language processing, autonomous driving and medical diagnostic systems, for example, are incredibly sophisticated and perform certain complex tasks with astonishing speed and accuracy. "Artificial general intelligence", sometimes known as "strong" AI, able to perform all functions of the human brain, still lies in the future. "Artificial super-intelligence" refers to a system whose capabilities exceed those of the human brain.

As the number of areas in which artificial intelligence systems are being applied grows, spreading into fields with significant potential impact on individual rights and freedoms and on systems of democracy and the rule of law, increasing and increasingly urgent attention has been paid to the ethical dimension.

Numerous proposals have been made by a wide range of actors for sets of ethical principles that should be applied to AI systems. These proposals are rarely identical, differing both in the principles that they include and the ways in which those principles are defined. Research has shown that there is nevertheless extensive agreement on the core content of ethical principles that should be applied to AI systems, notably the following:[65]

- *Transparency.* The principle of transparency can be interpreted widely to include accessibility, explainability and explicability of an AI system, in other words the possibilities for an individual to understand how the system works and how it produces its results.
- *Justice and fairness.* This principle includes non-discrimination, impartiality, consistency and respect for diversity and plurality. It further implies the possibility for the subject of an AI system's operation to challenge the results, with the possibility of remedy and redress.
- *Responsibility.* This principle encompasses the requirement that a human being should be responsible for any decision affecting individual rights and freedoms, with defined accountability and legal liability for those decisions. This principle is thus closely related to that of justice and fairness.
- *Safety and security.* This implies that AI systems should be robust, secure against outside interference and safe against performing unintended actions, in accordance with the precautionary principle.
- *Privacy.* Whilst respect for human rights generally might be considered inherent in the principles of justice and fairness and of safety and security, the right to privacy is particularly important wherever an AI system is processing personal or private data. AI systems must therefore respect the binding standards of the EU General Data Protection Regulation (GDPR) and the Council of Europe's data protection convention 108 (and the 'modernised' convention 108+), as applicable.

[64] For a wide-ranging overview of attempts to define 'artificial intelligence', see *AI Watch: Defining Artificial Intelligence – Towards an operational definition and taxonomy of artificial intelligence,* Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., and Delipetrev, B., European Commission Joint Research Centre, 2020.

[65] See *AI Ethics Guidelines: European and Global Perspectives,* Draft Report commissioned by the Council of Europe Ad Hoc Committee on Artificial Intelligence (CAHAI), Ienca & Vayena, March 2020.

The effective implementation of ethical principles in relation to AI systems requires an ‘ethics by design’ approach, including a human rights impact assessment so as to ensure compliance with established standards. It is not sufficient for systems to be designed on the basis of technical standards only and for elements to be added at later stages in an attempt to evince respect for ethical principles.

The extent to which respect for these principles should be built into particular AI systems depends on the intended and foreseeable uses to which those systems may be put: the greater the potential impact on public interests and individual rights and freedoms, the more stringent the safeguards that are needed. Ethical regulation can thus be implemented in various ways, from voluntary internal charters for the least sensitive areas to binding legal standards for the most sensitive. In all cases, it should include independent oversight mechanisms, as appropriate to the level of regulation.

These core principles focus on the AI system and its immediate context. They are not intended to be exhaustive or to exclude wider ethical concerns, such as democracy (pluralistic public involvement in the preparation of ethical and regulatory standards), solidarity (recognising the differing perspectives of diverse groups) or sustainability (preserving the planetary environment).