Submission to the Consultation on the "White Paper on Artificial Intelligence - a European approach to excellence and trust"

Frederike Kaltheuner
Tech Policy Fellow
Mozilla Foundation

June 2020

Background

I am writing this submission in my capacity as a 2019-2020 Mozilla Tech Policy Fellow. Mozilla awards fellowship grants to aid fellows in the pursuit of research and study. Fellowships are designed to benefit the public interest.

Overview

As an advocate for responsible technology, I welcome the Commission's approach to human-centric AI that is grounded in fundamental rights such as human dignity and privacy protection. I especially appreciate that the White Paper assesses the impact of AI systems not only from an individual perspective, but also from the perspective of society as a whole.

In this submission, I aim to provide further information on the fundamental rights risks that can arise from certain uses of AI. I will also comment on the extent to which the policy options set out in the White Paper are effective in protecting individuals and society as a whole from significant risks associated with the development and application of AI in Europe.

Based on my research and experience, I provide the following observations:

- The risk-based approach chosen by the Commission is narrow and leaves people unprotected
- For certain groups of people *any* application of AI, not just those considered "high-risk" comes with an inherent risk of discrimination and exclusion
- Mitigating significant harms and risks remains challenging
- Some applications and uses of AI are incompatible with fundamental rights
- The public sector has a special duty of care when it comes to developing and deploying AI

I will conclude this submission with a number of recommendations that I encourage the Commission to consider:

- Include 'high-risk' applications of AI in 'low-risk' sectors in the definition of 'high-risk' AI
- Conduct a comprehensive review of applicable existing legal frameworks and identify needs for updates with an urgent focus on non-discrimination legislation
- Ensure that DPAs, consumer protection authorities, equality bodies and human rights monitoring bodies are sufficiently trained and funded to monitor and enforce existing legislation
- Remove discussions of discrimination from the section about training data and create a separate section for discrimination, bias and unfairness
- Broaden requirements to take reasonable measures to prevent discrimination beyond prohibited discrimination
- Ensure that AI that is designed, developed, and deployed in the EU adheres to high scientific standards
- Establish red lines to ban applications and uses of AI that are inherently incompatible with fundamental rights
- Add special mandatory requirements that ensure, at the very least, that the public sector does not procure technology that is opaque and inscrutable because it relies on proprietary algorithms

The risk-based approach leaves people unprotected

Not every application of AI needs in-depth scrutiny. However, the specific risk-based approach taken by the Commission in the White Paper limits any future regulatory framework to a very narrow set of AI applications. As a result of this approach, individuals and society at large are left unprotected from the risks associated with some of the most common and widespread uses of AI.

The current definition of 'high-risk' leave too many gaps

The White Paper relies on a particularity narrow definition of risk, where both the sector and the intended use need to involve significant risks. While it is understandable that 'low-risk' applications of AI in 'high-risk' sectors do not require additional safeguards, it is unclear why 'high-risk' AI applications — especially those that pose significant risks to fundamental rights — should be excluded just because they are being deployed in 'low-risk' sectors.

From AI-driven consumer products, data brokers, and the online marketing and Ad-Tech industry, to the personalisation and recommendation systems that fuel social media platforms, **this definition leaves individuals and society at large unprotected from fundamental rights violations in the very sectors that have seen some of the earliest and most widespread adoption of AI.**

Recommendation: include 'high-risk' applications of AI in 'low-risk' sectors in the definition of 'high-risk' AI

Existing legislation and enforcement mechanisms don't provide effective protection

It is indeed the case that developers and deployers of AI are already subject to European legislation on fundamental rights (e.g. data protection, privacy, non-discrimination), consumer protection, and product safety and liability rules, in addition to sectoral legislation.

The underlying assumption behind a narrow definition of risk seems to be that existing legislation, if further adjusted and clarified, and combined with voluntary labelling and a strengthened liability regime, would offer enough protection. However, this is not necessarily the case and the White Paper does not sufficiently explain and justify this assumption.

Existing legislation will need more than clarification and adjustment, especially when it comes to non-discrimination. As it well documented, AI-driven identification, profiling and automated decision-making may lead to unfair, discriminatory, or biased outcomes. Individuals can be misclassified, misidentified, or judged negatively, and such errors or biases may disproportionately affect certain groups of people. **A consequence of this is that for certain groups of people *any* application of AI, not just those considered "high-risk" comes with an inherent risk of discrimination and exclusion.**

Direct and indirect discrimination is already prohibited in many treaties and constitutions, including Article 14 of the European Convention on Human Rights.[1] Similarity, non-discrimination law, in particular through the concept of indirect discrimination, prohibits many discriminatory effects of AI and automated decision-making.

In practice, however, enforcement remains challenging. One reason that the Commission identifies in the White Paper is that affected persons might lack the means to verify how a decision made with the involvement of AI was taken. The specific characteristics of many AI technologies, including opacity ('black box-effect'), complexity, unpredictability and partially autonomous behaviour make it difficult to identify and prove possible breaches of laws.

Without further mandatory legal requirements, such as those proposed by the Commission for high-risk applications of AI, it will remain exceptionally challenging — if not impossible — for individuals and society as a whole to effectively challenge fundamental rights violations in applications of AI that fall outside the narrow definition of high-risk. As I have argued above, this is especially concerning for high-risk uses of AI in sectors that are deployed 'low-risk', and reinforces the urgent need to broaden the definition of high-risk on which future regulatory frameworks are based.

Take, for instance, the case of discriminatory online ads, an area that I presume would fall outside the scope of the Commission's proposed definition of high-risk. In 2019 a study by researchers at Northeastern University, the University of Southern California, and nonprofit organisation Upturn revealed that **Facebook's ad delivery algorithm discriminates based on race and gender, even when advertisers are trying to reach a broad audience and use neutral targeting parameter**.[2] The reason for this kind of discrimination is the fact that Facebook's advertising delivery system automatically optimises who sees an ad. Discrimination of this kind are already challenging to prove for researchers, who in the case of the paper mentioned above spent over $8,500 on ads that reached millions of people to find evidence that discrimination has occurred.[3] **Without further mandatory legal requirements, individuals who are subjected to such discriminatory practices would be unable to even know or provide evidence that they have been discriminated against.**

Beyond challenges to effective enforcement, current non-discrimination law leaves gaps, as the Council of Europe has argued in a report about "Discrimination, artificial intelligence, and algorithmic decision-making".[4] From the nuanced proportionality test required to establish that indirect discrimination has occurred, to the concept of protected characteristics, which non-discrimination laws typically focus, **AI can lead to discrimination and unfairnesses that fall outside the scope of existing laws.**

---

[1] See e.g. Article 7 of the United Nations Declaration of Human Rights; Article 26 of the International Covenant on Civil and Political Rights; Article 21 of the Charter of Fundamental Rights of the European Union

[2] Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A. and Rieke, A., 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), pp.1-30.

[3] Robertson, A (2019) 'Facebook's ad delivery could be inherently discriminatory, researchers say', The Verge, Apr 4. Available at: https://www.theverge.com/2019/4/4/18295190/facebook-ad-delivery-housing-job-race-gender-bias-study-northeastern-upturn (Accessed June 10, 2020).

[4] Council of Europe (CoE) 2018. *Discrimination, artificial intelligence, and algorithmic decision-making*. Available at: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 (Accessed: 10 June 2020).

The Commission should conduct a comprehensive and systematic review of existing laws with a special focus on areas where the nature of AI techniques undermine existing laws, or make their effective enforcement challenging. **Discrimination law in particular need to be fit for purpose to protect people from new and changing forms of discrimination.** Part of this review should also include an assessment of the the capacity and technical expertise of regulatory bodies needed to effectively enforce existing laws in a world that is increasingly data-driven and automated. In order to do enforce existing laws and regulations that already apply to AI, Data Protection Authorities (DPAs), consumer protection authorities, equality bodies and human rights monitoring bodies need proper funding and technical expertise to enforce existent laws in the context of AI.

Recommendation: conduct a comprehensive review of applicable existing legal frameworks and identify needs for updates with an urgent focus on non-discrimination legislation. Ensure that DPAs, consumer protection authorities, equality bodies and human rights monitoring bodies are sufficiently trained and funded to monitor and enforce existing legislation.

Mitigating significant harms and risks remains challenging

In the previous two sections, I have argued that the proposed scope for future regulatory frameworks is too narrow. In this section, I will discuss the proposed mandatory legal requirements for high-risk AI. The underlying assumption seems to be that these requirements, if further specified through standards, can reduce significant risks with regards to safety, consumer rights and fundamental rights.

Bias and discrimination can occur during all parts of the

The paper discusses requirements to take responsible measures aimed at ensuring that uses of AI systems do not lead to outcomes entailing prohibited discrimination exclusively in the context of training data. Bias, unfairness and discrimination, however, can creep in during all phases of a project of AI development, long before the data is collected, as well as at many other stages of the deep-learning process.[5]

Recommendation: remove discussions of discrimination from the section about training data and create a separate section for discrimination, bias and unfairness

Prohibited discrimination is only one part of the problem

By focussing exclusively on prohibited discrimination, the White Paper fails to address new forms of unfairnesses, biases and discrimination that AI systems can (inadvertently) produce. As the Council of Europe has explained in a report on Discrimination, artificial intelligence, and algorithmic decision-making:

[5] Barocas, S. and Selbst, A.D., 2016. Big data's disparate impact. *Calif. L. Rev.*, *104*, p.671.

> *"AI also opens the way for new types of unfair differentiation (some might say discrimination) that escape current laws. Most non-discrimination statutes apply only to discrimination on the basis of protected characteristics, such as skin colour. Such statutes do not apply if an AI system invents new classes, which do not correlate with protected characteristics, to differentiate between people. Such differentiation could still be unfair, however, for instance when it reinforces social inequality."* [6]

By limiting mandatory legal requirements to prohibited discrimination, the White Paper fails to address new forms of discrimination, while also ignoring wider societal harms that can result from high-risk applications of AI.

Recommendation: broaden requirements to take reasonable measures to prevent discrimination beyond prohibited discrimination

Fixing bias, unfairness and discrimination remains a complex challenge

Fixing bias, unfairness and discriminations remains challenging, even when companies take "reasonable steps" to reduce prohibited discrimination. A good example is a recent paper on 'solving' discrimination in automated hiring systems.[7] The paper analyses how three prominent automated hiring systems in regular use in the UK, HireVue, Pymetrics and Applied, understand and attempt to mitigate bias and discrimination. The paper concludes:

> *"Claims and validation are often vague and abstract, if they are provided at all. Moreover, it is not clear how relevant stakeholders, not least job seekers, are able to access and understand information about how decisions about their eligibility might have been reached through AHSs. This makes it difficult to assess if and how discriminatory practices might have been part of the hiring process, and leaves little room for anyone to challenge the decision made. Given the transparency rights attributed to data subjects by the GDPR, this haziness as to transparency is unacceptable in theEU and UK. On the other hand, what approach to transparency is required by EU law, remains itself vague."*

This study highlights the weaknesses of existing legal frameworks in addressing the risks and harms associated with applications of AI. It also shows how far away even companies that claim to take "reasonable steps" to reduce prohibited discrimination are from being truly accountable.

A major concern in the products analysed in this paper is the use of emotion recognition technologies, a technology whose shaky scientific foundations have recently become subject to

---

[6] Council of Europe (CoE) 2018. *Discrimination, artificial intelligence, and algorithmic decision-making*. Available at: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 (Accessed: 10 June 2020).

[7] Sánchez-Monedero, J., Dencik, L. and Edwards, L., 2020, January. What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 458-468).

heightened scrutiny.[8] **There remains little to no evidence that these new affect-recognition products have any scientific validity**. A major review released this summer found that efforts to "read out" people's internal states from an analysis of facial movements alone, without considering context, are at best incomplete and at worst entirely lack validity.[9]

The White Paper states, that AI systems and high-risk AI applications must be, amongst other things, technically robust and accurate in order to be trustworthy. But no AI system can be safe, secure, and reliable if it is based on flawed scientific premises.

Recommendation: ensure that AI that is designed, developed, and deployed in the EU adheres to high scientific standards

Some applications of AI are incompatible with fundamental rights

Precisely because it remains challenging to reliably 'fix' bias, discrimination and lack of explainability in AI systems, some applications and uses of AI systems that are fundamentally incompatible with fundamental rights, which merit moratoriums or bans.

I support the points raised on this matter in the submission by AccessNow[10], which demands that the EU must establish red lines to ban applications of AI which are incompatible with fundamental rights. Some uses that are a good starting point for such a discussion are:

- use of AI to solely determine access to or delivery of essential public services (such as social security, policing, migration control);
- uses of AI which purport to identify, analyse and assess emotion, mood, behaviour, and sensitive identity traits (such as race, disability) in the delivery of essential services;
- uses of AI to make behavioural predictions with significant effect on people based on past behaviour, group membership, or other characteristics such as predictive policing;
- use of AI systems at the border or in testing on marginalised groups, such as undocumented migrants;
- use for autonomous lethal weapons and other uses which identify targets for lethal force (such as law and immigration enforcement);
- use for general-purpose scoring of citizens or residents, otherwise referred to as unitary scoring or mass-scale citizen scoring; and
- applications of automation that are based on flawed scientific premises, such as inferring emotion from facial analysis.

---

[8] Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas,Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, RashidaRichardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. AI Now 2019 Report. NewYork: AI Now Institute, 2019, https://ainowinstitute.org/AI_Now_2019_Report.html.

[9] Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M. and Pollak, S.D., 2019. Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), pp. 1-68.

[10] AccessNow, 2020. 'Trust and excellence — the EU is missing the mark again on AI and human rights'. Available at: https://www.accessnow.org/trust-and-excellence-the-eu-is-missing-the-mark-again-on-ai-and-human-rights/ (Accessed: 10 June 2020).

Recommendation: establish red lines to ban applications and uses of AI that are inherently incompatible with fundamental rights

The deployment of live face recognition, especially if used in public spaces and by law enforcement deserves special consideration. This technology is inherently indiscriminate as live facial recognition detects all faces on video footage and then compares the faces against watchlists. Given that such systems inevitably processes the biometric data of everyone, and since explicit and informed consent is impractical in public spaces, live facial recognition has the potential to fundamentally change the power relationship between people and the police – and even alter the very meaning of public space.[11]

The technology is also exceptionally invasive. In response to an announcement made by the Metropolitan Police Service on the use of live facial recognition, the UK Information Commissioner's Office remarked: "Never before have we seen technologies with the potential for such widespread invasiveness."[12]

At a time when major technology companies like IBM, Microsoft and Amazon have all committed to not sell facial recognition to law enforcement (at least temporarily), **it is unclear how the deployment of the technology fits with Europe's ambition to become a world leader in trustworthy AI**.

The public sector has a special duty of care

Some of the most high-risk applications of AI are being developed or deployed by the public sector. Last year the UN rapporteur on extreme poverty produced a devastating account of the "digital welfare state"[13], which argued that new digital technologies are deteriorating the interaction between governments and the most vulnerable in society. A recent court judgement in the Netherlands[14] ruled that an automated surveillance system for detecting welfare fraud violated basic human rights.

Given that public sector uses have an important signalling function for the private sector, **it would only make sense to ensure that only trustworthy AI is being developed and deployed by the public sector in Europe.** Given the limitations of the risk-based approach, however, the White Paper currently does not propose that *all* AI systems used in the public sector take responsible measures aimed at ensuring that uses of AI systems do not lead to outcomes entailing prohibited discrimination. Under the current proposal, the public sector would also be allowed to use proprietary AI systems that are opaque and thus difficult to audit and test.

---

[11] Kaltheuner, F. (2020) 'Facial recognition cameras will put us all in an identity parade', The Guardian, 27 Jan. Available at: https://www.theguardian.com/commentisfree/2020/jan/27/facial-recognition-cameras-technology-police (Accessed: 10 June 2020).

[12] Ranger, S. (2019) ' We must stop smiling our way towards a surveillance state', ZDNet, 3 Nov. Available at: https://www.zdnet.com/article/we-must-stop-smiling-our-way-towards-a-surveillance-state/ (Accessed: 10 June 2020).

[13] United National Human Rights Office of the High Commissioner (OHCHR) 2019, *Digital technology, social protection and human rights: Report.* Available at: https://www.ohchr.org/EN/Issues/Poverty/Pages/SRExtremePovertyIndex.aspx (Accessed: 10 June 2020).

[14] Henley, J, Booth, R. (2020) ' Welfare surveillance system violates human rights, Dutch court rules ', The Guardian, 5 Feb. Available at: https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules (Accessed: 10 June 2020).

Recommendation: add special mandatory requirements that ensure — at the very least — that the public sector does not procure technology that is opaque and inscrutable because it relies on proprietary algorithms

Thank you and we look forward to continuing to collaborate with the EU institutions to develop a strong framework for the development of a trusted AI ecosystem