The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems

Preamble

1. As machine learning systems advance in capability and increase in use, we must examine the positive and negative implications of these technologies. We acknowledge the potential for these technologies to be used for good and to promote human rights but also the potential to intentionally or inadvertently discriminate against individuals or groups of people. We must keep our focus on how these technologies will affect individual human beings and human rights. In a world of machine learning systems, who will bear accountability for harming human rights?

2. As the "ethics" discourse gains ground, this Declaration aims to underline the centrality of the universal, binding and actionable body of human rights law and standards, which protect rights and provide a well-developed framework for remedies. They protect individuals against discrimination, promote inclusion, diversity and equity, and safeguards equality. Human rights are "universal, indivisible and interdependent and interrelated."[1]

3. This Declaration aims to build on existing discussions, principles and papers exploring the harms arising from this technology. The significant work done in this area by many experts has helped raise awareness about and inform discussions about the discriminatory risks of machine learning systems. We wish to complement this work by reaffirming the role of human rights law and standards in protecting individuals and groups from discrimination and non-equality in any context. The human rights law and standards outlined in this Declaration provide a solid grounding for the developing ethical frameworks for machine learning.

4. From policing, to welfare systems, online discourse, and healthcare – to name a few examples – systems employing machine learning technologies can vastly and rapidly change or reinforce power structures or inequalities on an unprecedented scale and with significant harm to human rights. There is a substantive and growing body of evidence to show that machine learning systems, which can be opaque and include unexplainable processes, can easily contribute to discriminatory or otherwise repressive practices if adopted without necessary safeguards.

5. States and private actors should promote the development and use of these technologies to help people more easily exercise and enjoy their human rights. For example, in healthcare, machine learning systems could bring advances in diagnostics and treatments, while potentially making health services more widely available and accessible. States and private actors should further, in relation to machine learning and artificial intelligence more broadly,

[1] Vienna Declaration and Programme of Action, http://www.ohchr.org/EN/ProfessionalInterest/Pages/Vienna.aspx

promote the positive right to the enjoyment of the benefits of scientific progress and its applications as an affirmation of economic, social and cultural rights.[2]

6. The rights to equality and non-discrimination are only two of the human rights that may be adversely affected through the use of machine learning systems: privacy, data protection, freedom of expression, participation in cultural life, equality before the law, and meaningful access to remedy are just some of the other rights that may be harmed with the misuse of this technology. Systems that make decisions and process data can also implicate economic, social, and cultural rights; for example, they can impact the provision of services and opportunities such as healthcare and education, and access to opportunities, such as labour and employment.

Whilst this Declaration is focused on machine learning technologies, many of the norms and principles included are equally applicable to artificial intelligence more widely, as well as to related data systems. The declaration focuses on the rights to equality and non-discrimination. Machine learning, and artificial intelligence more broadly, impact a wider array of human rights, such as the right to privacy, the right to freedom of expression, participation in cultural life, the right to remedy, and the right to life.

Using the framework of international human rights law

7. **States have obligations to promote, protect and respect human rights; private sector, including companies, has a responsibility to respect human rights at all times. We put forward this Declaration to affirm these obligations and responsibilities.**

8. There are many discussions taking place now at supranational, state and regional level, in technology companies, at academic institutions, in civil society and beyond, focussing on how to make AI human-centric and the "ethics" of artificial intelligence. There is need to consider current and future potential human rights infringements, and how best to address them with better thinking about harm to rights, and regulatory and legal regimes.

9. Human rights law is a universally ascribed system of values based on the rule of law which provides established means to ensure that rights, including the rights to equality and non-discrimination, are upheld. Its nature as a universally binding, actionable set of standards is particularly well-suited for borderless technologies such as machine learning. Human rights law provides both standards and mechanisms to hold the public and private sectors accountable where they fail to fulfil their respective obligations and responsibilities to protect and respect rights. It also requires that everyone must be able to obtain an effective remedy and redress where their rights have been denied or violated.

10. The risks machine learning systems pose must be urgently examined and addressed at governmental level and by the private sector conceiving, developing and, deploying these systems. Government measures should be binding and adequate to protect and promote rights. Academic, legal and civil society experts should be able to meaningfully participate in these discussions, critique and advise on the use of these technologies. It is also critical that potential harms are identified and addressed and that mechanisms are put in place to hold accountable those responsible for harms.

[2] Article 15 of the International Covenant on Economic, Social and Cultural Rights (ICESCR).

The rights to equality and non-discrimination

11. **This Declaration focuses on the rights to equality and non-discrimination, critical principles underpinning all human rights.**

12. Discrimination is defined under international law as “any distinction, exclusion, restriction or preference which is based on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status, and which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise by all persons, on an equal footing, of all rights and freedoms.”[3] This list is non-exhaustive as the United Nations High Commissioner for Human Rights has recognized the necessity of preventing discrimination against additional classes.[4]

Preventing discrimination

13. **The public and the private sector have obligations and responsibilities under human rights law to proactively prevent discrimination. When prevention is not sufficient or satisfactory, discrimination should be mitigated.**

14. In employing new technologies, both the public and the private sector will likely need to find new ways to protect human rights, as new challenges to equality and representation of diverse individuals and groups arise. These types of technologies can exacerbate discrimination at scale.

15. Existing patterns of structural discrimination may be reproduced and aggravated in situations that are particular to these technologies – for example, machine learning system goals that create self-fulfilling markers of success and reinforce patterns of inequality, or issues arising from using non-representative or “biased” datasets.

16. All actors, public and private, must prevent and mitigate discrimination risks in the design, development and, application of machine learning technologies and that ensure that effective remedies are in place before deployment and throughout the lifecycle of these systems.

Protecting the rights of all individuals and groups and promoting diversity and inclusion diversity

17. This Declaration underlines that inclusion, diversity, and equity are key components to ensuring that machine learning systems do not create or perpetuate discrimination, particularly against marginalised groups. There are some groups for whom collecting data on discrimination poses particular difficulty, however, protections must extend to those groups as well.

[3] United Nations Human Rights Committee, General comment No. 18 (1989), para. 7.

[4] Tackling Discrimination against Lesbian, Gay, Bi, Trans, & Intersex People Standards of Conduct for Business https://www.unfe.org/standards/ .

18. Intentional and inadvertent discriminatory inputs throughout the design, development and, use of machine learning systems create serious risks for human rights; systems are for the most part developed, applied and reviewed by actors which are largely based in particular countries and regions, with limited input from diverse groups in terms of race, culture, gender, and socio-economic backgrounds. This can produce discriminatory results.

19. Inclusion, diversity and equity entails the active participation of, and meaningful consultation with, a diverse community to ensure that machine learning systems are designed and used in ways that respect non-discrimination, equality and other human rights.

The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems

Preamble

Duties of states: human rights obligations

20. States bear the primary duty to promote, protect, respect, and fulfill human rights. Under international law, states must not engage in, or support discriminatory or otherwise rights-violating actions or practices when designing or implementing machine learning systems in public context or through public-private partnerships.

21. States must adhere to relevant national and international laws and regulations that codify and implement human rights obligations protecting against discriminatory and other harms, for example data protection and privacy laws. States also have positive obligations to promote equality and other rights and protect against discrimination by the private sector, including through binding laws.

22. The obligations outlined in this section also apply to public use of machine learning in partnership with the private sector.

State use of machine learning systems

23. **States must ensure that existing measures to prevent against discrimination and other rights harms are updated to take into account and address the risks posed by machine learning technologies.**

24. Machine learning technologies are increasingly being deployed or implemented by public authorities in areas that are fundamental to the exercise and enjoyment of human rights, rule of law, due process, freedom expression, criminal justice, healthcare, access to social welfare benefits, and housing. While there may be beneficial opportunities to the use of these technologies in such contexts, there may also be a high risk of discriminatory or other rights-harming outcomes. To the extent discrimination cannot be eliminated, it is critical that States provide meaningful opportunities for remediation and redress of harms.

25. As confirmed by the Human Rights Committee, Article 26 of the International Covenant on Civil and Political Rights “prohibits discrimination in law or in fact in any field regulated and protected by public authorities”.[5] This is further set out in treaties dealing with specific forms of discrimination, in which states have committed to refrain from engaging in discrimination, and to ensure that public authorities and institutions “act in conformity with this obligation.”[6]

26. States must refrain from using or requiring the private sector to use tools that discriminate, lead to discriminatory outcomes, or otherwise harm human rights. States must take steps to mitigate and reduce the harms of discrimination from machine learning.

Identifying risks

[5] United Nations Human Rights Committee, General comment No. 18 (1989), para 12.

[6] See Convention on the Elimination of All Forms of Racial Discrimination, Article 2 (a), and Convention on the Elimination of All Forms of Discrimination against Women, Article 2(d).

27. Any state deploying machine learning technologies must thoroughly investigate systems for discrimination and other rights risks prior to development or acquisition, where possible, prior to use, and on an ongoing basis throughout the lifecycle of the technologies, in the contexts in which they are deployed. This may include:

    a. Conducting regular impact assessments, prior to public procurement, during development, at regular milestones and through the deployment and use of machine learning systems to identify potential sources of discriminatory or other rights-harming outcomes – for example, in algorithmic model design, in oversight processes, or in data processing.[7]
    b. Taking appropriate measures to mitigate risks identified through impact assessments, for example, mitigating inadvertent discrimination or underrepresentation in data or systems, ensuring dynamic testing methods and pre-release trials, ensuring that potentially affected groups and field experts have been included as actors with decision-making power in the design, testing, and review phases, and subjecting systems to independent expert review where appropriate.
    c. Subjecting systems to live, regular tests and audits, interrogate markers of success, and holistic independent reviews of systems in context of human rights harms in a live environment.
    d. Disclosing known limitations with the system in question. These might include, for example, confidence measures, known failure scenarios, and appropriate limitations on use.

Ensuring transparency and accountability

28. States must ensure and require accountability and maximum possible transparency around public sector use of machine learning systems. This must include explainability and intelligibility in the use of these technologies so that the impact on affected individuals and groups can be effectively scrutinised by independent entities, responsibilities established, and actors held to account. States should:

    a. Publicly disclose where machine learning systems are used in the public sphere, provide information that explains in clear and accessible terms how automated and machine-learning decision-making processes are reached, and document actions taken to identify, document and mitigate against discriminatory or other rights-harming impacts.
    b. Enable independent analysis and oversight by using systems that are auditable.
    c. Avoid using ”black box systems” that cannot be subjected to meaningful standards of accountability and transparency, and refrain from using them in high-risk contexts.[8]

Enforcing oversight

29. States must take steps to ensure public officials are aware of and sensitive to the risks of discrimination and other rights harms in machine learning systems. States should:

[7] AI Now Institute has outlined a practical framework for algorithmic impact assessments by public agencies, https://ainowinstitute.org/aiareport2018.pdf.

[8] AI Now Institute, AI Now Report 2017, https://ainowinstitute.org/AI_Now_2017_Report.pdf.

a. Proactively adopt diverse hiring and equitable compensation practices, and engage in consultations to assure diverse perspectives so that those involved in the design, implementation, and review of machine learning represent a range of backgrounds and identities.
b. Ensure that public bodies carry out training in human rights and data analysis for officials involved in the procurement, development, use, and review of machine learning tools.
c. Create mechanisms for independent oversight, including by judicial authorities when necessary.
d. Ensure that machine learning supported decisions meet international accepted standards of due process.

30. As research and development of machine learning systems are being largely driven by the private sector, in practice States will often rely on private contractors to design and implement these technologies in a public context. In such cases, States must not relinquish their own obligations around preventing and ensuring accountability and redress for discrimination and other human rights harms in delivery of services.

31. Any state authority procuring machine learning technologies from the private sector should maintain relevant oversight and control over the use of the system, and require the third party to carry out human rights due diligence to identify, prevent and mitigate against discrimination and other human rights harms, and publicly account for their efforts in this regard.

Promoting equality

32. States have a duty to take proactive measures to eliminate discrimination.[9]

33. In the context of machine learning and wider technology development, one of the most important priorities for states is to promote programs to increase diversity, inclusion, and equity in the education and hiring in science, technology, engineering, and mathematics sectors. Such efforts serve as ends in themselves and help mitigate against discriminatory outcomes. States can also invest in research into ways to mitigate human rights harms in machine learning.

Holding private actors to account

34. International law clearly sets out the duty of states to protect human rights; this includes ensuring the right to non-discrimination by private actors.

35. According to the UN Committee on Economic, Social and Cultural Rights, “States parties must therefore adopt measures, which should include legislation, to ensure that individuals and entities in the private sphere do not discriminate on prohibited grounds".[10]

[9] The UN Committee on Economic, Social and Cultural Rights states that in addition to refraining from discriminatory actions, States parties should take “concrete, deliberate and targeted measures to ensure that discrimination in the exercise of Covenant rights is eliminated.” UN Committee on Economic, Social and Cultural Rights, general comment 20.

[10] UN Committee on Economic, Social and Cultural Rights, general comment 20.

36. States should put in place regulation compliant with human rights law for oversight of the use of machine learning by the private sector in contexts that present risk of discriminatory or other rights-harming outcomes, recognising technical standards may be complementary to regulation. In particular, non-discrimination, data protection, privacy and other areas of law on the national and regional level expand upon and reinforce international human rights obligations applicable to machine learning.

37. States must guarantee access to effective remedy for all individuals.

Responsibilities of private sector: human rights due diligence

38. The private sector has a responsibility that exists independent of state obligations to respect human rights.[11] As part of fulfilling this responsibility, private sector needs to take ongoing, proactive, and reactive steps to ensure that they do not cause or contribute to human rights abuses – a process called 'human rights due diligence'.[12]

39. Private sector entities that develop and deploy machine learning systems should follow a human rights due diligence framework in order to avoid fostering or entrenching discrimination and to respect human rights more broadly through the use of their systems.

40. Public sector entities developing machine learning are subject to the responsibilities listed above.

41. There are three core steps to the process of human rights due diligence:[13]

1. Identify potential discriminatory outcomes

42. During the development and deployment of any new machine learning technologies, non-state actors and the private sector should assess the risk that the system will result in discrimination. The risk of discrimination and the harms will not be equal in all applications, and the actions required to address discrimination will depend on the context. The private sector must be careful to identify not only direct discrimination, but also indirect forms of differential treatment which may appear neutral at face value, but lead to discrimination.

43. When mapping risks, private actors should take into account risks commonly associated with machine learning systems, including incomplete or "biased" training data, and those that arise in the design and deployment of algorithms. Private actors should consult with relevant stakeholders in an inclusive manner, including affected groups, organizations that work on human rights, equality and discrimination, as well as independent human rights and machine learning experts.

[11] *See* UN Guiding Principles on Business and Human Rights and additional supporting documents.

[12] *See* Council of Europe's Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries available at https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14.

[13] World Economic Forum, How to Prevent Discriminatory Outcomes in Machine Learning, http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf.

2. Take effective action to prevent and mitigate discrimination and document responses

44. After identifying human rights risks, the second step is to prevent those risks. For developers of machine learning systems, this requires:

    a. Correcting for discrimination, both in the design of the model and the impact of the system, and deciding which training data to use.
    b. Pursuing diversity, equity and, other means of inclusion in machine learning development teams. This will help to identify and prevent inadvertent discrimination.
    c. Submit systems that have a significant risk of resulting in human rights abuses to independent third party audits.

45. Where the risk of discrimination or other rights violations has been assessed to be too high or impossible to mitigate the private sector should consider not deploying a machine learning application.

46. Another vital element of this step is documenting the effectiveness of the private sector's response on impacts that emerge during the course of implementation and over time. This requires regular, ongoing quality assurances checks and real time auditing through the design, testing and deployment stages to monitor the application for discriminatory impacts, and correct errors and harms as appropriate. This is particularly important given the risk of feedback loops that can exacerbate and entrench discriminatory outcomes.

3. Be transparent about efforts to identify, prevent, and mitigate against discrimination in machine learning

47. Transparency is a key component of human rights due diligence, and involves "communication, providing a measure of transparency and accountability to individuals or groups who may be impacted and to other relevant stakeholders."[14]

48. Private sector entities that develop and implement machine learning applications should explain the process of identifying risks, the risks that have been identified, and the concrete steps taken to prevent and mitigate identified human rights risks. This may include:

    a. In instances where there is a risk of discrimination, publishing technical specification with details of the machine learning application and its functions, including samples of the training data used and details of the source of data.
    b. Establishing mechanisms to ensure that where discrimination has occurred as a result of a decision-making algorithm relevant parties, including affected individuals, are informed of the harms and how they can challenge a decision or outcome.

[14] UN Guiding Principles on Business and Human Rights, principle 21.

The right to an effective remedy

49. The right to justice is a vital element of international human rights law.[15] Under international law, victims of human rights violations or abuses must have access to prompt and effective remedies, and those responsible for the violations must be held to account.

50. Companies and private entities designing and implementing machine learning applications should take action to ensure individuals and groups have access to meaningful remedy and redress. This may include, for example, creating clear, independent, and visible processes for redress following adverse individual or societal effects, and designating roles in the entity responsible for the timely remedy of such issues subject to accessible and effective appeal and judicial review.

51. The use of machine learning systems where people's rights are at stake may pose challenges for ensuring the right to remedy. The opacity of some systems means individuals may be unaware how decisions which affect their rights were made, and whether the process was discriminatory. In some cases, the public body or private entity involved may itself be unable to explain the decision-making process.

52. The challenges are particularly acute when automated systems that make or enforce decisions are used within the justice system, the very institutions which are responsible for guaranteeing rights, including the right to access to remedy.

53. The measures already outlined around identifying, documenting, and responding to discrimination, and being transparent and accountable about these efforts, will help state bodies to ensure that individuals have access to effective remedies. In addition, states should:

    a. Ensure that if machine learning is to be used in the public sector, such use is carried out in line with standards of due process.
    b. Act cautiously on the use of machine learning system in the justice systems given the risks for fair trial and litigants rights.[16]
    c. Outline clear lines of accountability for the development and implementation of machine learning applications and clarify which bodies or individuals are legally responsible for decisions made through the use of such systems.
    d. Put in place effective penalties and sanctions for public or private bodies responsible for discriminatory outcomes through the use of machine learning systems where they have failed to take appropriate action to prevent or mitigate such impacts. This may be possible using existing laws and regulations or may require developing new ones.

[15] See for example Article 8, Universal Declaration of Human Rights; Article 2 (3), International Covenant on Civil and Political Rights; Article 2, International Covenant on Economic, Social and Cultural Rights; Committee on Economic, Social and Cultural Rights. *General Comment No. 3: The Nature of States Parties' Obligations (Art. 2, Para. 1, of the Covenant*) (1990) UN Doc E/1991/23 [5]; Article 6, International Convention on the Elimination of All Forms of Racial Discrimination; Article 2, Convention on the Elimination of All Forms of Discrimination against Women and UN Committee on Economic, Social and Cultural Rights (CESCR), *General Comment No. 9: The domestic application of the Covenant*, 3 December 1998, E/C.12/1998/24, available at: http://www.refworld.org/docid/47a7079d6.html.

[16] See ProPublica, Machine Bias https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Conclusion

54. The signatories of this Declaration call on the private and public sector to uphold their obligations and responsibilities under human rights laws and standards, in particular to avoid discrimination in the use of machine learning systems.

55. We call on states and the private sector to work together and play an active and committed role in protecting individuals and groups against discrimination. When deploying machine learning systems, they must take meaningful measures to promote accountability and human rights including, but not limited to, equality and non-discrimination as per their obligations and responsibilities under international human rights law and standards.

56. Technological advances must uphold our human rights. We are at a crossroads where those with the power must act now to protect human rights, including the rights to non-discrimination and equality – and help safeguard the human rights that we are all entitled to now, and for future generations.

Drafting committee members

Anna Bacciarelli and Joe Westby, Amnesty International
Estelle Massé, Drew Mitnick and Fanny Hidvegi, Access Now
Boye Adegoke, Paradigm Initiative Nigeria
Frederike Kaltheuner, Privacy International
Malavika Jayaram, Digital Asia Hub
Yasodara Córdova, Researcher
Solon Barocas, Cornell University
William Isaac, HRDAG