The Toronto Declaration: Protecting the right 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 impact of this technology on human rights. We acknowledge the potential for machine learning and related systems to be used to promote human rights, but are increasingly concerned about the capability of such systems to facilitate intentional or inadvertent discrimination against certain individuals or groups of people. We must urgently address how these technologies will affect people and their rights. In a world of machine learning systems, who will bear accountability for harming human rights?

2. As discourse around ethics and artificial intelligence continues, this Declaration aims to draw attention to the relevant and well-established framework of international human rights law and standards. These universal, binding and actionable laws and standards provide tangible means to protect individuals from discrimination, to promote inclusion, diversity and equity, and to safeguard 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 of and inform discussions about

---

[1] UN Human Rights Committee, *Vienna Declaration and Programme of Action*, 1993, http://www.ohchr.org/EN/ProfessionalInterest/Pages/Vienna.aspx

the discriminatory risks of machine learning systems.[2] We wish to complement this existing work by reaffirming the role of human rights law and standards in protecting individuals and groups from discrimination in any context. The human rights law and standards referenced in this Declaration provide solid foundations for developing ethical frameworks for machine learning, including provisions for accountability and means for remedy.

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

5. States and private sector actors should promote the development and use of machine learning and related technologies where they help people exercise and enjoy their human rights. For example, in healthcare, machine learning systems could bring advances in diagnostics and treatments, while potentially making healthcare services more widely available and accessible. In relation to machine learning and artificial intelligence systems more broadly, states should promote the positive right to the enjoyment of developments in science and technology as an affirmation of economic, social and cultural rights.[3]

6. We focus in this Declaration on the right to equality and non-discrimination. There are numerous other human rights that may be adversely affected through the use

---

[2] For example, see the FAT/ML *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*; IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design*; *The Montreal Declaration for a Responsible Development of Artificial Intelligence*; The Asilomar AI Principles, developed by the Future of Life Institute.

[3] The International Covenant on Economic, Social and Cultural Rights (ICESCR), Article 15 https://www.ohchr.org/EN/ProfessionalInterest/Pages/CESCR.aspx

and misuse of machine learning systems, including the right to privacy and data protection, the right to freedom of expression and association, to participation in cultural life, equality before the law, and access to effective remedy. Systems that make decisions and process data can also undermine economic, social, and cultural rights; for example, they can impact the provision of vital services, such as healthcare and education, and limit access to opportunities like employment.

7. While this Declaration is focused on machine learning technologies, many of the norms and principles included here are equally applicable to technologies housed under the broader term of artificial intelligence, as well as to related data systems.

Index of Contents

Using the framework of international human rights law

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

9. 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 the ethics of artificial intelligence and how to make technology in this field human-centric. These issues must be analyzed through a human rights lens to assess current and future potential human rights harms created or facilitated by this technology, and to take concrete steps to address any risk of harm.

10. Human rights law is a universally ascribed system of values based on the rule of law. It provides established means to ensure that rights are upheld, including the rights to equality and non-discrimination. Its nature as a universally binding, actionable set of standards is particularly well-suited for borderless technologies. Human rights law sets standards and provides mechanisms to hold public and private sector actors 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 effective remedy and redress where their rights have been denied or violated.

11. The risks that machine learning systems pose must be urgently examined and addressed at governmental level and by private sector actors who are conceiving, developing and deploying these systems. It is critical that potential harms are identified and addressed and that mechanisms are put in place to hold those responsible for harms to account. 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, and critique and advise on the use of these technologies.

The right to equality and non-discrimination

12. **This Declaration focuses on the right to equality and non-discrimination, a critical principle that underpins all human rights.**

13. 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."[4] This list is non-exhaustive as the United Nations High Commissioner for Human Rights has recognized the necessity of preventing discrimination against additional classes.[5]

Preventing discrimination

14. **Governments have obligations and private sector actors have responsibilities to proactively prevent discrimination in order to comply with existing human rights law and standards. When prevention is not sufficient or satisfactory, and discrimination arises, a system should be interrogated and harms addressed immediately.**

15. In employing new technologies, both state and private sector actors will likely need to find new ways to protect human rights, as new challenges to equality and representation of and impact on diverse individuals and groups arise.

16. Existing patterns of structural discrimination may be reproduced and aggravated in situations that are particular to these technologies – for example, machine learning

---

[4] United Nations Human Rights Committee, General comment No. 18, UN Doc. RI/GEN/1/Rev.9 Vol. I (1989), para. 7

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

system goals that create self-fulfilling markers of success and reinforce patterns of inequality, or issues arising from using non-representative or biased datasets.

17. All actors, public and private, must prevent and mitigate against discrimination risks in the design, development and application of machine learning technologies. They must also ensure that there are mechanisms allowing for access to effective remedy in place before deployment and throughout a system's lifecycle.

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

18. This Declaration underlines that inclusion, diversity and equity are key components of protecting and upholding the right to equality and non-discrimination. All must be considered in the development and deployment of machine learning systems in order to prevent discrimination, particularly against marginalised groups.

19. While the collection of data can help mitigate discrimination, there are some groups for whom collecting data on discrimination poses particular difficulty. Additional protections must extend to those groups, including protections for sensitive data.

20. Implicit and inadvertent bias through design creates another means for discrimination, where the conception, development and end use of machine learning systems is largely overseen by a particular sector of society. This technology is at present largely developed, applied and reviewed by companies based in certain countries and regions; the people behind the technology bring their own biases, and are likely to have limited input from diverse groups in terms of race, culture, gender, and socio-economic backgrounds.

21. Inclusion, diversity and equity entails the active participation of, and meaningful consultation with, a diverse community, including end users, during the design and application of machine learning systems, to help ensure that systems are created and used in ways that respect rights – particularly the rights of marginalised groups who are vulnerable to discrimination.

Duties of states: human rights obligations

22. States bear the primary duty to promote, protect, respect and fulfil 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 a public context or through public-private partnerships.

23. States must adhere to relevant national and international laws and regulations that codify and implement human rights obligations protecting against discrimination and other related rights harms, for example data protection and privacy laws.

24. States have positive obligations to protect against discrimination by private sector actors and promote equality and other rights, including through binding laws.

25. The state obligations outlined in this section also apply to public use of machine learning in partnerships with private sector actors.

State use of machine learning systems

26. **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.**

27. Machine learning systems 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 of expression, criminal justice, healthcare, access to social welfare benefits, and housing. While this technology may offer benefits in such contexts, there may also be a high risk of discriminatory or other rights-harming outcomes. It is critical that states provide meaningful opportunities for effective remediation and redress of harms where they do occur.

28. 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".[6] 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".[7]

29. States must refrain altogether from using or requiring the private sector to use tools that discriminate, lead to discriminatory outcomes, or otherwise harm human rights.

30. **States must take the following steps to mitigate and reduce the harms of discrimination from machine learning in public sector systems:**

i. Identify risks

31. 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 throughout 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.[8]
    b) Taking appropriate measures to mitigate risks identified through impact assessments – for example, mitigating inadvertent discrimination or underrepresentation in data or systems; conducting dynamic testing methods

---

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

7 For example, 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).

8 The AI Now Institute has outlined a practical framework for algorithmic impact assessments by public agencies, https://ainowinstitute.org/aiareport2018.pdf. Article 35 of the EU's General Data Protection Regulation (GDPR) sets out a requirement to carry out a Data Protection Impact Assessment (DPIA); in addition, Article 25 of the GDPR requires data protection principles to be applied by design and by default from the conception phase of a product, service or service and through its lifecycle.

and pre-release trials; ensuring that potentially affected groups and field experts are included as actors with decision-making power in the design, testing and review phases; submitting systems for independent expert review where appropriate.

c) Subjecting systems to live, regular tests and audits; interrogating markers of success for bias and self-fulfilling feedback loops; and ensuring holistic independent reviews of systems in the context of human rights harms in a live environment.

d) Disclosing known limitations of the system in question - for example, noting measures of confidence, known failure scenarios and appropriate limitations of use.

ii. Ensure transparency and accountability

32. 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 these systems at all in high-risk contexts.[9]

---

[9] The AI Now Institute at New York University, *AI Now 2017 Report*, 2017, https://ainowinstitute.org/AI_Now_2017_Report.pdf

iii. Enforce oversight

33. 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:
    a) Proactively adopt diverse hiring 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 for due process.

34. As research and development of machine learning systems is largely driven by the private sector, in practice states 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 discrimination and ensuring accountability and redress for human rights harms in the delivery of services.

35. 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

36. **States have a duty to take proactive measures to eliminate discrimination.**[10]

---

[10] The UN Committee on Economic, Social and Cultural Rights affirms that in addition to refraining from discriminatory actions, “State parties should take concrete, deliberate and targeted measures to ensure that discrimination in the exercise of Covenant rights is eliminated.” – UN

37. In the context of machine learning and wider technology developments, one of the most important priorities for states is to promote programs that increase diversity, inclusion and equity in the science, technology, engineering and mathematics sectors (commonly referred to as STEM fields). Such efforts do not serve as ends in themselves, though they may help mitigate against discriminatory outcomes. States should also invest in research into ways to mitigate human rights harms in machine learning systems.

Holding private sector actors to account

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

39. 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".[11]

40. 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 addition, non-discrimination, data protection, privacy and other areas of law at national and regional levels may expand upon and reinforce international human rights obligations applicable to machine learning.

41. States must guarantee access to effective remedy for all individuals whose rights are violated or abused through use of these technologies.

Committee on Economic, Social and Cultural Rights, General Comment 20, UN Doc. E/C.12/GC/20 (2009) para. 36

[11] UN Committee on Economic, Social and Cultural Rights, General Comment 20, UN Doc. E/C.12/GC/20 (2009) para. 11

Responsibilities of private sector actors: human rights due diligence

42. Private sector actors have a responsibility to respect human rights; this responsibility exists independently of state obligations.[12] As part of fulfilling this responsibility, private sector actors need 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'.[13]

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

44. **There are three core steps to the process of human rights due diligence:**

    i. Identify potential discriminatory outcomes
    ii. Take effective action to prevent and mitigate discrimination and track responses
    iii. Be transparent about efforts to identify, prevent and mitigate against discrimination in machine learning systems.

i. Identify potential discriminatory outcomes

45. During the development and deployment of any new machine learning technologies, non-state and private sector actors 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

---

[12] See UN Guiding Principles on Business and Human Rights and additional supporting documents

[13] 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, https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14

context. Actors 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.

46. When mapping risks, private sector actors should take into account risks commonly associated with machine learning systems – for example, training systems on incomplete or unrepresentative data, or datasets representing historic or systemic bias. 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.

ii. Take effective action to prevent and mitigate discrimination and track responses

47. 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 in deciding which training data to use.
    b) Pursuing diversity, equity and other means of inclusion in machine learning development teams, with the aim of identifying bias by design and preventing inadvertent discrimination.
    c) Submitting systems that have a significant risk of resulting in human rights abuses to independent third-party audits.

48. Where the risk of discrimination or other rights violations has been assessed to be too high or impossible to mitigate, private sector actors should not deploy a machine learning system in that context.

49. Another vital element of this step is for private sector actors to track their response to issues that emerge during implementation and over time, including evaluation of the effectiveness of responses. This requires regular, ongoing quality assurances checks and real-time auditing through design, testing and deployment stages to monitor a system for discriminatory impacts in context and situ, and to correct errors

and harms as appropriate. This is particularly important given the risk of feedback loops that can exacerbate and entrench discriminatory outcomes.

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

50. 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]

51. Private sector actors that develop and implement machine learning systems should disclose 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) Disclosing information about the risks and specific instances of discrimination the company has identified, for example risks associated with the way a particular machine learning system is designed, or with the use of machine learning systems in particular contexts.
    b) In instances where there is a risk of discrimination, publishing technical specification with details of the machine learning and its functions, including samples of the training data used and details of the source of data.
    c) Establishing mechanisms to ensure that where discrimination has occurred through the use of a machine learning system, relevant parties, including affected individuals, are informed of the harms and how they can challenge a decision or outcome.

The right to an effective remedy

52. 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

---

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

[15] For example, see: Universal Declaration of Human Rights, Article 8; International Covenant on Civil and Political Rights, Article 2 (3); International Covenant on Economic, Social and Cultural Rights, Article 2; Committee on Economic, Social and Cultural Rights, General Comment No. 3:

prompt and effective remedies, and those responsible for the violations must be held to account.

53. Companies and private sector actors designing and implementing machine learning systems should take action to ensure individuals and groups have access to meaningful, effective remedy and redress. This may include, for example, creating clear, independent, 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.

54. 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 sector actors involved may itself be unable to explain the decision-making process.

55. The challenges are particularly acute when machine learning systems that recommend, 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 effective remedy.

56. The measures already outlined around identifying, documenting, and responding to discrimination, and being transparent and accountable about these efforts, will help states to ensure that individuals have access to effective remedies. In addition, states should:
    a) Ensure that if machine learning systems are to be deployed in the public sector, use is carried out in line with standards of due process.

The Nature of States Parties' Obligations, UN Doc. E/1991/23 (1990) Article 2 Para. 1 of the Covenant; International Convention on the Elimination of All Forms of Racial Discrimination, Article 6; Convention on the Elimination of All Forms of Discrimination against Women and UN Committee on Economic, Social and Cultural Rights (CESCR), Article 2, General Comment No. 9: The domestic application of the Covenant, E/C.12/1998/24 (1998) http://www.refworld.org/docid/47a7079d6.html

b) Act cautiously on the use of machine learning systems in justice sector given the risks to fair trial and litigants' rights.[16]
c) Outline clear lines of accountability for the development and implementation of machine learning systems and clarify which bodies or individuals are legally responsible for decisions made through the use of such systems.
d) Provide effective remedies to victims of discriminatory harms linked to machine learning systems used by public or private bodies, including reparation that, where appropriate, can involve compensation, sanctions against those responsible, and guarantees of non-repetition. This may be possible using existing laws and regulations or may require developing new ones.

Conclusion

57. The signatories of this Declaration call for public and private sector actors to uphold their obligations and responsibilities under human rights laws and standards to avoid discrimination in the use of machine learning systems where possible. Where discrimination arises, measures to deliver the right to effective remedy must be in place.

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

59. Technological advances must not undermine our human rights. We are at a crossroads where those with the power must act now to protect human rights, and help safeguard the rights that we are all entitled to now, and for future generations.

---

[16] For example, see: Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner for ProPublica, *Machine Bias*, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

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, The Human Rights Data Analysis Group

This Declaration was published on 16 May 2018 by Amnesty International and Access Now, and launched at RightsCon 2018 in Toronto, Canada.