Artificial Intelligence and Human Rights. A Challenging Approach on the Issue of Equality

PhD Laura Stănilă*
Senior Lecturer
Faculty of Law
West University Timişoara

Abstract

Artificial Intelligence itself, its increased development and expanded use in all dimensions of social life, tends to rise very complex questions on the interactions between humans and AI systems. AI, with its algorithms, seems to provide models of perfect conduct aiming perfect objectivity. But is this really possible or are we to fall in a hidden trap, ready to lose one of our most important human rights – the right to equality and non-discrimination. Recent research had demonstrated that, despite careful programming, AI has generated situations of discriminatory labelling of certain social groups, in total opposition with our standards and goals. Can we prevent such results or the model of conduct imposed for AI is eventually going to fail?

Keywords: *Artificial Intelligence (AI), human rights, non-discrimination, right to equality, bias, risk assessment*

1. Artificial Intelligence (AI) – a new actor in the area of social relations

AI is a new actor in the area of social relations, having all the answers, simplifying decision making process for humans, due to its perfect mathematic algorithms. By "intelligence", Risse means "the ability to make predictions about the future and solve complex tasks (...) ability demonstrated by machines, in smart phones, tablets, laptops, drones, self-operating vehicles or robots that might take on tasks ranging from household support, companionship of sorts, even sexual companionship, to policing and warfare"[1].

Artificial intelligence systems aim to improve and are constantly changing the ways of action in the activity of companies and public authorities all around the world, and leading to a potential interference with human rights. As Andersen pointed out, "data protection laws and safeguards for accountability and transparency, may be able to mitigate some of the worst uses known today, but more work is necessary to safeguard human rights as AI technology gets more sophisticated and expands into other areas[2].

* E-mail: laura.stanila@e-uvt.ro.

[1] M. Risse, *Human rights and Artificial Intelligence. An Urgently Needed Agenda*, Carr Center for Human Rights Policy, USA, 2018, p. 2, https://carrcenter.hks.harvard.edu/files/cchr/files/human rightsai_designed.pdf, accessed on 10.12.2018.

[2] L. Andersen, 2018, *Human rights in the Age of Artificial Intelligence*, Access Now, https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf, p. 37.

But under the denomination of "Artificial Intelligence" (AI), there is a sum of systems, devices, technologies and algorithms. In fact, the term "Artificial Intelligence" was called "umbrella term"[3] under which AI systems may be classified into two categories[4]: *a) knowledge-based systems* and *b) systems with the capacity of continuously improving the decision-making performance.*

a) The knowledge-based systems include "expert systems" which use formal logic and coded rules to engage in reasoning. E.g. commercial tax preparation software, healthcare diagnostic decision support algorithms. These kind of AI systems are reasoning optimal decisions based on defined rules within a specific domain and usually cannot learn or automatically leverage the information they have accumulated over time to improve the quality of their decision-making process.

b) The systems with the capacity of continuously improving the decision-making performance use statistical learning in order to achieve the so-called "machine learning" and "deep learning" e.g. self-driving vehicles, facial recognition systems used by police forces, automate translation systems, algorithms that tell you what to watch next on video streaming services. This category of AI systems has been criticized as not being reliable at the individual level: "For example, deep learning computer vision systems can classify an image almost as accurately as a human; however, they will occasionally make mistakes that no human would make – such as mistaking a photo of a turtle for a gun"[5]. AI in the form of "neural networks" are increasingly used in technologies like self-driving cars, in order to be able to see and recognize objects. Such systems could be used to identify explosives in public spaces security lines, but, as studies have showed, they are making mistakes. A research team has showed that they were not only able to fool a neural network into thinking that a gun was another object, but they could actually determine the AI system to classify a physical object as anything they wanted. The research team has slightly changed the object's texture, so a bomb would get classified as a tomato, or could potentially even render an object entirely invisible for AI. The consequences in using such a system would be disastrous[6].

Another doctrinal classification divides AI systems into four categories[7] more comprehensive for the "AI non-educated" public:

1) systems that think like humans (e.g. cognitive architectures and neural networks);

2) systems that act like humans (e.g. pass the Turing test, knowledge representation, automated reasoning, and learning);

3) systems that think rationally (e.g. logic solvers, inference, optimization);

4) systems that act rationally (e.g. intelligent software agents and embodied robots that achieve goals via perception, planning, reasoning, learning, communicating, decision-making, and acting).

---

[3] F. Rasso, H. Hilligoss, V. Krishnamurthy, C. Bavitz, L. Kim, *Artificial Intelligence & Human Rights. Opportunities and Risks*, Berkman Klein Center for Internet and Society at Harvard University, 2018, p. 10, https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf, acces sed on 12.12.2018.

[4] *Idem*, pp. 10-11.

[5] A. Conner-Simons, *Fooling Neural Networks w/3D-Printed Objects*, MIT Computer Science, https://www.csail.mit.edu/news/fooling-neural-networks-w3d-printed-objects, accessed on 12.12.2018.

[6] *Idem*.

[7] S.J. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall Series in Artificial Intelligence, Englewood Cliffs, N.J Prentice Hall, 1995, pp. 31-53. https://www.cin.ufpe.br/~tfl2/ artificial-intelligence-modern-approach.9780131038059.25368.pdf, accessed on 2.12.2018.

2. Human rights and AI – are there any perils?

As reliance on artificial agents continues to grow, what are the consequences and risks of such an "addiction"? A better understanding of attitudes toward and interactions with algorithms is essential, due to "the aura of objectivity and infallibility today's culture ascribes to algorithms", as Osoba and Welser observed[8].

AI is widely used in the business sector and, because of that, the risks of its interfering with human rights was probably observed in the first place, certain counter-actions being triggered by the parties involved.

United Nations Guiding Principles on Business and Human Rights adopted by United Nations in 2011[9] is an international document seeking to provide an efficient risk management for enterprises involved in business relations from the point of view of interactions between human rights and business. AI is widely used by business enterprises with important risks for specific human rights and that is why we consider useful to present some of the guiding principles contained by this important document, which has not the power of law, but sets useful standards in order to safeguard human rights in the business sector.

Business enterprises may be involved with adverse human rights impacts either through their own activities or as a result of their business relationships with other parties. According to Principle 13, the responsibility to respect human rights requires that business enterprises:

a) Avoid causing or contributing to adverse human rights impacts through their own activities, and address such impacts when they occur;

b) Seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts.

In order to identify, prevent, mitigate and account for how they address their adverse human rights impacts, business enterprises should carry out human rights due diligence. The process should include assessing actual and potential human rights impacts, integrating and acting upon the findings, tracking responses, and communicating how impacts are addressed.

According to Principle 17, human rights due diligence[10]:

a) Should cover adverse human rights impacts that the business enterprise may cause or contribute to through its own activities, or which may be directly linked to its operations, products or services by its business relationships;

b) Will vary in complexity with the size of the business enterprise, the risk of severe human rights impacts, and the nature and context of its operations;

c) Should be ongoing, recognizing that the human rights risks may change over time, as the business enterprise's operations and operating context evolve.

---

[8] O. Osoba, W. Welser IV, *An Intelligence in Our Image. The Risks of Bias and Errors in Artificial intelligence*, 2017 RAND Corporation, p. iii, https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf, accessed on 10.12.2018.

[9] The Guiding Principles on Business and Human Rights seek to provide an authoritative global standard for preventing and addressing the risk of adverse human rights impacts linked to business activity. Full text in several languages available at www.ochr.org; English version available at https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf, accessed 19.12.2018.

[10] Principle 17 of the Guiding Principles on Business and Human Rights.

Human rights risks are understood to be the business enterprise's potential adverse human rights impacts. Potential impacts should be addressed through prevention or mitigation, while actual impacts – those that have already occurred – should be a subject for remediation in accordance to Principle 22 of UN United Nations Guiding Principles on Business and Human Rights.

Another sector of increasingly extended use of AI is the criminal justice system. The results of using AI in this field could also lead to dramatic breaches of human rights. The system of rights and guarantees for defendants and convicts could be easily eluded by the so-called risk-assessment tools which are in fact AI systems used to ease the decision-making process for humans in different phases of a case: pre-arrest phase, conviction phase etc. Risk assessments tools are usually used in different legal systems (U.S., Canada, and U.K.) to inform decisions about pre-trial detention, sentencing, and parole, due to their positive impact on the rights of individuals accused and convicted for crimes. Nevertheless, flaws or unknown limitations in the operation of such systems may produce harmful effects on the rights of indicted or convicted persons. In recent years, the use of algorithmic risk assessment tools in the criminal justice systems tends to spread and to be over-used. "All such tools automate the analysis of whatever data has been inputted into the system. Most of these tools still rely on manually-inputted data from questionnaires similar to those that were part and parcel of the last generation of risk-assessment tools, while newer tools are fully automated and rely on information that already exists in various government databases"[11].

As AI systems constantly developed, they needed more sophisticated data, so statisticians began to consider both static factors, such as a defendant's age and gender, as well as dynamic factors, such as a defendant's skill set or psychological profile[12].

Several risk assessment inventories are currently in use, one of the most used being the Level of Service Inventory-Revised (LSI-R). LSI-R is an assessment instrument originally developed in Canada through collaboration between academics, correctional psychologists and probation officers, being widely used in probation and custodial settings in Canada. It is in use also in other countries such as Scotland, and in about 20 probation services in England, Wales and the Channel Islands. It is the product of about 20 years' development, and a considerable amount of research has been carried out on its psychometric properties and its capacity to predict reconviction and various other correctional-relevant outcomes in North America[13].

Risk assessment tools in the U.S. criminal justice system have been critiqued as inherently unfair due to the disproportionate targeting of minority individuals and communities by the police. This, in turn, raises the risk that such tools will miscalculate the risk of recidivism for individuals from minority versus majority communities. This may also have deleterious effects on the rehabilitation of offenders from minority communities by impacting their access to cultural programming and their opportunities for parole[14].

---

[11] F. Rasso *et al.*, *op. cit.*, p. 22.

[12] T. Mathiesen, *Selective Incapacitation Revisited*, Law and Human Behavior, issue 22, no. 4 (1998), pp. 455-469.

[13] P. Raynor, J. Kynch, C. Roberts, S. Merrington, *Risk and need assessment in probation services: an evaluation*, Home Office Research Study 211, 2000, p. viii, https://www.researchgate.net/publication/267419138_Risk_and_Need_Assessment_in_Probation_Services_An_Evaluation/download, accessed on 12.12.2018.

[14] K.C. Land, *Automating Recidivism Risk Assessment: Should We Stay or Should We Go?*, Criminology & Public Policy issue 16, no. 1 (February 2017), pp. 231-233, https://doi.org/10.1111/1745-9133.12271, accessed on 12.12.2018.

Another risk assessment tool is COMPAS. COMPAS is one of the fourth-generation risk assessment instruments currently being utilized in US. COMPAS can be used to predict a variety of outcomes, and provides separate estimates for violence, recidivism, failure to appear, and community failure. Also similar to other fourth-generation tools, the COMPAS is both guided by theory (e.g. social learning theory, low self-control theory, strain theory, and social control theory) and provides gender specific calibrations[15].

But studies have shown that COMPAS has perpetuated racial bias, being used by U.S. state courts in making bail and sentencing decisions. The results of the study had shown how African-American offenders were labelled as "high-risk" at twice the rate of Caucasians[16].

Risk assessment tools as AI systems were qualified as unpredictable and opaque, an increasing awareness raising that specific human rights are to be particularly affected[17].

The use of algorithms may lead to rights violations or may undermine the effective enjoyment of these human rights in the following cases:

1. Fair trial and due process

Automated processing techniques and algorithms in crime prevention and the criminal justice system may affect the presumption of innocence and other procedural rights of the defendant. In the present high-risk society, following the terrorist attacks that took place in Europe and U.S., online social media platforms were and are used to identify potential terrorists and to become efficient in the fight against terrorism[18]. They are also used to identify accounts that generate extremist content. From this point of view, we must agree that there are consequences for the freedom of expression, but also for fair trial standards – art. 6 of the ECHR[19] –, notably the presumption of innocence, the right to be informed promptly of the cause and nature of an accusation, the right to a

---

[15] T. Blomberg, W. Bales, K. Mann, R. Meldrum, J. Nedelec, *Validation of the COMPAS risk assessement clasification instrument*, 2010, Center for Criminology and Public Policy Research, Florida, p. 8, http://criminology.fsu.edu/wp-content/uploads/Validation-of-the-COMPAS-Risk-Assessment-Classification-Instrument.pdf, accessed on 10.12.2018.

[16] J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks*, ProPublica, 23 May 2016, Study available on https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, accessed on 12.12.2018.

[17] Council of Europe, *Algorithms and Human Rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*, DGI (2017)12, Prepared by the Committee of Experts on Internet Intermediaries, (MSI-NET), p. 10, https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5, accessed on 15.12.2018.

[18] Lindsey Andersen, 2018, *Human rights in the Age of Artificial Intelligence*, Access Now, p.22, https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf, accessed on 12.12.2018.

[19] Article 6 par.1 and 2 of the ECHR – Right to a fair trial

1. In the determination of his rights and obligations, or of any criminal charge against him, everyone is entitled to a fair and public hearing within a reasonable time by an independent and impartial tribunal established by law. Judgment shall be pronounced publicly but the press and public may be excluded from all or part of the trial in the interests of morals, public order or national security in a democratic society, where the interests of juveniles or the protection of the private life of the parties so require, or to the extent strictly necessary in the opinion of the court in special circumstances where publicity would prejudice the interests of justice.

2. Everyone charged with a criminal offence shall be presumed innocent until proved guilty according to law.

fair hearing and the right to defend oneself in person. Risk assessment tools may contribute, as shown, to prejudicial and discriminatory decision-making.

2. Privacy and data protection

Due to the fact that all risk assessment tools are using algorithms in order to collect and sort all sorts of data and images, serious questions about breaching right to respect private and family life provided by art. 8 of the ECHR[20], including the right to data protection. "Algorithms are used in online tracking and profiling of individuals whose browsing patterns are recorded by *cookies*. Moreover, behavioural data is processed from smart devices, such as location and other sensor data through apps on mobile devices, raising increasing challenges for privacy and data protection". Scholars emphasized that ML ("machine learning") models were developed, that can accurately estimate a person's age, gender, occupation, and marital status just from their cell phone location data. They were also able to predict a person's future location from past history and the location data of friends[21].

At the same time, everyone could notice that AI is enabling more invasive surveillance tools. The negative impact of AI-powered surveillance would be felt most acutely by the marginalized populations who are disproportionately targeted by security forces. Also, because permanent monitoring of the general population is neither necessary, nor proportionate to the goal of public safety or crime prevention, this will lead to the breach of the right to privacy with certainty[22].

3. Non-discrimination

The right to enjoy all human rights and fundamental freedoms without discrimination and in full equality is one of the most important rights in the democratic countries.

In a study prepared for the Council of Europe on algorithms and human rights, it was emphasized that "search algorithms and search engines by definition do not treat all information equally. While processes used to select and index information may be applied consistently, the search results will typically be ranked according to perceived relevance. Accordingly, different items of information will receive different degrees of visibility depending on which factors are taken into account by the ranking algorithm"[23].

A biased algorithm that systematically discriminates one group in society, for example based on their age, sexual orientation, race, gender or socio-economic standing, may raise considerable concerns not just in terms of the access to rights of the individual

---

[20] Article 8 of the ECHR – Right to respect for private and family life:

1. Everyone has the right to respect for his private and family life, his home and his correspondence.

2. There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

[21] S.M. Bellovin, R.M. Hutchins, T. Jebara, S. Zimmeck, *When enough is enough: Location tracking, mosaic theory, and machine learning,* New York University Journal of Law and Liberty, 8(2), (2014) 555 – 6, https://digitalcommons.law.umaryland.edu/fac_pubs/1375, accessed on 10.12.2018.

[22] L. Andersen, *op. cit.*, p. 21.

[23] Council of Europe, *Algorithms and Human Rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*, DGI (2017)12, Prepared by the Committee of Experts on Internet Intermediaries, (MSI-NET), p. 26, https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5, accessed on 15.12.2018.

end-users or customers affected by these decisions, but also for society as a whole. Some authors have even suggested that online services which use personalised rating systems are inherently likely to lead to discriminatory practices[24].

3. The question of equality

In order to develop ML process, the computers must be "fed" with huge amount of data consisting in texts, images or recordings of sounds (e.g. human voice) – and then adding a classifier to this data (e.g. The computer is shown an image of a woman working in an office and then labelling this as woman office worker. In time, the computer will learn to recognise similar images and be able to associate these images with women working in an office and eventually make predictions for things such as job candidate screening or making loan approvals[25]).

The platform LinkedIn was reported because highly-paid jobs were not displayed as frequently for searches by women as they were for men, because of the way its algorithms were written. And the algorithms were written that way because, in the beginning, man users of LinkedIn were predominantly looking for the high-paying jobs, so the ML ended up proposing these jobs to men – thus discriminating women and reinforcing the bias against them[26].

Also, automated testing and analysis of Google's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs[27].

Another research has showed that two prominent research-image collections – including one supported by Microsoft and Facebook – display a predictable gender bias in their depiction of activities such as cooking and sports. Images of shopping and washing are linked to women, for example, while coaching and shooting are tied to men[28]. If a photo set generally associates women with housework, software trained by studying those photos and their labels create an even stronger association with it[29].

The criminal justice system, increasingly relying on risk assessment tools as showed before, makes no exception on the discrimination issue and the breach of equality between citizens, being qualified as "in crisis"[30]. Studies have revealed

---

[24] R. Alex, S. Luke, *Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers,* July 30, 2016, International Journal of Communication, 10, p. 3777. Available at https://ssrn.com/abstract=2686227, accessed on 10.12.2018.

[25] B. Büchel, *Artificial intelligence could reinforce society's gender equality problems*, March 1st, 2018, The Conversation UK, http://theconversation.com/artificial-intelligence-could-reinforce-societys-gender-equality-problems-92631, accessed on 12.12.2018.

[26] H. Reese, *Bias in machine learning, and how to stop it*, TechRepublic 18 November 2016, https://www.techrepublic.com/article/bias-in-machine-learning-and-how-to-stop-it/, accessed on 10.12.2018.

[27] S. Gibbs, *Women less likely to be shown ads for high-paid jobs on Google, study shows*, 8 July 2015, The Guardian, https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study, accessed on 12.12.2018.

[28] T. Simonite, *Machines Taught by Photos Learn a Sexist View of Women*, Wired, 21.08.2017, https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/, accessed on 10.12.2018.

[29] B. Büchel, *op. cit.*

[30] V. Southerland, *With AI and Criminal Justice, the Devil is in the Data*, ACLU, 9 April 2018, https://www.aclu.org/issues/privacy-technology/surveillance-technologies/ai-and-criminal-justice-devil-data, accessed on 15.12.2018.

"persistent racial disparities at every stage, a different kind of justice for the haves and the have nots, and a system that neither rehabilitates individuals nor ensures public safety"[31].

When person is arrested in U.S., for example, he or she will be usually subjected to a pre-trial risk assessment tool in order to help the judge decide whether to incarcerate that person pending trial or to release that person. Some U.S. states have used these pre-trial AI tools at the sentencing and parole decision stage, in an attempt to predict the likelihood that someone will commit a new offense if released from prison[32].

But the negative consequences of this use are serious, because of the biases they reinforce and perpetuate. As pointed out by Southerland, "all risk assessment tools generally rely on historical, actuarial data". If the data is not accurate or is obtained due to some illegal or exaggerate conduct (e.g. policemen in a city are usually arresting people from a certain community in a period of time. If that data is fed to the computer, the predictive model will suggest that people from that community are generally more likely to commit crimes so the results of the predictive test will perpetuate the discriminating pattern). If the algorithms are not "cleaned" of the waste (discriminatory patterns), the use of such AI tools will produce more harm than benefit for the justice system.

A study revealed that Black defendants were more likely to be wrongly labelled high risk than white defendants, drawing attention on the perils of AI "scoring"[33]. "In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing. Rating a defendant's risk of future crime is often done in conjunction with an evaluation of a defendant's rehabilitation needs. The Justice Department's National Institute of Corrections now encourages the use of such combined assessments at every stage of the criminal justice process"[34].

The study revealed that only 20 percent of the people predicted to commit violent crimes actually reiterated their criminal conduct. Also, in forecasting who would re-offend, the algorithm made mistakes with African-American and white defendants at roughly the same rate but in very different ways. The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labelling them this way at almost twice the rate as white defendants. White defendants were mislabelled as low risk more often than black defendants[35]. As a matter of fact, 23.5% of white people were labelled higher risk, but didn't re-offend in opposition with 44.9% African-American. At the same time, 47.7% of white defendants were mislabelled as lower risk, yet DID re-offend, while 28.0% African-American labelled lower risk actually reiterated their criminal conduct[36].

---

[31] *Idem.*

[32] *Idem*.

[33] J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks*, ProPublica, 23 May 2016, Study available on https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, accessed on 12.12.2018.

[34] *Idem*.

[35] J. Angwin *et al.*, *op. cit.*

[36] *Idem*.

4. What is to be done to preserve the human rights in their everyday confrontation with AI?

In order to determine AI to ensure equality between citizens, the first common-sense suggestions would be[37]:

a) Diverse groups and viewpoints must be represented in all stages of AI development.

b) Data will almost always be skewed, so methods must be developed to detect and correct for bias.

c) Standardization of the human assumptions or rulesets implicit in any AI model is a good start to developing best practices.

Scholars have also proposed other general policy approaches that could reduce risks of AI affecting human rights[38]:

a) Creating a comprehensive data protection legislation which can anticipate and mitigate many human rights risks posed by AI.

b) Imposing high standards for the Government use of AI, including open procurement standards, human rights impact assessments, full transparency, and explainability and accountability processes.

c) Establishing internal ethics policies by the Companies and develop transparency, explainability, and accountability processes.

d) Encouraging research conducted into the potential human rights harms of AI systems and make investments in creating structures to respond to these risks.

The adoption of Toronto Declaration[39] is also an important step in drawing attention on the issue of human rights posed by AI. The act was adopted by a coalition of human rights and technology groups in May 2018 and establishes machine learning standards, calling on both governments and tech companies to ensure that algorithms respect basic principles of equality and non-discrimination. The Toronto declaration focuses on the obligation to prevent machine learning systems from discriminating, and in some cases violating, existing human rights law. The declaration was announced as part of an annual gathering of digital and human rights groups.

The Toronto Declaration uses the concept of "human rights due diligence" and proposes three core steps to be followed by private sector[40]:

1. Identifying potentially adverse outcomes for human rights. Private-sector actors should assess risks an AI system may cause or contribute to human rights violations. In doing this, actors must: identify both direct and indirect harm as well as emotional,

---

[37] S. Venkatachalam, *AI and Equality: Let's Get It Right This Time*, Forbes Tchnology Council, 28 July 2017, https://www.forbes.com/sites/forbestechcouncil/2017/07/28/ai-and-equality-lets-get-it-right-this-time/#5c17d8c34de8, accessed on 10.12.2018.

[38] L. Andersen, *Human rights in the Age of Artificial Intelligence*, 2018, Access Now, p. 30, https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf, accessed 10.12.2018.

[39] Full text of Toronto Declaration, available at https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf, accessed 12.12.2018.

[40] The statement calls on governments and companies to ensure that machine learning applications respect the principles of equality and non-discrimination. The document articulates the human rights norms that the public and private sector should meet to ensure that algorithms used in a wide array of fields – from policing and criminal justice to employment and education – are applied equally and fairly, and that those who believe their rights have been violated have a meaningful avenue to redress.

social, environmental, or other non-financial harm; consult with relevant stakeholders in an inclusive manner, particularly any affected groups, human rights organizations, and independent human rights and AI experts; if the system is intended for use by a government entity, both the public and private actors should conduct an assessment.

2. Taking effective action to prevent and mitigate the harms, as well as track the responses. After identifying human rights risks, private-sector actors must mitigate risks and track them over time.

3. Being transparent about efforts to identify, prevent, and mitigate the harms in AI systems. Transparency to all individuals and groups impacted as well as other relevant stakeholders is a key part of human rights due diligence, and involves communication. In practice, this means that private-sector actors must:

Although states have the primary duty to provide access to formal remedy in the case of human rights violations[41], in order to establish appropriate mechanisms for accountability and remedy, private-sector actors should take additional action to ensure access to meaningful, effective non-judicial remedy and redress including:

• Internal responsibility for development and implementation of an AI system.

• Commitment by third parties developing AI systems for third parties to clearly delineating responsibility and accountability between vendor and client, including the vendor's obligation to ensure proper training of the risks of the system as well as to mitigate the risk of function creep and misuse of an AI system.

• Creation of clear, transparent processes by which an individual can directly submit complaints and seek redress for human rights harms in a timely manner.

References

1. Andersen, L., *Human rights in the Age of Artificial Intelligence*, Access Now, 2018, https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf;

2. Angwin, J., Larson, J., Mattu S., Kirchner, L., *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks*, ProPublica, 23 May 2016, Study available on https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing;

3. Bellovin, S.M., Hutchins, R.M., Jebara, T., Zimmeck, S., *When enough is enough: Location tracking, mosaic theory, and machine learning*, New York University Journal of Law and Liberty, 8(2), 2014, https://digitalcommons.law.umaryland. edu/fac_ pubs/1375;

4. Blomberg, T., Bales, W., Mann, K., Meldrum, R., Nedelec, J., *Validation of the COMPAS risk assessement clasification instrument*, 2010, Center for Criminology and Public Policy Research, Florida, http://criminology.fsu.edu/wp-content/uploads/Validation-of-the-COMPAS-Risk-Assessment-Classification-Instrument.pdf;

5. Büchel, B., *Artificial intelligence could reinforce society's gender equality problems*, March 1st, 2018, The Conversation UK, http://theconversation.com/artificial-intelligence-could-reinforce-societys-gender-equality-problems-92631;

[41] According to art. 42 of the *Toronto Declaration*, the responsibility to respect human rights of the private sector actors exists independently of state obligations in this field.

6. Conner-Simons, A., *Fooling Neural Networks w/3D-Printed Objects*, MIT Computer Science, https://www.csail.mit.edu/news/fooling-neural-networks-w3d-printed-objects;

7. Council of Europe, *Algorithms and Human Rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*, DGI (2017)12, Prepared by the Committee of Experts on Internet Intermediaries, (MSI-NET), https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5;

8. European Convention on Human Rights, https://www.echr.coe.int/Documents/Convention_ENG.pdf;

9. Gibbs, S., *Women less likely to be shown ads for high-paid jobs on Google, study shows*, 8 July 2015, The Guardian, https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study;

10. Land, K.C., *Automating Recidivism Risk Assessment: Should We Stay or Should We Go*?, Criminology & Public Policy issue 16, no. 1 (February 2017), https://doi.org/10.1111/1745- 9133.12271;

11. Mathiesen, T., *Selective Incapacitation Revisited*, Law and Human Behavior, issue 22, no. 4 (1998);

12. Osoba, O., Welser IV, W., *An Intelligence in Our Image. The Risks of Bias and Errors in Artificial intelligence*, 2017 RAND Corporation, https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf;

13. Rasso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C., Kim, L., *Artificial Intelligence & Human Rights. Opportunities and Risks*, Berkman Klein Center for Internet and Society at Harvard University, 2018, https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf;

14. Raynor, P., Kynch, J., Roberts C., Merrington, S., *Risk and need assessment in probation services: an evaluation*, Home Office Research Study 211, 2000, https://www.researchgate.net/publication/267419138_Risk_and_Need_Assessment_in_Probation_Services_An_Evaluation/download;

15. Reese, H., *Bias in machine learning, and how to stop it*, TechRepublic 18 November 2016, https://www.techrepublic.com/article/bias-in-machine-learning-and-how-to-stop-it/;

16. Risse, M., *Human rights and Artificial Intelligence. An Urgently Needed Agenda*, Carr Center for Human Rights Policy, USA 2018, https://carrcenter.hks.harvard.edu/files/cchr/files/humanrightsai_designed.pdf;

17. Rosenblat, A., Stark, L., *Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers,* July 30, 2016, International Journal Of Communication, 10, https://ssrn.com/abstract=2686227;

18. Russell S.J., Norvig, P., *Artificial Intelligence: A Modern Approach*, Prentice Hall Series in Artificial Intelligence, Englewood Cliffs, N.J Prentice Hall, 1995, https://www.cin.ufpe.br/~tfl2/artificial-intelligence-modern-approach.9780131038059.25368.pdf;

19. Simonite, T., *Machines Taught by Photos Learn a Sexist View of Women*, Wired, 21.08.2017, https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women;

20. Southerland, V., *With AI and Criminal Justice, the Devil is in the Data*, ACLU, 9 April 2018, https://www.aclu.org/issues/privacy-technology/surveillance-technologies/ai-and-criminal-justice-devil-data;

21. Toronto Declaration, https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.

22. United Nations, Guiding Principles on Business and Human Rights, www.ochr.org.; English version: https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf;

23. Venkatachalam, S., *AI and Equality: Let's Get It Right This Time*, Forbes Technology Council, 28 July 2017, https://www.forbes.com/sites/forbestechcouncil/2017/07/28/ai-and-equality-lets-get-it-right-this-time/#5c17d8c34de8.