ISSN 1982-0496

APPLICABILITY OF HUMAN RIGHTS CONTROL MECHANISMS IN ALGORITHMIC DECISION-MAKING CASES

APLICABILIDADE DOS MECANISMOS DE CONTROLE DOS DIREITOS HUMANOS EM CASOS DE TOMADA DE DECISÕES POR ALGORITMOS.

Dominika Iwan

PhD e Professora assistente da Universidade da Silésia em Katowice, Polônia. Pesquisa se concentra no Direito Internacional Público, Direito Internacional Humanitário e Direito Internacional dos Direitos Humanos em particular.

ABSTRACT

A The purpose of this paper is to address impacts of algorithmic decision-making (hereinafter: ADM) on human rights by means of human rights control mechanisms. The fundamental research question with respect to intersection between ADM and human rights compliance is whether ADM-related human rights violations can be addressed by any of the human rights control mechanisms. Existing human rights treaties were adopted in a pre-digital era but nowadays human rights exist online even more than offline. In this sphere, there is a difference in state obligations and business responsibilities while deploying ADM tools. Whereas states are at the frontline of respecting, protecting and fulfilling their human rights obligations, private entities are seemed to be free to develop and use ADM for their commercial purposes. As a result, ADM-related human rights implications do not exist in vacuum. International human rights law, both universal and regional, have resources to address some breaches resulting from ADM.

DOI: 10.25192/issn.1982-0496.rdfd.v26i22286

1. INTRODUCTION

The paper seeks to assess human rights impacts of algorithmic decision-making (hereinafter: ADM) through human rights control mechanisms. The fundamental research questionwith respect to intersection between new technologies and human rights compliance is whether related human rights violations can be addressed by any of the procedural guarantees of international human rights law.

The methodology used was in-desk research on treaty and resolution bodies which could proceed with potential ADM-related human rights violations. The quantitative method was used to examine whether, if any, ADM-related cases have been held before human rights control mechanisms so far. The qualitative method allowed to verify a significance of non-discrimination and the right to privacy while using ADM systems, as well as the meaning of control mechanisms for human rights compliance.

The paper consists of three parts. Firstly, the definition and examples of ADM, as well as human rights considerations are presented. Then, it is analysed which of the human rights control mechanisms address ADM. Ideally, the focus should be paid on individual and inter-state complaints, as being primary expected to address and further prevent human rights violations. Unfortunately, there has been little, if any,explicit jurisprudence concerning ADM and AI so far. A significant contribution has been made within the framework of resolution bodies, particularly before the Human Rights Council (hereinafter: HRC) and special procedures established therein. Several special rapporteurs have already examined human rights impacts of increasing digitization of welfare states and of ADM tools in a private sphere.Eventually, there are some reactive and decentralised (either regional or domestic) efforts to control or supervise pre or post development and use of ADM. The paper concludes with final remarks evaluating control mechanisms that addressADM. In this sphere, there is a difference in state obligations and business responsibilities while deploying ADM tools. Whereas states are at the frontline of respecting, protecting and fulfilling their human rights obligations (collectively, states are further trustees of international human rights law[1]), private entities seem to be free to develop and use ADM for their commercial

purposes[2]. It is firstly because the private entities bear human rights responsibilities rather than human rights obligations, and, secondly, big-tech companies usually perform as intermediaries between an interested customer and a final user. As a result, human rights in ADM systems do not exist in vacuum, and both universal and regional human rights lawhave substantive and institutional resources to address, at least some, breaches resulting from ADM systems.

2. RELATED TERMINOLOGY AND CHALLENGES

New technologies, including AI, robotics, big data, Internet of Things, biotechnology, and algorithms, have been used both in public and private sphere. AI and ADM become key (also increasingly legal) concepts of the rapidly approaching digital era (called the Fourth Industrial Revolution)[3], which span geographical borders. Frontiers of technology are often shifted towards human rights violations, repression and censorship. And both these technologies(AI and ADM) present great challenges not only with regard to ethical dilemmas (such as a secession of human decisions to a machine), but for human rights in general. While ethics is diverse and varies even within a single state, international human rights law seems more appropriate to protect human beings against (sometimes severe) consequences of new technologies. The very nature of human rights opts in favour of international human rights law to address ADM, since human rights are universal, indivisible, and inherent to everyone. Similarly to new technologies, human rights span the geographical borders. Every human being is born with and possesses the same rights, regardless of the place they live, gender, race, religion, or other grounds.Notably, the World Economic Forum considered human rights the ‘hard edge’ of values and a central value for shaping ethical framework and normative standards in the systemic change of the Fourth Industrial Revolution. Therefore, it international human rights law presents a proper framework for discussing emerging technologies to make them more sustainable.

An example of such new technology, being an integral part of the science on Artificial Intelligence[4], is ADM. This is a step-by-step mathematical operation of a

[4]AI is considered a branch of computer science that creates systems able to perform some human tasks (narrow AI),for example voice recognition, autonomous cars, data analysis. AI is also defined as a strategy aiming at developing machines that would replace basically every human performance (general AI).

computer, by which a system producesa numerical answer (an outcome).ADM uses data and statistical analyses to classify people for the purpose of assessing their eligibility for a benefit (for example, social) or penalty (risk assessment, crime prevention). AI enthusiasts argue that algorithms cost less and are faster than humans[5]. ADM are first and foremost developed by private entities, forming a fundament of big-tech industry, such as Amazon, Google, Apple, Facebook. Examples involve employment screening, insurance eligibility, pricing algorithms, targeted advertising, job online advertising. Google searching engines are probably the most powerful tools used to shape our preferences in a wide spectrum of interests. Nonetheless, the use of ADM in the private sphere contributes to an increase interest among states that heavily rely on automatization of civil services, thus transcending into digital welfare states. Public sector uses ADM to mass surveillance (e.g. Skynet in China), assess pregnancy risks among teenagers (e.g. several provinces in Argentina), grant members of population with social benefits (e.g. SyRI in Denmark, profiling unemployment in Poland), support asylum services (e.g. Roborder in the EU), or determine penalty (e.g. COMPASS in the USA). The occurrence of ADM tools is not new. An old example reaches 1980s when a British medical school used ADM to assess students admissions, in which ADM program was trained on files prepared by the university employees in previous recruitments. The application of the program resulted in discrimination against women and persons with an immigrant background.

At the same time, ADM are distinguished from machine learning, with the former being an outcome of the latter. Machine learning is a process by which probabilistic algorithms are trained to improve performance over time. The process is possible thanks to an element of probability and an increase access to data. Nowadays, it becomes more and more clear that humans, by believing in neutrality of technology, become victims of their own success. Big-tech companies, pressured to block certain digital contents on their platforms, are often unable to properly catch the complexity of social interdependencies. For example, a YouTube tool designed to identify inappropriate content, blocked videos of independent media groups that documented serious human rights violations and war crimes in Syria. As a result, many videos documenting human rights and humanitarian violations were lost. Symptoms of challenges relate to a lack of transparency at the data and system level, poor quality

of input data, lack of accountability, etc. All of these can result in a violation of non-discrimination principle, or interference with the right to privacy, among others. These two rights are considered defensive curtains in ensuring the sustainability of the Fourth Industrial Revolution.

Many scholars argue that ADM systems are discriminatory, unfair, or at least biased. Algorithmic bias has been a subject of interest of human rights treaty and political bodies, including committees and special rapporteurs. A real-world example of possible ADM-related discrimination is the Northpointe Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). A risk assessment algorithm used in the U.S. criminal justice system turned out to be discriminating against race. The prohibition of discrimination occupies a particular place in international human rights law. All universal and regional human rights treaties include the prohibition of discrimination, whereas jurisprudence of international courts and tribunals refer to the customary, *ergaomnes*or even peremptory character of the principle of non-discrimination. The prohibition of discrimination has various consequences for states, which have to refrain from any activities that directly or indirectly aim at discrimination (either legal or factual) of a person under their jurisdiction. Consequently, states must not enact any legislation or pursue any procedure that discriminate the person against any of the protected grounds set forth in the treaties to which they are bound (a negative duty to respect). States shall further take affirmative action to prevent and to react to acts of third parties (both negative and positive duty to protect). Eventually, states shall provide every one with access (allocate resources) to their human rights (a positive duty to fulfil). Subsequently, each state obligation (to respect, protect and fulfil) concerning non-discrimination apply to ADM systems that are used by the state or under which jurisdiction the private entity develops or uses ADM systems. Non-compliance with any of these obligations gives rise to state responsibility, that can be invoked directly by individuals or another state before universal and regional human rights bodies. Discrimination has different, sometimes very subtle, facets. Pursuant to para. 7 of the General Comment No. 18 of the Human Rights Committee, discrimination means “any distinction, exclusion, restriction or preference which is based on any ground (…) 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"[6]. Direct discrimination (disparate treatment) occurs when a person is treated worse than another person because of a protected characteristic. On the other hand, indirect discrimination (disparate impact) occurs when apparently neutral laws or policies are applied in the same way for everyone but effects are disadvantaging, without justification, a group who share protected characteristic, and then a person is indirectly discrimination as a part of the group.

The source of discrimination in ADM systems is firstly caused by data provided to train algorithms. Data origin from different sources that can be biasedthemselves, so it seems that avoiding certain protected characteristic from being collected for the purposes of ADM would prevent discrimination. Nonetheless, research has revealed that simply removing discriminatory-related data (such as race, gender, age) is counterproductive since this type of data can be derived from other personal data. Another claim is that developers of ADM systems are biased or discriminatory. As a consequence, labelling reflects developer's biases and prejudices, because many developers are white men from the Global North. The problem also mirrors a lack of sufficient infrastructure among states of the Global South, which leads to underrepresentation in data sets of individuals residing in these states. In this sense, the Special Rapporteur on racism indicates that individuals from states of the Global South are least digitally ready to access digital platforms, particularly in times of COVID-19 pandemic. This is why, for example, Facebook has introduced new policies aiming at increasing diversity at the system level by hiring people of other backgrounds, such as women. Furthermore, non-discriminatory or unbiased algorithms do not necessarily equal data-driven algorithms. Kleinberg et al. distinguish situations in which ADM discrimination can and cannot occur. System design can result in discriminatory outcome, either through choice of outcome (attributing weight to different outcomes), choice of predictors (inputting variables given to train algorithms), or choice of training procedure (including dataset and structure of training data). According to these authors, discrimination is unlikely to occur in other algorithms' behaviours.They claim that choosing input variables is the data- (not human-)driven process, and access to this process, along with the outcome and the training data, allows to detect potential discriminatory intent of the designer. Last but not least,

[6] Human Rights Committee, *CCPR General Comment No. 18: Non-discrimination.*

human decision-making also involves biases, which can be either unconscious or conscious, so discrimination is not always a result of ADM system itself. Therefore, the right to non-discrimination would not always be violated while using the ADM systems.

The subsequent regime addressing ADM-related human rights violations can be achieved through data protection law. Borgesius argues that the shortcomings of non-discrimination law are supplemented by dataprotection law, as construedmainly upon the right to private life (as derived from art. 12 of the UDHR, art. 17 of the ICCPR, art. 8of the ECHR). The digital sphere has been brought to the attention of the UN General Assembly that adopted several resolutions devoted to state obligations and business responsibilities concerning the right to privacy. In the current state of financing models, the private entities are encouraged to collect personal data for commercial purposes (particularly for pricing algorithms and targeted advertising). The UN Secretary-General's Roadmap for digital cooperation of 2020 underpins the necessity of change in these financing models. The first legally binding international document referring to automated processing of personal data was adopted within the Council of Europe's framework (the Convention 108 of 1981). From the perspective of privacy implication of ADM, the important change was made in 2018 with the adoption of the Protocol amending the Convention 108 (open for signature on 8 November 2001). The objectives of the Protocol focus on free flow of data and respect for human rights and human dignity while caring for economic growth and sustainability. The Protocol sets up independent bodies to ensure supervision over processing of personal data.

The EU model of ADM governance is regulated in the General Data Protection Regulation (hereinafter: GDPR). Art. 22 of the GDPR stipulates that data subjects shall not be subject to decisions based solely on ADM. Art. 17 of the GDPR further introduces the right to be forgotten[7]. Nonetheless, the EU data protection law has its shortcomings. Firstly, control bodies are not equipped with mechanisms of sanctions. Secondly, the law applies to personal data, so ADM are partly out of the scope of application. The right to private life is separated from the right to the protection of

[7] A part of the civil society argues that the GDPR has mandated the right to explanation, as derived from safeguards against ADM, notification duties, or the right to access to information, set forth in the GDPR. Explanation is to concern system functionality and specific decisions either prior to or before decision-making process. See: A.D. Selbst, J. Powles, 'Meaningful information and the right to explanation,' [2017] 7 *International Data Privacy Law* 4, 233. With a critical note see: S. Wachter, B. Mittelstadt, and L. Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation', [2017] 7 *International Data Privacy Law* 2, 76, p. 78.

personal data. Since ADM systems do not collect personal data itself, they only create predictive models and label persons based on their personal data. Therefore, predictive models, as not related to any identifiable personal information, are not covered by the GDPR.

A symptom-like solution would be to incorporate the prohibition of discrimination in other laws, such as consumer law, competition law, criminal law, administrative law, where ADM is used. Also, self-regulation contributes to a better protection of the right to non-discrimination but it is rather a soft instrument adopted on a case-by-case basis. Therefore, human rights control mechanisms play an important role in safeguarding individuals against unlawful effects of AI. Within international human rights law framework, it will be now examined how effective particular control mechanisms in relation to ADM systems are.

3. CONTROL MECHANISMS

There are four general categories of human rights control mechanisms, namely: international complaints, individual complaints, local examinations, and independent reviews. Although most of them depend on the explicitacceptance of an interested state, procedural guarantees of human rights do not rely solely on claims against violating states, but on every legal path preventing or remedying human rights violations (at domestic and international level). Although national guarantees should serve as principal for ensuring human rights compliance, they are often ineffective. Therefore, international guarantees, as being subsidiary to domestic legislation, are inevitable to ensure that states comply with their human rights obligations. These procedures are particularly relevant in the context of ADM. ADM systems span geographical boundaries but human rights exist there. The control mechanisms usually do not apply to the private entities, but this gap is filled by civil societies worldwide. NGOs complement the sphere of business and human rights by reporting on human rights adverse impacts in the business, including by tools of naming and shaming.

Human rights control mechanisms can intervene on individual cases, examine human rights situations of a state, create bodies to collect evidence on human rights violations, provide early warning or implementation guides for states and non-states actors. International guarantees of human rights protection are divided into treaty- and resolution-based (political) bodies. The treaty monitoring bodies consist of courts and

tribunals, as well as committees, all of which are created under the framework of a particular human rights treaty[8]. They review a state party mainly on the basis of individual and inter-state complaints, and their decisions are binding upon states. Political bodies cover both states and topics, and are usually linked to the HRC, which appoints individual mandate holders (independent human rights experts, or special rapporteurs), and working groups. The HRC also pursues the Universal Periodic Review, under which all UN member states are reviewed every 5 years on their human rights records. Therefore, there are plenty of mechanisms that can address human rights violations resulting from the use of ADM tools.

Committee on the Elimination of Racial Discrimination prepared a General recommendation on preventing and combating racial profiling by law enforcement officials. The General recommendation No. 36 of 2020 directly refers to algorithmic biases and discrimination in racial profiling with the latter violating international human rights law.The concern has also been articulated by the Committee against Torture in the context of algorithmic profiling used for law enforcement purposes (such as predictive policing, risk assessments, surveillance technologies, DNA testing), because these systems in fact create a profile (and a generalisation) of a person based on their characteristics. According to the General recommendation No. 36 of the CERD, racial profiling occurs when it is committed by law enforcement authorities, without reasonable justification or objective criteria, based solely on protected grounds or in intersection with other protected grounds (such as religion, gender, disability, age, etc.), and is used in a context of law enforcement procedures (for example, combating terrorism or controlling immigration). The document concludes with a list of recommendations for both states and the private entities, with the pressure put into states' legislative and policy-related measures, education, monitoring, and accountability for algorithmic profiling. In the context of accountability, the CERD underpinned the role of complaints on discriminatorypractices in law enforcement procedures. An example of an individual complaint concerning data processing is the latest case Big Brother Watch and Others v the United Kingdom involved data sharing between intelligence services. The European Court of Human Rights dismissed the

[8] For example, the Human Rights Committee, the Committee on Economic, Social and Cultural Rights, the Committee against Torture, the Committee Eliminating Discrimination against Women, the Committee Eliminating Racial Discrimination, the European Court of Human Rights, the Inter-American Court of Human Rights, the African Tribunal on Human and People's Rights.

claim that art. 8 of the ECHR was violated due to sufficient justification for using mass surveillance and intelligence sharing for security purposes. The justification referred to difficulties in investigating terrorist and criminal threats from abroad.

Within the UN system a variety of human rights challenges arising out of ADM and AIhave been vividly discussed. Thesepolitical mechanisms are the oldest and the most universal tools for ensuring procedural guarantees for human rights compliance. Following competences of the UN General Assembly, a primary political task has been attributed to the HRC. It appoints thematic and state rapporteurs, as well as review individual complaints on wide-spread human rights violations irrespective of the place they were committed. States can declare standing invitations by which they announce that they will always accept all special procedures, and consequently experts and rapporteurs can undertake country visits without a need for a separate invitation. Controversially, the membership in the HRC is exclusively granted to states, which politically impact subjects and objects of attention.

Being not necessarily a part of the UN political system of human rights protection, special procedures have been operationalised in the area of digitalisation. The HRC used several permanent special procedures to examine areas of intersection between human rights and new technologies. K. Annan described these procedures as the ‘crown jewel’ of the UN human rights system.These special procedures, albeit appointed by states members of the HRC, involve independent and impartial experts that are not employed by the UN. Although they do not receive any salary for their work, observers are frontline troops of the UN human rights system. Their status is highly important because the mandate often comment on politically controversial issues. There are several determinants of special procedures’ impact and influence. Along with independence and impartiality, special procedures deliver expertise, flexibility, accessibility, cooperation and follow-up, that allow to reach wider audience, credibility, and consequently have a significant positive impact on human rights globally. It does not mean that observers are fully successful in monitoring, protecting and promoting human rights, because not every state is willing to cooperate and respond with the observers requests. Nonetheless, non-binding character of reports concerning ADM contributes to raising awareness in the area of human rights.Challenges of digital world have been presented, for example, in the report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance in 2020. The report indicated that existing social

inequities are exacerbated by digital technologies, giving rise to both direct and indirect discrimination, as well as intersectional discrimination having holistic or systemic effect on protected grounds[9]. Pursuant to the report, ADM systems reproduce, sometimes implicit, biases extracted from big data sets.

Another important part of the HRC's mandate is the Universal Periodic Review, which relies on the cooperation and dialogue with the state concerned. The UPR supplements human rights treaty mechanisms, therefore, among already existing instruments, itcan be used to address ADM impacts on human rights. The UPR enables reviewing human rights records of all UN member states. What makes it slightly different from other control mechanisms is the possibility of verification of the widest human rights situations, since the UPR is not limited to a particular human rights treaty.A suggested enforcement procedure for ensuring the respect for the prohibition of discrimination with regard to ADM, also used in the UPR,is the instrument of 'naming and shaming'. This is a public indication that a person, entity or a state have behaved unlawfully. In international human rights law, it remains an essential strategy for careful documentation and publicisation of human rights abuses, which also serves as a reliable source for the accountability.Naming and shaming allows to operationalise human rights in the wider context of international community covering states and non-state actors. This procedure further adapts to challenges of new technologies, enabling progress of the Fourth Industrial Revolution and taking into account human rights, by developing soft instruments for ensuring human rights compliance. Members of the international community, including the private entities, care about their public relations and like to be a good example or pioneers for other entities not only in the context of available technology, but also in human rights.

Another mechanism devoted to effectively optimising benefits and risks of digital technologies is based on the follow-up procedures. These procedures aim at ensuring that human rights recommendations are actually implemented. This particular instrument often create media attention, openness of state authorities to address human rights problems, impetus for taking steps to improve human rights compliance, new resources for an increase cooperation in the area of human rights. These procedures are not reserved for resolution-based mechanisms, but for both treaty and resolution bodies. Being a soft instrument for increasing human rights compliance,

[9] *The Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance: Racial discrimination and emerging digital technologies,* para. 4.

follow-up procedures are remote from lives of ordinary people. Nonetheless, they open governments and civil society to document, share and disseminate results of their actions. The High-level Panel on Digital Cooperation was convened by the UN Secretary General in 2018, and issued a report 'The Age of Digital Interdependence'. The report indicated that digital inclusion does not only cover access to the Internet and digital technologies. This inclusion should rely on policy frameworks that take into consideration economic and social inclusion (including digital literacy, inclusive and holistic digital access) to leave no one behind. As a result, the UN Secretary-General submitted to the UN General Assembly a Roadmap with recommendations for global digital cooperation. The UN Secretary-General stressed an importance of digital connectivity since the lack of access to the Internet poses direct risks to individuals in terms of health and life (especially in times of COVID-19 pandemic).

An important contribution to human rights adverse impacts in ADM systems results from the business activities. Although the private entities do not possess human rights obligations (they do not adhere to any international human rights treaties), they are responsible for human rights compliance too. The UN Guiding Principles on Business and Human Rights of 2011 divide issues concerning business responsibilities into three sections: 1) state's duty to protect;2) business responsibility to respect; as well as 3) state and business responsibilities to grant access to remedies. This soft law instrumenthas grown out of society expectations towards business as entities acting for the benefit of society as a whole. The first section develops state duty to protect against human rights violations within its territory or jurisdiction by third parties. The duty to protect covers prevention, investigation, punishment and redress for these violations committed by all business enterprises domiciled in the state territory or jurisdiction. The private entities that are owned, controlled or supported by a state, require taking additional steps on the state's side as long as the nexus between the state and the private entity entails that acts of the latter can be attributed to the former in terms of state responsibility. This legal or factual link is particularly important in providing support and services involving ADM systems by the private entities to states' authorities. It implies a stricter human rights due diligence with respect to the tools purchased by the state.

The responsibility to respect human rights lying onthe private entities means that they should avoid causing – what is called in the UN Guiding Principles – human rights impacts through their own activities. Business should address human rights

impacts attached to its activities. Importantly, business responsibilities to protect human rights, including through human rights due diligence, depend on its size. Big-tech companies obviously fall into an increase due diligence since they operate with a great amount of data concerning human beings. It means that their activities, including ADM systems, should be audited from the human rights perspective, among others. For example, after pressures from the US Congress and civil society, Facebook conducted civil rights audit in 2020, in which both discriminatory and privacy risks were assessed. Nonetheless, the Facebook's approach to human rights issues is considered reactive and selective (by not taking into consideration human rights impacts outside the U.S.), and some representatives of the civil society considered it as "nothing more than a PR exercise". Still, big-tech companies have been increasingly under pressure to inform on their human rights impacts, especially in using AI tools and ADM systems.

4. DECENTRALISED APPROACHES TO HUMAN RIGHTS IMPACTS

Nationallaws are at the frontline of fulfilling human rights by states. Here, some aspects of ADM systems have been or will be regulated in the near future. Thedomestic laws are very diverse but their substance aims at protecting personal information and increasing transparency in decision-making by imposing obligations on the private entities. Some states invoke long-standing data processing regulations, other adopt totally new legal instruments to protect individuals in the digital era. Despite the tools used in addressing new technologies, the mere reference to states actions and that they do not remain silent on the topic is a good sign.

In 2018, the Canadian Supreme Court considered the case of risk assessment tools used for indigenous persons, and the reliability of data used therein. The claimant, Mr. Ewert, argued that the assessment tools used in custody cases were trained on inaccurate data that did not include indigenous persons. However, the Supreme Court ruled out that the burden of proof concerning doubts about the accuracy of data lied with the alleged victim. This is what the Directive on Automated Decision-Making (hereinafter: DADM) is to ease since individuals have gained access to source coding, transparency of information, and supervision of ADM outcomes. The DADM was adopted in 2019 by the Treasury Board of Canada, and regulates deployment of ADM systems,developed or procured after April 1, 2020,to comply with fairness and due

process. It applies to all governmental and quasi-judicial institutions of Canada and requires reducing negative outcomes of AI.Under the DADM, AI is understood as "*information technology that performs tasks that would ordinarily require biological brainpower to accomplish, such as making sense of spoken language, learning behaviours, or solving problems*". The act refers to procedural fairness that is considered a guiding principle of decision-making, but the degree of fairness depends on two factors, namely the significance of decision,and decision's impact on rights and interests of individuals. The Government of Canada prepared an algorithmic impact assessment (AIA) tool that is to support developing and implementing phases of ADM systems. The AIA tool consists of detailed questions concerning various aspects of decision's impacts (including the right to privacy and equal treatment of women and men), and provides an organ with guidance on steps that are required prior to development or implementation of the system. Depending on the impact on rights of individuals or communities, the DADM divides decisions into four levels (no, moderate, high, and very high impact), which further translates to quality and quantity of involved governmental or non-governmental experts. In practical terms, the DADMenables the Assistant Deputy Minister responsible for programs involving ADM, among others, to notify that a decision will be made with ADM system, provide a meaningful explanations on decisions, release source code, document ADM decisions, monitor the data used by ADM systems for factors that unfairly impact decisions. The DADM with the AIA tool creates a framework for government only, and hence exists outside the judiciary. This clear distinction should be assessed positively since judges have remained active in verifying fairness of the outcomes of ADM from the perspective of not only domestic laws of Canada, but also state obligations resulting from international law and human rights law in particular. The DADM does not apply to national security systems, therefore, surveillance tools used to prevent crimes, for example,are left outside the scope of the Directive. Effects of directives set forth by the Treasury Board are mandatory to their addressees but form instructions to fulfill policy objectives rather than legal obligations in terms of responsibility because they do not create individual rights that can proceed with legal action. This soft instrument does not prevent state authorities from developing ADM systems in governmental applications,but creates red lines that nonetheless must be taken into account by decision-makers.

A far-reaching laws are intended in the US withthe adoption of the Algorithmic Accountability Act. The laws impose the private entities,either generating over $50

million per year, possess information on at least a million people, with an obligation to assess high-risk systems that involve personal data, including systems based on machine learning and AI. For the purpose of the bill, the high-risk systems concern all systems that may contribute to discrimination, facilitate evaluation of consumers' behavior, raise privacy concerns, involve personal data concerning race, religion and gender, among others.According to MacCarthy, the proposed Act strikes especially against content moderation algorithms, but when their actions are taken mistakenly can result in disparate impacts against vulnerable groups. Although heading towards the right direction, the proposal has been criticized for its selective framing that differentiate domestic obligations of companies depending on their size but irrespective of the risks decision-making in general (either algorithmic or human) would engender. The bill would further unfeasibly require impact assessments on every updates made to software. By not making public results of impact assessments, individuals would not be aware of potential risks of ADM systems and provide companies with feedback.

A reactive state approach is encountered in response to media revelations worldwide, for example poor track records on the right to privacy. The Facebook's Cambridge Analytica– political consulting company – was intended to collect and share data of Facebook users with third parties (even beyond Facebook's control), and manipulating presidential elections in the US. The revelations resulted in legal sues in several states against Facebook. The protection of personal information has also been a subject of the UK and Australian co-investigation on Clearview AI Inc. The company deployed a facial recognition app that collected biometrics of individuals without their consent, including by scraping images from other platforms. The Clearview platform allows users to upload a photo and link it to other photos collected from the Internet. However, the company cooperated with law enforcement companies and individuals from around the world in data sharing activities. The investigation aims at protecting personal information of UK and Australian citizens "*in a globalized data environment".* Similar investigation pursued by Canada resulted in removing Clearview from the country, and ruling that the company and the local police breached the Canadian law. On the one hand, the reactive approach can be considered as insufficient in protecting individuals against breaches of their human rights. On the other, such revelations and investigations are necessary to raise awareness among public, civil society, stakeholders and decision-makers. For example, following the Cambridge Analytica scandal, California adopted the Consumer Privacy Act that aims at handling and

collecting data about Californians by several US-based companies. This events push the society to pursue a real change in domestic legal systems to better protect against both state and business adverse human rights impacts.

Another way to control human rights impacts takes place before domestic courts. A landmark rulinghas been made by the Hague District Court in the SyRI case. The case involved risk models used by the government predominantly in areas with higher concentration of vulnerable groups. The court noted that SyRI carried out risks of discriminatory effects due to insufficient transparency and verifiability. The case dealt with an increase of digital welfare states, but is not the only example of a system used in welfare assessment for purposes of social scoring (another example is Prometeus in the Latin American states). The ruling of the Dutch court has been applauded by the UN Special Rapporteur on extreme poverty and human rights, as contributing to preventing spying on individuals and being a role-model precedent for other courts. India places itself on the other end of the spectrum. The Government of India uses Aadhaar - the largest biometric identification system that was initially implemented without a legal basis. However, the system was declared constitutional by the Supreme Court of India because of its vital and inevitable character in digitizing modern states. Therefore, the Supreme Court referred to a justification for the usage of the system.

5. CONCLUSIONS

Material guarantees of ADM's compliance cannot be met without the reinterpretation of the effective methods of state and business compliance with international human rights law. This is the state's treaty obligation to ensure that human rights are respected, protected and fulfilled, also with regard to business. Because of that, states shall take positive steps and affirmative action in relation to ADM. This can be done for example by conducting a public debate with private entities and engineers developing ADM system. It will contribute to the creation of trust and promotion of the effective and lawful use of AI in State activities.

The human rights control mechanisms differ in personal and substantive scope of application. Some, like individual and inter-state complaints, better perform in a limited and sterile number of instances, while others, like reports and reviews, allow to take a wider perspective that contributes to slow but voluntary change. All these

mechanisms are necessary for democratising and increasing interdependencies among the members of the international community as a whole. One can argue that state interests prevail over individual interests. Still, the concept of human rights long precedes the development of AI and robotics. International human rights law sets out the core rules for the protection of human dignity, whilst colliding with state interests. The material guarantees of compliance with international human rights law (and hence the prohibition of discrimination) in relation to ADM systems can be found *inter alia* in the level of autonomy of such robots, the assurance of a meaningful human control, and responsibility or liability for human rights violations[10]. Beyond doubt, the scope of the substantive guarantees of all human rights prescribed in these treaties are contextual. However, when non-discrimination and privacy will be given the central part in the development of any new technologies, the paradigmatic balance between state interests and human rights interests prevails for the dignity of a person.

Acknowledgments

The paper is a result of author's participation in a Post-Doctoral Program in "New Technologies and Law" that was conducted by the Mediterranea International Centre for Human Rights Research (DipartimentoDiGiES, "di Eccellenza in Italia" – Universita "Mediterranea" di Reggio Calabria).

References

Canada, Treasury Board Secretariat, *Directive on Automated Decision-Making*, 1 April 2019 (1 April 2021, *Ottawa:* ***Treasury Board Secretariat***) <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592> accessed 18 August 2021.

Canada, Algorithmic Impact Assessment Tool (1 April 2021, ***Government of Canada***), <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html> accessed 18 August 2021.

C. Orwat, '**Risks of Discrimination through the Use of Algorithms**' (Federal Anti-Discrimination Agency, Berlin: 2020). See also: B. Goodman, S. Flaxman, 'European Union regulations on algorithmic decision-making and a <<right to explanation>>,'

[10]AccessNow, 'Human Rights in the Age of Artificial Intelligence' (*AccessNow*, 2018), <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf> accessed 3 August 2021, p. 18.

[2017] 38 *AI Magazine* 3, 50, p. 53; X. Renzhe et al., 'Algorithmic Decision Making with Conditional Fairness,' [2020] *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. N. Criado, J.M. Such, 'Digital Discrimination,' in: K. Yeun and M. Lodge (eds), *Algorithmic Regulation* (Oxford, OUP: 2019),pp. 82-97.

Committee against Torture, *Concluding observations on the combined third to fifth periodic reports of the United States of America*, 19 December 2014**, UN Docs** CAT/C/USA/CO/3-5, para. 26.

Committee on the Elimination of Racial Discrimination, *General recommendation No. 36 (2020) on* ***preventing and combating racial profiling by law enforcement officials***, 17 December 2020, UN Doc CERD/C/GC/36, para. 12.

Convention for the Protection of Human Rights and Fundamental Freedoms (adopted 4 November 1950, entered into force 3 September 1953), as amended by Protocols Nos 11 and 14, Council of Europe Treaty Series No. 5.

Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (adopted 28 January 1981, entered into force 1 October 1985), European Treaty Series No. 108.

C. Kuner, D. Jerker B. Svantensson, F.H. Catem O. Lynskey, and C. Millard, **'Machine learning with personal data: is data protection law smart enough to meet the challenge?**,' [2017] 7 *International Data Privacy Law* 1, 1, p. 2.

Discrimination, *General recommendation No. 36 (2020) on preventing and combating racial profiling by law enforcement officials*, 17 December 2020, **UN Doc** CERD/C/GC/36, para. 6.

D.J. Brand, '**Algorithmic Decision-making and the Law**,' [2020] 12 JeDEM 1, 114, p. 117.

D.J. Karp, '**What is the responsibility to respect human rights? Reconsidering the <<respect, protect, and fulfil**>> framework,' [2020] 12 *International Theory* 1, 83, p. 89.

D. Sadler, **'Privacy office's Clearview AI inquiry still going'** (23 June 2021, *InnovationAus*), <https://www.innovationaus.com/privacy-offices-clearview-ai-investigation-still-going/> accessed 24 August 2021.

European Court of Human Rights, *Carvahlo Pinto de Sousa Morais v Portugal*, Judgment 25 July 2017, Application No. 17484/15, Concurring Opinion of Judge Yudkivska, p. 23.

E.M. Aswad, 'The Future of Freedom of Expression Online,' [2018] 17 ***Duke Law & Technology Review* 1**, 26, p. 39.

European Court of Human Rights, *Big Brother Watch and Others vs the United Kingdom* (judgment of 25 May 2021, App. Nos. 58170/13 and 24960/15, para. 501.

F.J. Zuiderveen Borgesius, 'Strengthening **legal protection against discrimination by algorithms and artificial intelligence**,' [2020] 24 *The International Journal of Human Rights* 10, 1572, p. 1574.

How do we build an ethical framework for the Fourth Industrial Revolution' (7 November 2018, ***World Economic Forum***), <https://www.weforum.org/agenda/2018/11/ethical-framework-fourth-industrial-revolution> accessed 6 August 2021.

Human Rights Committee, ***CCPR General Comment No. 18****: Non-discrimination.*

Human Rights Council, ***Resolution 17/4: Guiding Principles on Business and Human Rights*****,** 16 June 2011, HR/PUB/11/04.

Human Rights Council**,** ***Resolution 5/1***, 18 June 2007, UN Docs A/HRC/RES/5/1. As of 17 August 2021, standing invitations were notified by the 127 UN member states.

Human Rights Council, ***Resolution 17/4: Guiding Principles on Business and Human Rights*****.** AccessNow, 'Human Rights in the Age of Artificial Intelligence' (*AccessNow*, 2018), <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf> accessed 3 August 2021, p. 18.

Human Rights Committee**,** ***CCPR General Comment No. 18: Non-discrimination***, 37th Session, adopted 10 November 1989, para. 1.

India, the Supreme Court of India, *Justice K.S. Puttaswamy et al (Retd) vs Union of India et al.*

International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976), 999 UNTS 171.

Inter-American Court of Human Rights, *Juridical Condition and Rights of Undocumented Migrants*, para. 103

J. Larson. S. Mattu, L. Kirchner, and J. Angwin, '**How We Analyzed the COMPAS Recidivism Algorithm**' (23 May 2016, *ProPublica*), <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> accessed 6 August 2021.

J. New, '**How to Fix the Algorithmic Accountability** Act' (23 September 2019, *Center for Data Innovation*), <https://datainnovation.org/2019/09/how-to-fix-the-algorithmic-accountability-act/> accessed 24 August 2021

J. Kleinberg, J. Ludwig, S. Mullainathan, and C.S. Sunstein, **'Discrimination in the Age of Algorithms,**' [2018] *Journal of Legal Analysis* 10, 113, p. 138.

L.W. Murphy, **'Facebook's Civil Rights Audit – Final Report'** (8 July 2020, *Facebook*), <https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf> accessed 27 July 2021, p. 63.

L.W. Murphy, **'Facebook's Civil Rights Audit – Final Report'**.
Debating on the size of Facebook,C. Newton and Z. Schiffer stressed that the audit excluded other Facebook's owned platforms (such as Instagram and WhatsApp), as well as human rights impacts outside the U.S. See: C.Newton, Z. Schiffer, 'What a damning civil rights audit missed about Facebook. In Ignoring Facebook's size, it gave the company a free pass to continue operating mostly as is' (10 July 2020, *The Verge*), <https://www.theverge.com/interface/2020/7/10/21318718/facebook-civil-rights-audit-critique-size-congress> accessed 19 August 2021.

M. Balcerzak, **'Procedury ochrony praw człowieka i kontroli wykonywania zobowiązań przez państwa,'** in: B. Gronowska, T. Jasudowicz, M. Balcerzak, M. Lubiszewski, and R. Mizerski (eds), *Prawa człowieka i ich ochrona* (Dom Organizatora, Toruń: 2010), p. 167.

M. Hardt, E. Price, and N. Srebro, '**Equality of Opportunity in Supervised Learning'** (7 October 2016, *ArXiv*), <https://arxiv.org/pdf/1610.02413.pdf> accessed 27 July 2021, p. 1. See also: M. Altman, A. Wood, and E. Vayena, 'A Harm-Reduction Framework for Algorithmic Fairness,' [2018] 16 IEEE Security & Privacy 3, 34, p. 38.

M. MacCarthy**, 'An Examination of the Algorithmic Accountability Act of 2019**,' (24 October 2019), <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3615731> accessed 24 August 2021, p. 9.

M. Isaac, '**Facebook's Decisions Were** <<Setbacks for Civil Rights,>> Audit Finds' (8 July 2020, *NY Times*), <https://www.nytimes.com/2020/07/08/technology/facebook-civil-rights-audit.html> accessed 19 August 2021.

Netherlands, **The Hague District Court, *NJCM c.s. vs De Staat der Nederlanden (SyRI),*** Judgment, 5 February 2020, C-09-550982-HA ZA 18-388, ECLI:NL:RBDHA:2020:865 <https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878> accessed 20 July 2021.

Office of the High Commissioner for Human Rights, ***Guiding Principles on Business and Human Rights. Implementing the United Nations"Protect, Respect and Remedy"*** *Framework*, New York and Geneva, 2011, p. 7.

Office of the High Commissioner for Human Rights, 'Landmark ruling by Dutch court

stops government attempts to spy on the poor – **UN expert'** (5 February 2020, *OCHR*), <https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25522> accessed 24 August 2021.

Office of the High Commissioner for Human Rights, **'A Practical Guide for Civil Society: How To Follow Up On United Nations Human Rights Recommendations'** (OCHR, Geneva: 2016), p. 17-18.

Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (adopted 10 October 2018, not yet in force), European Treaty Series No. 223.

P. Zialcita, '**Facebook Pats $643,000 Fine For Role in Cambridge Analytica Scandal'** (30 October 2019, *NPR*), <https://www.npr.org/2019/10/30/774749376/facebook-pays-643-000-fine-for-role-in-cambridge-analytica-scandal> accessed 24 August 2021. See also: BBC News, 'Facebook <<to be fined $5bn over Cambridge Analytica scandal>>' (12 July 2019, *BBC News*), <https://www.bbc.com/news/world-us-canada-48972327> accessed 24 August 2021.

R. Fuller, **'Enforcing International Law: States, IOs, and Courts as Shaming Reference Groups**,' [2014] 39 *Brooklyn Journal of International Law* 1.

Report of the UN Secretary-General's High-level Panel on Digital Cooperation: The Age of Digital Interdependence**,** 10 June 2019 (*United Nations*), <https://www.un.org/en/pdfs/HLP%20on%20Digital%20Cooperation%20Report%20Executive%20Summary%20-%20ENG.pdf> accessed 11 August 2021.

Report of the Secretary-General: Road map for digital cooperation: ***implementation of the recommendations of the High-level Panel on Digital Cooperation***.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 4 May 2016, Official Journal 2016 L 119/1.

S. Gregory, '**Cameras Everywhere Revisited: How Digital Technologies and Social Media Aid and Inhibit Human Rights Documentation and Advocacy'**, [2019] *Journal of Human Rights Practice* 11, 373, p. 386-87.

S.P. Subedi, '**Protection of Human Rights through the Mechanism of UN Special Rapporteurs,'** [2011] 33 *Human Rights Quarterly* 1, 201, p. 228.

T. Dias Oliva, 'Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression', [2020] ***Human Rights Law Review*** **20**, 607, p. 608.

T. Dias Oliva, **‘ Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression**’, p. 609.

The Human Rights Committee sentenced that racial profiling violates international human rights law in a case Williams Lecraft v Spain. **The Committee considered racial profiling as unlawful discrimination**. See: Human Rights Committee, *Rosalind Williams Lecraft v Spain,* Views adopted 27 July 2009, Communication No. 1493/2006, 96th session, 13-31 July 2009, views, CCPR/C/96/D/1493/2006. See also: Committee on the Elimination of Racial

The Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance: Racial discrimination and emerging digital technologies, 18 June 2020, UN Docs A/HRC/44/57, para. 20.

T. Piccone, ‘**Human Rights Special Procedures: Determinants of Influence**,’ [2014] *Proceedings of the Annual Meeting (American Society of International Law* 108, *The Effectiveness of International Law,* 288, pp. 288-291.

The Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance: Racial discrimination and emerging digital technologies, para. 4.

T. Scassa, ‘**Administrative Law and the Governance of Automated Decision-Making: A Critical Look at Canada’s Directive on Automated Decision-Making**,’ [2021] 54 *University of British Columbia Law Review* 1 (forthcoming)

UNESCO, World Commission on the Ethics of Scientific Knowledge and Technology, *Preliminary study on the ethics of Artificial Intelligence,* Paris, 26 February 2019, SHS/COMEST/EXTWG-ETHICS-AI/2019/1. See also: International Telecommunication Union, *AI for Good Summit*, <https://aiforgood.itu.int />.

UN Committee on Economic, Social and Cultural Rights, *General Comment No. 12: **The Right to Adequate Food** (Art. 11 of the Covenant),* adopted on 12 May 1999, E/C.12/1999/5, para. 15. See also: D.J. Karp, ‘What is the responsibility to respect human rights? Reconsidering the <<respect, protect, and fulfil>>framework,’ p. 86.

United Nations**, ‘Press release: Annan calls on Human Rights Council to strive for unity, avoid familiar fault lines’** (29 November 2006, *United Nations*), <https://news.un.org/en/story/2006/11/201202-annan-calls-human-rights-council-strive-unity-avoid-familiar-fault-lines> accessed 17 August 2021. See also: S.P. Subedi, ‘Protection of Human Rights through the Mechanism of UN Special Rapporteurs,’ [2011] 33 *Human Rights Quarterly* 1, 201, p. 203.

UN General Assembly, ***Resolution 68/167 (2013): Right to Privacy in the Digital Age***, 21 18 December 2013, UN Docs A/RES/68/167 (2013). See also: UN General Assembly, *Resolution 73/179 (2018): The right to privacy in the digital age*, 17 December 2018, UN Docs A/RES/73/179 (2018).

UN General Assembly, *Resolution 217 A (III): Universal Declaration of Human Rights,* 10 December 1948, **UN Docs A/RES/217 A (III).**

USA, ***H.R. 2231: Algorithmic Accountability Act of 2019*** (4 November 2019, *Congress*), <https://www.congress.gov/bill/116th-congress/house-bill/2231/text> accessed 19 August 2021.

UK, Information Commissioner's Office, '**The Office of the Australian Information Commissioner and the UK's Information Commissioner's Office open joint investigation into Clearview AI Inc.** (9 July 2020, *Information Commissioner's Office*), <https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2020/07/oaic-and-ico-open-joint-investigation-into-clearview-ai-inc/> accessed 19 August 2021.

US California, ***California Consumer Privacy Act of 2018*** [1798.100 – 1798.188.100] (*California Legislative Information*), <https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5> accessed 24 August 2021.

With reference to the protection from racial discrimination as raising *ergaomnes* obligations: International Court of Justice, *Barcelona Traction, Light and Power Company, Limited*, Judgment, 5 February 1970, [1970] ICJ Reports, p. 3, para. 34.

With reference to peremptory character of the principle of non-discrimination and the principle of equality before the law: Inter-American Court of Human Rights, *Juridical Condition and Rights of Undocumented Migrants*, Advisory Opinion, 17 September 2003, OC-18/03 2003, para. 101.

World Conference Against Racism, Racial Discrimination, Xenophobia and Related Intolerance, *Declaration and Programme of Action,* 2002, para. 72. See also: **Committee on the Elimination of Racial Discrimination, *General recommendation No. 36*,** para. 13.

World Economic Forum, Global Agenda Council on Values (2014-2016), '**Values and the Fourth Industrial Revolution. Connecting the Dots Between Value, Balues, Profit and Purpose,'** Geneva, September 2016 (14 November 2016, *World Economic Forum*)<https://www.weforum.org/whitepapers/values-and-the-fourth-industrial-revolution-connecting-the-dots-between-value-values-profit-and-purpose/> accessed 6 August 2021. See also: H. Sutcliffe, A.-M. Allgrove,

Recebido em 29/08/2021
Aprovado em 30/08/2021
Received in 08/29/2021
Approved in 08/30/2021