UNIVERSITY OF ESSEX DISSERTATION

SCHOOL OF LAW

LLM/MA IN: INTERNATIONAL HUMAN RIGHTS LAW
STUDENT’S NAME: ALEXANDRA ZIAKA
SUPERVISORS’S NAME: SHELDON LEADER
DISSERTATION TITLE Businesses’ corporate responsibility for a human rights-centred Artificial Intelligence --- --- ---

COMMENTS: (PLEASE WRITE BELOW YOUR COMMENTS)

MARK:	
SIGNATURE:	DATE:

UNIVERSITY OF ESSEX

SCHOOL OF LAW

LLM on International Human Rights Law

2018-2019

Supervisor: Sheldon Leader

DISSERTATION

Businesses' corporate responsibility for a human rights-centred Artificial Intelligence

Name: Alexandra Ziaka

Registration Number (optional): 1807647

Number of Words: 19,712

Date Submitted: 20-09-19

TABLE OF CONTENTS

List of abbreviations

AI	Artificial Intelligence
AIA	Algorithmic Impact Assessment
CoE	Council of Europe
DPIA	Data Protection Impact Assessment
EU	European Union
GDPR	General Data Protection Regulation
GNI	Global Network Initiative
HRBDT	Human Rights, Big Data and Technology Project of the University of Essex
HRDD	Human Rights Due Diligence
HRESIA	Human Rights, Ethical and Social Impact Assessment
HRIA	Human Rights Impact Assessment
OECD	Organisation of Economic Co-operation and Development
OHCHR	Office of the High Commissioner for Human Rights
PIA	Privacy Impact Assessment
Special Rapporteur on Freedom of Expression	UN Special Rapporteur on the Promotion and protection of the right to freedom of opinion and expression
UN	United Nations
UNGPs	UN Guiding Principles on Business and Human Rights
US	United States of America

I. INTRODUCTION

Artificial intelligence (hereinafter ‘AI’) is the new hype of our era. Our newsfeed on Facebook, the famous voice assistant of Amazon, Alexa, the spam filter in our emails, smart cities, self-driving cars, search engines, smart clothes, credit scoring algorithms, they all have something in common and this is AI.

There is not one single form of AI but all AI systems have three elements in common: sensors, operational logic and actuators.[1] According to the Independent High-Level Expert Group on AI which has been mandated by the European Commission and sought to come up with a definition of AI, “*AI systems are software (and possibly also hardware) systems designed by human that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal”.*[2] The most known type of AI today is machine learning. Machine learning enables a system to learn from examples, data and experiences.[3] More specifically, the algorithms of machine learning systems are given a certain task to do, and in parallel they are fed with loads of data, from which they can create patterns after making correlations.[4] The algorithm[5] constantly learns through this process and improves. Although we will use the term ‘AI’ in an inclusive way along the dissertation, most of the examples refer to machine learning systems.

AI has certainly benefited humans in many ways. In the health sector, for instance, AI is currently used to diagnose diseases and help doctors give the appropriate prescription to patients. Also, AI has contributed to the development of agriculture as it can help farmers choose the appropriate crops and

[1] OECD, *Artificial Intelligence in Society* (OECD Publishing 2019) 22 <https://www.oecd-ilibrary.org/docserver/eedfee77-en.pdf?expires=1568930907&id=id&accname=oid051805&checksum=B82141F6397F76C33F77524BFB3F4F0E> accessed 20 September 2019.

[2] Independent High-Level Expert Group on Artificial Intelligence, ‘A Definition of AI: Main Capabilities and Disciplines’ (April 2019) 6 <https://www.aepd.es/media/docs/ai-definition.pdf> accessed 20 September 2019.

[3] The Royal Society, ‘Machine learning: the power and promise of computers that learn by example (April 2017) 16 <https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf> accessed 20 September 2019.

[4] ibid 19.

[5] In the context of this dissertation, algorithms are ‘*encoded procedures for transforming input data into the desired output, based on specific calculations*’ from Tarleton Gillespie, Pablo J. Boczkowski and Kirsten A. Foot, *The relevance of algorithms, Media technologies: Essays on communication, materiality, and society* (MIT Press 2014) 167.

adapt to new environments.[6] However, the attention of the legal literature mainly focuses on the negative effects. Nemitz compares AI to nuclear power. He says that it took many tragedies to understand the real impact of nuclear power, and our society cannot afford to do the same with AI, due to the potentially irreversible repercussions.[7]

The nearly incontrollable pace of technological developments has led to a shift of power balances.[8] Many public services are now completed through the use of AI systems.[9] Businesses working on the design, development, operation and sale of AI systems, engage into activities that traditionally belong to the state, eg. design of algorithms for resource allocation or prediction of criminal recidivism and this way they have gained a lot of power. The term "tech companies", which is used throughout the dissertation, is considered to include all the above-mentioned businesses.

While tech companies certainly do not have an *obligation* to respect, protect and fulfil human rights, their *responsibility* to respect human rights remains. The issue of human rights-centred AI becomes more urgent especially in States which lack strong human rights legal framework. In that case, the irresponsible use of AI can exacerbate the vulnerabilities of people and expose them in violations of greater scale[10] without high prospect of a remedy.

However, AI which puts human rights at the centre may conflict with the commercial interests of tech companies. The aim of this dissertation is to identify the special characteristics of AI which make the fulfilment of tech companies' responsibility to respect human rights very challenging and further recommend ways with which tech companies can take proactive measures to ensure that the AI puts human rights at the centre of its design. More particularly, it argues that tech companies have a responsibility to design AI in a way that it incorporates human rights law principles and standards.

[6] Access Now, 'Human Rights in the Age of Artificial Intelligence' (November 2018) 14 <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf> accessed 20 September 2019.

[7] Paul Nemitz, 'Profiling the european citizen: why today's democracy needs to look harder at the negative potential of new technology than at its positive potential' in Emre BayamlioğLu, Irina Baraliuc, Liisa Janssens and Mireille Hildebrandt (Eds), *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen* (AUP 2018).

[8] Lorna McGregor, Vivian Ng, Ahmed Shaheed, Elena Abrusci, Catherine Kent, Daragh Murray and Carmel Williams, 'The Universal Declaration of Human Rights at 70 - Putting Human Rights at the heart of the Design, Development, and Deployment of Artificial Intelligence' (*Human Rights Big Data and Technology – University of Essex*, December 2018) 42 <https://48ba3m4eh2bf2sksp43rq8kk-wpengine.netdna-ssl.com/wp-content/uploads/2018/12/UDHR70_AI.pdf> accessed 20 September 2019.

[9] Access Now (n 6) 14.

[10] Dunstan Allison-Hope and Mark Hodge, 'Artificial Intelligence: A Rights-Based Blueprint for Business-Paper 1: Why a Rights-Based Approach?' (*Business Social Responsibility*, August 2018) 10 <https://www.bsr.org/reports/BSR-Artificial-Intelligence-A-Rights-Based-Blueprint-for-Business-Paper-01.pdf> accessed 20 September 2019.

Chapter II will set the tone of the discussion by referring to the special characteristics of AI and its implications on the right to privacy, freedom of expression and non-discrimination. Although there is a wide range of human rights that can be interfered with by AI, we will delve into the implications on these three rights not only because of space limitation, but also because these rights are most often restricted and their violation leads to the violation of other rights too.

Chapter III will analyse the challenges that tech companies face in the context of human rights due diligence (hereinafter 'HRDD'), which is a fundamental component of corporate responsibility to respect human rights. We will explore whether elements of other types of impact assessment can inform the human rights impact assessment (hereinafter 'HRIA') in the context of AI and we will identify ways with which tech companies can live up to the expectations of a proactive HRDD.

Chapter IV will shift the focus on the human rights by design, which entails that human rights considerations should be incorporated in the design of services and products. The importance of infusing human rights concerns already in the design of AI systems will be on the spotlight. It will be shown that the role of design is fundamental in guaranteeing respect for human rights, as it can work as a form of regulation and thus prohibit risks since the beginning of AI's lifecycle. We will then explore the challenges of ensuring transparency, which is a subset of human rights by design, through designing explainable algorithms. Although explainability is important, it entails a lot of trade-offs that can work as a disincentive for tech companies. This dissertation argues that qualified transparency could serve as a meeting point for the two competing interests - human rights interests and commercial interests.

II. The particular challenges of AI on human rights

The aim of this chapter is to set the colours for the upcoming discussion. The chapter will first go through the characteristics of AI which distinguish it from other types of technology, while the second part will examine how AI impacts the right to privacy, freedom of expression and non-discrimination. These characteristics of AI and the implications on human rights will serve as an indication of why businesses need to be extra cautious when conducting their HRDD and discharging their corporate responsibility to respect human rights.

A. Special characteristics distinguishing AI from other technologies

AI has some unique characteristics that distinguish it from other types of technologies. Below we will discuss what makes AI so particular and why it requires special treatment.

One of the main characteristics of AI is that the accuracy of its predictions depends on the availability of large datasets.[11] One of the main data protection principles is the data minimisation, according to which the personal data shall be "*adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed"*.[12] Thus, it seems that the data minimisation principle cannot really apply in the context of AI, since the more the data, the more effective the algorithm will be.

The efficiency of AI depends not only on the availability of large datasets, but also on the accuracy of those data. Thus, the quality of the training data, ie the historic data with which the algorithm is trained, is crucial for the inferences produced by the algorithm.[13] Data which are inaccurate, incomplete or irrelevant can have a detrimental effect on the efficient operation of the algorithm.[14]

The operation of machine learning systems is based on the concept of correlation and statistical probability.[15] As mentioned earlier, machine learning systems are trained on large data sets and they

[11] Dimitra Kamarinou, Christopher Millard and Jatinder Singh, 'Machine learning with personal data' in Ronald Leenes, Rosamunde van Brakel, Serge Gutwirth and Paul De Hert (eds), *Data Protection and Privacy- The Age of Intelligent Machines,* vol. 10 (Hart Publishing 2017) 100.

[12] European Parliament and Council Regulation 2016/679/EU 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 [2016] OJ L119/1 ('GDPR'), art 5§1(c).

[13] Aaron Rieke, Miranda Bogen, David G. Robinson, 'Public Scrutiny of Automated Decisions: Early Lessons and Emerging Decisions' (*Omidyar Network and Upturn*, February 2018) 12 https://www.omidyar.com/sites/default/files/file_archive/Public%20Scrutiny%20of%20Automated%20Decisions.pdf accessed 20 September 2019.

[14] ibid.

[15] Lorna McGregor, Daragh Murray And Vivian Ng, 'International Human Rights Law as A Framework For Algorithmic

tend to identify patterns and make correlations between the variables in a dataset and draw inferences.[16] The correlations reflect a relation between the data but not a causation.[17] Based on past behaviours, the system makes a prediction assuming that there will be no change in the future.[18] What is more, these correlations are based on data derived from group behaviour but they ultimately determine the decision about an individual.[19] This is problematic as individuals are not evaluated on grounds of their merit, but on the basis of their membership in a certain group.[20]

AI systems are known for their dynamic nature and the fact that they are not easily controlled. This is obvious especially in the case of machine learning systems, which gradually learn through the process, as they use their inferences as new input data.[21] This leads to continuous changes of the decisional rule.[22]

Another common characteristic of AI is opacity. Frank Pasquale has compared algorithms to "black boxes".[23] He has emphasised how companies can scrutinise every detail of our life, yet they manage to get away from scrutiny through using opaque algorithms.[24] Consequently, individuals who are affected by the AI's prediction, e.g. those who were not given a loan, may not be able to understand the reasons behind that decision. Sometimes opacity does not allow even the computer scientists who designed and developed the algorithms, to understand their logic.[25]

AI is also characterised by discreteness in the sense that the different parts of the AI system may be designed by different people in different places without coordination.[26] Additionally, matters get more complex, if we think that an algorithm may be developed by a company and be trained on certain training

Accountability' (2019) 68 International and Comparative Law Quarterly 309, 316.

[16] David Lehr and Paul Ohm, 'Playing with the Data: What Legal Scholars Should Learn about Machine Learning' (2017) 51 University of California Davis Law Review 653, 671.

[17]Mireille Hildebrandt, 'Defining Profiling: A New Type of Knowledge?' in Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen:* 17 *Cross-Disciplinary Perspectives (Springer 2008)18.*

[18] Cathy O'neil, Weapons of Math Destruction: How Big Data increases inequality and threatens democracy (Penguin Random House UK 2016) 155.

[19] McGregor, Murray and Ng (n 15) 316.

[20] Eyal Benvenisti, 'Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance?' (2018) 19 European Journal of International Law 9, 60.

[21] McGregor, Murray and Ng (n 15) 310.

[22] Kamarinou, Millard and Singh (n 11) 110.

[23] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015)

[24] ibid 9.

[25] Will Knight, 'The Dark Secret at the heart of AI' (2017) 120 MIT Technology Review 55, 57.

[26] Matthew U. Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016(29 Harvard Journal of Law & Technology 353, 369.

data, but then it may be sold to a third party which will feed the algorithm with different input data.[27] This characteristic makes the attribution of harm very challenging since it is difficult to identify at which point something got wrong.

All the above-mentioned characteristics, jointly or separately, are responsible for AI's impact on human rights, which will be discussed below.

B. Implications of AI on human rights

AI applications impact many human rights, like privacy, freedom of expression, non-discrimination, right to health, right to employment, freedom of assembly, and freedom of association.[28] This subchapter will solely focus on the impact of AI on privacy, freedom of expression and non-discrimination.

1. Privacy

The right to privacy translates into the right to respect for private life.[29] Warren and Brandeis first defined privacy as the right to be left alone,[30] as early as in 1890.[31] Even though since then the notion has been significantly enlarged, the right to be left alone still maintains its principal role. Solove found that privacy has been classified in six categories as: 1) the right to be left alone, as described by Warren and Brandeis 2) limited access to oneself from access by others, 3) secrecy, 4) control over personal information, which is linked with the protection of personal data, 5) personhood which has to do with the protection of individuality and dignity and 6) intimacy, which ensures that a person has control over his/her intimate relationships of his/her life.[32] Nowadays there is a tendency to conflate the right to

[27] McGregor, Murray and Ng (n 15) 318.

[28] McGregor, Ng, Shaheed, Abrusci, Kent, Murray and Williams (n 8)10.

[29] Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) ('UDHR'), art 12; International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 ('ICCPR'), art 17; Convention for the Protection of Human Rights and Fundamental Freedoms (adopted 4.11.1950, entered into force 03.09.1953) ETS 005 ('ECHR'), art 8; Charter of Fundamental Rights of the European Union [2012] OJ C 326/391 ('EU Charter'), art 7; American Convention on Human Rights (adopted 22.11.1969, entered into force 18.07.1978) OAS Treaty Series No 36 ('ACHR'), art 11; African Charter on Human and Peoples' Rights (adopted 27.06.1981, entered into force 21.10.1986) 21 ILM 58 ('African Charter'), art 6.

[30] Warren, Brandeis, 'The Right to Privacy' (1890) 4 Harvard Law Review 193.

[31] Pieter Kleve and Richard De Mulder, 'Privacy protection and the right to information: in search of a new symbiosis in the information age' in Sylvia Mercado Kierkegaard (ed), Cyberlaw, Security & Privacy (Ankara Bar Association Press 2007) 338

[32] Daniel J. Solove, *Understanding Privacy* (Harvard University Press 2008) 12-13.

privacy with the right to protection of personal data, however it is important to stress that they are not identical,[33] as data protection is narrower.[34]

Privacy serves as a 'gatekeeper' for other rights.[35] In other words, the violation of privacy can lead to the violation of other rights, such as freedom of expression, non-discrimination, right to political participation, right to employment and right to health. The collection of massive amounts of data can lead to the creation of profiles, which are further used for the operation of AI. For example, predictive algorithms collect data coming from online and offline activities of individuals and make inferences about them.[36] Based on those findings, public and private actors take important decisions that have an impact on them.[37] For example, the posts of somebody on social media can negatively influence the prediction of an algorithm used for recruitment.

The business model of many companies depends on data exploitation, in the sense that the more data they collect and further sell, the more profit they have. These models raise a lot of privacy issues, as very often the data collection takes place without the individuals' consent or data are collected for a certain purpose and are subsequently used for different purposes. AI systems overcome the barrier of consent by collecting mainly non-personal data which do not fall under the scope of data protection laws. Although at first glance the collection of non-personal data seems innocent, non-personal data when combined, can ultimately lead to re-identification of personal data and even worse, sensitive data, which require a more sophisticated treatment.[38] Thus, even if AI systems gather pieces of allegedly anonymised information, when all these pieces come together they can breed new data,[39] which can

[33] Monika Zalnieriute, 'An International constitutional moment of data privacy in the times of mass- surveillance' (2015) 23 International Journal of Law and Information Technology 99, 104.

[34] HRC, 'General Comment No. 16: Article 17 (Right to Privacy) The Right to Respect of Privacy, Family, Home and Correspondence, and Protection of Honour and Reputation' (8 April 1988) UN Doc CCPR/C/21/Add.6, para 10.

[35] McGregor, Ng, Shaheed, Abrusci, Kent, Murray and Williams (n 8) 14.

[36] Danielle Keats Citron and Frank Pasquale, 'The Scored Society: Due Process for automated predictions' (2014) 89 Washington Law Review 1, 3.

[37] Ibid.

[38] According to art 7 GDPR, special categories of data include "*personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation";* Article 19 and Privacy International, 'Privacy and Freedom of Expression In the Age of Artificial Intelligence' (April 2018) 27 https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf accessed 20 September 2019.

[39] Committee of Experts on Internet Intermediaries of Council of Europe, 'Algorithms and Human Rights - Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications' DGI(2017)12 (*Council of Europe*, March 2018) 13 https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5 accessed 20 September 2019.

give an insight into a person's lifestyle, preferences and even his/her personal thoughts.[40] The most alarming issue is that individuals do not realise that their data are collected and accessed by different actors.[41] This is the so-called informational asymmetry between the information that companies collect and what the individual perceives.[42]

However, there is another case in which the right to privacy can be interfered with, without involving the processing of one's own data. It is what the Human Rights, Big Data and Technology Project of the University of Essex (hereinafter 'HRBDT') has called the "tyranny of minority".[43] Tyranny of minority refers to cases in which particular individuals have not given their consent for a processing of their data, but the inferences from the processing of other people's data ultimately have an impact on them. When individuals have not allowed for their data to be processed by AI systems, the outputs of the algorithm influence them based on data of other individuals.

2. Freedom of expression

AI can also impact the right to freedom of expression, which encompasses the freedom to seek, receive and impart information and ideas of all kinds.[44] Although it is an individual right, it also has a collective dimension.[45] More specifically, freedom of expression also consists of the right to hear the views of others, exchange ideas with others and the right to be informed.[46]

Freedom of expression is interlinked with the right to privacy.[47] Privacy is the precondition for the effective enjoyment of the freedom of expression and freedom of expression is the means with which individuals can self-develop.[48] The pervasive character of AI which is linked with the way that data are

[40] McGregor, Ng, Shaheed, Abrusci, Kent, Murray and Williams (n 8) 14.
[41] Council of Europe Commissioner for Human Rights, Human Rights Comment: Safeguarding human rights in the era of artificial intelligence (3 July 2018) https://www.coe.int/en/web/commissioner/-/safeguarding-human-rights-in-the-era-of-artificial-intelligence accessed 20 September 2019.
[42] Article 19 and Privacy International (n 38) 18.
[43] HRBDT, 'Background Paper on Consent Online' (June 2019) 9 https://hrbdt.ac.uk/wp-content/uploads/2019/06/19.06.09-Background-Paper-on-Consent-Online.pdf accessed 20 September 2019.
[44] ICCPR, art 19§2; ECHR, art 10; EU Charter, art 11; ACHR, art 13; African Charter, art 9.
[45] Dominic McGoldrick, 'Thought, Expression, Association, and Assembly' in Daniel Moeckli, Sangeeta Shah and Sandesh Sivakumaran (eds), International Human Rights Law (3rd edition, Oxford University Press 2018) 217.
[46] ibid 218.
[47] Article 19 and Privacy International (n 38) 5.
[48] ibid.

massively collected, creates a sense that we are constantly being watched.[49] How many times have we discussed orally about something with a friend, and the next day there is an advertisement about it on our Facebook newsfeed. Besides that, personal information of users can be used for many other reasons which can have significant effects on the individual. For example, an algorithm used for recruitment may take into account the browsing history of a candidate or his/her posts on Facebook and eliminate him/her. All these can seriously affect the freedom of expression, as they may lead people to censor themselves for fear that whatever they do, will have consequences on multiple aspects of their lives.

AI is also used for content moderation which can have a chilling effect on freedom of expression. For example, search algorithms define the top results of our search on Google.[50] The fact that each user may eventually have access to different kind of information which is tailored to his/her preferences, can lead to the creation of echo chambers.[51] In other words, people are exposed to ideas that are compatible with their own ideas, without having the opportunity to be exposed to opposing views. This situation can seriously interfere with the right to receive information, as a subset of the right to freedom of expression.

Disinformation constitutes one of the most alarming issues nowadays. AI systems can be used to spread false information with a view to disorientate the public .[52] They are used by political parties which try to sabotage their opponents or from governments which aim to manipulate the public opinion.

3. Non-discrimination

Non-discrimination and its counterpart, equality, constitute a general principle of human rights

[49] Martjin van Otterlo, 'A machine learning view on profiling' in Mireille Hildebrandt and Katja de Vries (eds), *Privacy, Due Process and the Computational Turn* (Routledge 2013) 41.
[50] Access Now (n 6) 23.
[51] Katja de Vries and Mireiller Hildebrandt, 'Introduction: Privacy, due process and the computational turn at a glance' in Mireille Hildebrandt and Katja de Vries (eds), *Privacy, Due Process and the Computational Turn* (Routledge 2013) 1.
[52] Access Now (n 6) 16.

protection.[53] The Human Rights Committee has established that discrimination must be understood as *"distinction, exclusion, restriction or preference which is based on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status, and which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise by all persons, on an equal footing, of all rights and freedoms*".[54] The way in which AI systems work has shown that they disproportionately affect vulnerable groups such as women, specific racial, ethnic or religious groups, disabled people and LGBTQ.[55] Importantly, non-discrimination, like privacy, also serves as a gatekeeper for other rights, such as the right to employment and the right to health.[56]

Even when a decision-making process depends partially and not fully on the inference of the AI system, in cases for example when a bank has to decide whether an individual qualifies for a loan, and consults a credit scoring algorithm, there is a risk that what is supposed to be just an advice, eventually determines the result.[57] There is an assumption that the algorithm is objective, and people tend to depend on it at a large extent, due to lack of time or skills.[58]

Criminal justice is one of the most problematic areas. AI is used to predict criminal recidivism, ie the probability of committing a crime in the future. These systems are used to consult judges in their decisions about sentencing.[59] ProPublica conducted a study which revealed that a machine learning system which made predictions about criminal recidivism in the US was biased against black people, as it wrongly flagged them at almost twice the percentage of white defendants.[60]

[53] HRC, 'General Comment 18' (10 November 1989) UN Doc HRI/GEN/1/Rev.9 (Vol. I), §1; UN Vienna Declaration and Program of Action (1993) UN Doc A/CONF.157/23, para 15.

[54] ibid para 7

[55] Access Now (n 6) 18.

[56] McGregor, Ng, Shaheed, Abrusci, Kent, Murray and Williams (n 8) 11.

[57] Frederik Zuiderveen Borgesius, 'Discrimination, artificial intelligence, and algorithmic decision-making' (*Council of Europe*, 2018) 8 https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 accessed 20 September 2019.

[58] Committee of Experts on Internet Intermediaries of Council of Europe, 'Algorithms and Human Rights' (n 39) 8

[59] Access Now (n 6) 15.

[60] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, 'Machine Bias' (*ProPublica*, 23 May 2016) https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing accessed 20 September 2019.

AI can lead to a discriminatory treatment when it is trained on historic biased data.[61] In that case the algorithm reproduces the already existing discrimination. This can be illustrated with an example on university admissions.[62] For example, if a university used to prefer male students instead of women and immigrants and the algorithm for the selection of ideal candidates is trained on these historic biased data, it will infer that the ideal output is a male student.

Even if an AI system is not trained on historic biased data and there is no attribute which relates to one of the protected grounds under the non-discrimination principle, it may be based on proxies which eventually lead to bias.[63] For example, postal codes can work as proxies for low income or race. A credit scoring algorithm which takes postal code into account may disqualify a candidate who lives in a poor neighbourhood, because in this neighbourhood there is a high percentage of failure to pay off loans.

A flawed AI system can create a continuously discriminating environment for individuals and restrict their life opportunities.[64] For example, the connections of a person on social media can be the reason for not being accepted to a job. The lack of job can lead to impoverishment and create the need for a loan, which based on these facts, may not be given.

It follows from the above that AI can have significantly negative impact on the right to non-discrimination.

To summarise, this chapter introduced the special characteristics of AI and the implications on human rights by its use. What distinguishes AI from other technologies is its dependency on massive amounts of data, its dynamic process which leads to constant changes, its opaque nature and its discreteness. The right to privacy is often impacted by the use of AI, since very often data are collected and used by

[61] Solon Barocas and Andrew D. Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671, 680.

[62] Borgesius (n 57) 11

[63] ibid 13.

[64] Citron and Pasquale, 'The Scored Society' (n 36) 33.

AI systems without the consent of individuals and even non-personal data can easily lead to the identification of personal data. AI can also have a chilling effect on freedom of expression, as the feeling of ongoing surveillance leads to self-censorship. Also, content moderation can lead to the creation of echo chambers which further creates a bubble where individuals are not exposed to different ideas. Finally, AI can lead to discriminatory treatment, as it may discriminate against vulnerable groups of people and can exacerbate already existing bias.

III. CORPORATE RESPONSIBILITY TO RESPECT HUMAN RIGHTS

Many years have passed since Milton Friedman stated that the only social responsibility of the companies is to increase their profits on condition that they do not violate the competition rules with deception or fraud.[65] Now companies have a responsibility to respect human rights, but this responsibility does not equal to States' obligation to respect, protect and fulfil human rights. Thus, companies still do not have direct obligations under international human rights law.[66] The choice of "responsibility" instead of "obligation" or "duty" reflects the pragmatic approach that Ruggie adopted while drafting the UN Guiding Principles on Business and Human Rights (hereinafter 'UNGPs),[67] which still constitute the most authoritative framework[68] regarding the companies' responsibility to respect human rights. The former Special Rapporteur wanted to find a way to make companies more respectful of human rights but without setting very high standards that states would not accept and companies would not embrace. Although UNGPs' soft law nature renders their enforcement more challenging than if they constituted hard law, their adoption has been considered a great achievement.[69] Their moderate character serves as a foundation towards the establishment of direct human rights obligations of companies under international human rights law.[70]

This chapter will review the content of this corporate responsibility to respect human rights with a special focus on HRDD, as it is considered to be its most central component.[71] It will explain why HRDD gets more challenging when it comes to tech companies which design, develop and deploy AI, and will provide recommendations on how HRDD could be shaped so that it sufficiently addresses the challenges of AI.

[65] Milton Friedman, 'The Social Responsibility of Business is to increase its profits' *New York Times* (New York, 13 September 1970) <https://graphics8.nytimes.com/packages/pdf/business/miltonfriedman1970.pdf> accessed 20 September 2019.

[66] Robert McCorquodale, 'International Human Rights Law Perspectives' in Lara Blecher, Nancy Kaymar Stafford and Gretchen C. Bellamy (eds), Corporate Responsibility for Human Rights Impacts: New Expectations and Paradigms (American Bar Association 2014) 64.

[67] Human Rights Council, 'UN Guiding Principles on Business and Human Rights' (2011) UN Doc A/HRC/17/31 ('UNGPs'); Bjorn Fasterling and Geert Demuijnck, 'Human Rights in the Void? Due Diligence in the UN Guiding Principles on Business and Human Rights' (2013) 116 Journal of Business Ethics 799, 800.

[68] Marco Fasciglione, 'The enforcement of corporate Human Rights Due Diligence' (2016) 10 Human Rights & International Legal Discourse 94, 98.

[69] Nadia Bernaz, *Business and Human Rights: History, law and policy- Bridging the accountability gap* (Routledge 2017) 195.

[70] John Gerard Ruggie, *Just Business – Multinational Corporations and Human Rights* (W.W. Norton & Company 2013) 124.

[71] Fasciglione (n 66) 104.

A. Human Rights Due Diligence

The UNGPs are divided into three pillars. Pillar I is dedicated to the States' duty to protect human rights from interferences of third parties. Pillar II focuses on the corporate responsibility to respect human rights while Pillar III calls both States and companies to ensure that victims of corporate-related violations have access to remedies.

The corporate responsibility to respect human rights requires that companies avoid causing or contributing to adverse human rights impacts through their business activities and seek to prevent or mitigate adverse human rights impacts that are directly linked to their business activities by their business relationships.[72] In order to discharge their responsibility to respect human rights, companies are expected to adopt a human rights policy, conduct a HRDD and provide procedures for effective remediation.[73] This subchapter will set out the fundamentals of HRDD which will serve as basis for the next subchapters which will focus on the special case of HRDD in the context of AI.

HRDD is defined as *"an ongoing management process that a reasonable and prudent enterprise needs to undertake, in the light of circumstances (including sector, operating context, size and similar factors) to meet its responsibility to respect human rights"*.[74] HRDD is a process with which companies identify, prevent, mitigate and account for how they address the adverse human rights impacts generated by their business activities.[75] The human rights impacts include those that the company has caused, contributed to or are directly linked to its business relationships.[76] The company has four different missions to accomplish: a company should (1) assess the actual and potential human rights risks, (2) integrate its findings in its business activities and take measures to mitigate those risks, (3) track its performance and (4) communicate the risks and the results of its conduct.

1. Characteristics of Human Rights Due Diligence

HRDD should accumulate a range of characteristics in order to comply with UNGPs. First, it should be proactive. The 'Protect, Respect and Remedy Framework' states that HRDD is considered to be *"a*

[72] UNGPs, Principle 13.
[73] UNGPs, Principle 15.
[74] OHCHR, 'The Corporate responsibility to respect human rights: An Interpretative Guide' (2012), 6.
[75] UNGPs, Principle 17.
[76] UNGPs Principle 17(a).

comprehensive, proactive attempt to uncover human rights risks, actual and potential".[77] Companies should not wait until a human rights risk materialises in order to take retroactive measures. Thus, HRDD serves as a set of prophylactic measures that the companies should deploy early in order to prevent any adverse human rights risks.[78]

HRDD is context-specific. The UNGPs appear to be quite flexible with regard to the scale and complexity of the measures taken in the context of HRDD. The size of the company, the risk of severe human rights impacts and the nature as well as the context of the company's operations are the factors which will ultimately determine the scale of HRDD.[79] The context is subject to change, and thus the scope of HRDD will vary according to the circumstances.[80]

Connected to that, HRDD should run in an ongoing basis.[81] The fact that the actual or potential human rights impacts have been assessed at the beginning of the product's or service's lifecycle does not mean that they will remain the same. Therefore, companies have to repeat their assessment in regular intervals.

2. The different stages of Human Rights Due Diligence

2.1 Assessment through human rights impact assessment

The first stage of HRDD is the assessment of the actual and potential human rights risks, which is realised through a HRIA. Through HRIA the company can identify the underlying human rights risks so that it addresses them in advance.[82] Thus, its value is critical as it is the precondition for the other components of HRDD.

HRIA must take place before any critical stages in the company's business activities - prior to a new activity or business relationship, important decisions and changes affecting its operation, such as

[77] UNHRC, 'Business and Human Rights: Towards Operationalizing the 'Protect, Respect and Remedy' Framework (22 April 2009) UN Doc A/HRC/11/13, para 71.

[78] Björn Fasterling, 'Human Rights Due Diligence as Risk Management: Social Risk Versus Human Rights Risk' (2017) 2 Business and Human Rights Journal 225, 228.

[79] UNGPs, Principle 17(b).

[80] Robert Mccorquodale, Lise Smit, Stuart Neely and Robin Brooks, 'Human Rights Due Diligence in Law and Practice: Good Practices and Challenges for Business Enterprises' (2017) 2 Business and Human Rights Journal 195, 199.

[81] Anita Ramasastry, 'Corporate Social Responsibility Versus Business and Human Rights: Bridging the Gap Between Responsibility and Accountability' (2015) 14 Journal of Human Rights 237, 247.

[82] Mccorquodale, Smit, Neely and Brooks, 'Human Rights Due Diligence in Law and Practice' (n 80) 205.

change in the policy framework or social tensions and during the lifecycle of the activity or relationship.[83] It identifies the risks of a certain business activity as well as the affected stakeholders, it enlists the relevant human rights standards and projects how the specific activity or relationship will lead to the adverse human rights impacts on these groups.[84] The most challenging part though is that the company needs to involve the affected stakeholders and other relevant stakeholders. There is a wide range of stakeholder groups that the company should consider in order to assess whether its business activities could have an impact on them, such as the company employees, the supply chain workers, consumers and users and vulnerable or marginalised groups.[85] The company is expected to conduct a meaningful consultation with them.[86] The company should also engage with relevant stakeholders, such as civil society, companies of the same sector in order to receive their insight on the assessment of human rights impacts.

HRDD also includes the identification and assessment of the adverse human rights impacts which could potentially be caused by entities which have a business relationship with the company.[87] However, these risks need to be relevant to the service or product of the company.[88] The UNGPs have recognised that when companies have a big value chain, it may seem unreasonable and unrealistic to ensure[89] that the HRDD covers all of the entities included. In order to tackle this, the company must identify the general problematic area, the operating context of the value chain which creates concerns and prioritise the emerging human rights impacts.

2.2 Integrating

After HRIA, the business needs to integrate its findings in its business operations and identify ways to prevent and mitigate the identified risks.[90] If the company causes or may cause a human rights impact, then it is expected to take appropriate measures to prevent or discontinue it.[91] In case where the company contributes or its activities are directly linked to the adverse human rights impacts, then the

[83] UNGPs, Principle 18.

[84] UNGPs, Principle 18.

[85] European Commission, 'ICT Sector Guide on Implementing the UN Guiding Principles on Business and Human Rights' (June 2013) 12 <https://www.ihrb.org/pdf/eu-sector-guidance/EC-Guides/ICT/EC-Guide_ICT.pdf> accessed 20 September 2019.

[86] UNGPs, Principle 18.

[87]ibid.

[88] OHCHR, 'An Interpretative Guide' (n 74) 39.

[89] UNGPs, Principle 17.

[90] UNGPs, Principle 19.

[91] UNGPs, Principle 19.

leverage of the company comes into play, which, according to the interpretative guide of UNGPs, is *“the ability to effect change in the wrongful practices of the party that is causing or contributing to the impact”*.[92] According to the UNGPs, if the company contributes or may contribute to the human rights impact, it should prevent or discontinue the impact, but given that the business is not the only actor that contributes to the human rights impact, it should use its leverage to mitigate the human rights impact.[93] If its operations, products or services are directly linked to the adverse human rights impacts, then again it has to exercise its leverage to mitigate the impacts.[94] However, the level of its involvement will depend on the degree of leverage that the company has on the other entity which causes the impact, the importance of this business relationship to the company, the severity of the abuse and whether the discontinuation of the relationship would pose additional human rights risks.[95]

2.3 Tracking

According to Principle 20 of the UNGPs, a company needs to track its response to the adverse human rights impacts. As discussed, HRDD is not a one-off process. A company can track its responses through the use of qualitative and quantitative indicators as well as through feedback from internal and external sources, including the affected stakeholders.[96] Qualitative and quantitative indicators can help the company realise its progress in dealing with human rights impacts and identify problematic areas. Indeed, there are indicators and benchmarks which have proved to be very helpful for tracking human rights progress, such as the Standards of the Global Reporting Initiative,[97] the Human rights Compliance Assessment tool of Danish Institute for Human Rights[98] and the Corporate Human Rights Benchmark.[99] Likewise, the feedback of internal and external sources can benefit the tracking process. Employees can give their feedback internally as long as there is no risk of reprisal, should an employee point something wrong in the services or products of the company.[100] Additionally, the conduct of audit can track the company’s response.[101] Also, the operational-level grievance mechanisms and the

[92] OHCHR, ‘An Interpretative Guide’ (n 74) 48.
[93] UNGPs, Principle 19.
[94] ibid.
[95] ibid.
[96] ibid Principle 20.
[97] Global Reporting Initiative, ‘GRI 102: General Disclosures’ (2016).
[98] Danish Institute of Human Rights, ‘Human Rights Indicators for Business platform’ (2016) <https://www.business-humanrights.org/sites/default/files/HRCA_INTRODUCTION.pdf> accessed 20 September 2019.
[99] Corporate Human Rights Benchmark <https://www.corporatebenchmark.org/> accessed 20 September 2019.
[100] OHCHR, ‘An Interpretative Guide’ (n 74) 54.
[101] UNGPs, Principle 20.

affected stakeholders are a precious source of feedback, since they are the ones who are directly affected by the actual or potential human rights impacts. The company can consider the type of complaints and identify patterns in them, and use these data to correct the flaws and adapt its operations.[102]

2.4 Communicating

One of the components of HRDD is communication. It is the means with which the company gives information on how it addresses human rights impacts ensuring transparency and accountability,[103] and can be realised through in-person meetings, online dialogues, consultation with affected stakeholders and formal public reports.[104] Businesses are called to provide sufficient information through reporting with a view to enabling the evaluation of its response to adverse impacts.

B. Particular challenges for the HRDD in the context of AI

This subchapter will scan the particular challenges that AI causes for HRDD.

AI is said to be the fourth industrial revolution due to the large-scale changes it has brought in everyday life.[105] Huge investments are taking place in the name of innovation. However, innovation entails that constant changes take place so as to achieve optimisation, which can lead to changes in the risks. Thus, the risks for human rights will constantly change too. This change is also linked with the nature of AI per se. Its dynamic nature makes it difficult to predict its logic. The algorithm constantly learns from the data and changes accordingly[106] making the risk assessment even more difficult. Adding to that, the opaque character of the algorithms can create more difficulties in the effort to assess the risks emanating from their use.[107]

[102] UNGPs, Principle 29.
[103] Karin Buhmann, 'Neglecting the Proactive Aspect of Human Rights Due Diligence? A Critical Appraisal of the EU's Non-Financial Reporting Directive as a Pillar One Avenue for Promoting Pillar Two Action' (2018) 3 Business and Human Rights Journal 23, 24.
[104] UNGPs, Principle 21.
[105] Sean Gallagher, 'The fourth Industrial revolution emerges from AI and the Internet of Things' (*Ars Technica*, 18 June 2019) <https://arstechnica.com/information-technology/2019/06/the-revolution-will-be-roboticized-how-ai-is-driving-industry-4-0/> accessed 20 September 2019.
[106] Nicholas Diakopoulos, 'Accountability in Algorithmic Decision Making' (2016) 59 Communications of the ACM 56, 60.
[107] Knight (n 25) 57.

Another issue that makes HRIA in the context of AI even more complex is that it needs to take into account different factors depending on the geographic and cultural context.[108] The potential risks and the affected groups of stakeholders will be not be the same in geographically and culturally different States. For example, in a State with high level of patriarchy, it may be that women are not employed in higher positions the same way that men do. In these situations, tech companies which develop the algorithms may need to assess differently the risks for women as a specific affected group and take more mitigating measures in comparison to another State where the gender gap is not so vast.

Two characteristics of AI challenge the consultation with the affected groups of stakeholders. First, AI systems can affect millions of individuals. This is a characteristic that relates to ICT sector in general.[109] For example, the algorithms are trained with data of millions of individuals – often without their consent for the specific processing of their data - leading to an impact on the right to privacy. Also, the inferences of an algorithm which has been trained on biased data, can lead to a discriminatory treatment of individuals. Thus, many individuals could be potentially affected but it is extremely difficult to identify them at the beginning or even during the lifecycle of AI. Second, AI classifies people into groups, taking into account various elements that do not reflect the classical grounds on which discrimination is usually based, such as race, colour, sex, language, religion, political or other opinion among others.[110] An AI system can categorise some people in the same group not because of their common ethnicity, but because of their customer habits.[111] This issue renders the identification of the affected groups problematic,[112] and thus it has negative effect on the affected stakeholders. People who get categorised in these new groups may even not know that they belong there, they do not know their peers and consequently they cannot organise themselves, and appoint a representative to support their interests in the consultation with the tech companies.[113]

[108] Dillon Reisman, Jason Schultz, Kate Crawford and Meredith Whittaker, 'Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability' (AI Now*,* April 2018) 18 <https://ainowinstitute.org/aiareport2018.pdf> accessed 20 September 2019.

[109] Mccorquodale, Smit, Neely and Brooks, 'Human Rights Due Diligence in Law and Practice' (n 80) 210.

[110] ICCPR, art. 2§1.

[111] Alessandro Mantelero*, '*AI and Big Data: A blueprint for a human rights, social and ethical impact assessment' (2018) 34 Computer Law & Security Review 754, 763.

[112] Anton Vedder, 'Why Data Protection and Transparency are not enough when facing social problems of Machine Learning in a Big Data Context' in Emre BayamlioğLu, Irina Baraliuc, Liisa Janssens and Mireille Hildebrandt (Eds), *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen* (AUP 2018).

[113] Mantelero (n 111) 764.

The scope of HRDD should also cover the risks for adverse human rights impacts generated by the company's business relationships in relation to its products and services. As in other web-based service companies, the value chain can be very wide. It can include suppliers, resellers, customers and end-users.[114] The big data supply chain poses significant challenges, as different actors collect the data, store and use them[115] and it is very challenging to identify the actors involved in each stage of the process. In addition, the discreteness of AI can make the HRDD over the value chain even more challenging, as the different components of AI may be developed in different places, by different companies in different time.[116]

One issue that requires more attention, is that the State can be the end-user of the AI system. In fact, States tend to use more and more AI in the public services, like in resource allocation, criminal recidivism, health sector, and facial recognition. In this case, both the company and the State should conduct HRIA.[117] States may actually misuse AI and will ultimately impact human rights of individuals under their jurisdiction. However, again AI's nature makes it difficult to assess the relevant impacts.

It follows from the above that tech companies which design and develop AI face numerous challenges when it comes to conducting a HRDD. Nevertheless, it is necessary to find effective ways to guarantee that a proactive HRDD takes place.

C. Human Rights Due Diligence in the context of AI

The special nature of AI raises numerous issues which render the HRDD particularly challenging. Although there are several reports on the implementation of the UNGPs in the ICT sector,[118] there is still not much work addressing the implementation of UNGPs in the context of AI. The UN Human rights has just launched the B-Tech project which will investigate how companies can effectively use UNGPs to address human rights impacts of digital technologies. The second focus area is HRDD and end-use.

[114] European Commission, 'ICT Sector Guide' (n 85) 32.

[115] HRBDT, ''Background Paper on Consent Online' (n 43) 6.

[116] Scherer (n 26) 369.

[117] Allison-Hope and Hodge, 'Paper 1: Why a Rights-Based Approach?' (n 10).

[118] European Commission, 'ICT Sector Guide' (n 85); Business Social Responsibility, 'Applying the UN Guiding Principles on Business and Human Rights to the ICT industry' (September 2012) <http://www.bsr.org/reports/BSR_Guiding_Principles_and_ICT_2.0.pdf> accessed 20 September 2019; Institute for Human Rights and Business, 'Telecommunications and Human Rights: An Export Credit Perspective' (February 2017) <https://www.ihrb.org/uploads/reports/IHRB%2C_Telecommunications_and_Human_Rights_-_An_Export_Credit_Perspective%2C_Feb_2017.pdf> accessed 20 September 2019.

The draft scoping paper, which sets out the topics for the upcoming consultations, uses examples of AI in order to show the urgency of elaborating more on the HRDD in the context of digital technologies.[119]

This subchapter will suggest how companies should address the challenges generated by AI. Given the ongoing consultation and the absence of a comprehensive guide for the implementation of UNGPs dealing specifically with AI, there are no clear-cut answers. Given that the HRIA is an integral part of HRDD and all the next stages are closely entwined with it, we will first focus on that. We will review other types of impact assessments, like privacy impact assessment (hereinafter 'PIA'), data protection impact assessment (hereinafter 'DPIA'), algorithmic impact assessment (hereinafter 'AIA') and human rights, ethical and social impact assessment (hereinafter 'HRESIA') and we will underline what elements of those assessments could potentially be integrated in the HRIA in the context of AI. Afterwards, we will discuss some more specific aspects of HRDD that require special consideration.

In general, HRDD is not a one-off exercise. The unpredictable nature of AI calls for an ongoing HRDD[120] which constantly assesses the risks and ensures that the company takes mitigation measures.

1. Human rights impact assessment of AI

In this subchapter, we will first explore whether the HRIA in the context of AI could borrow elements from other types of impact assessments, and then we will go through some propositions on the form of HRIA.

1.1 Drawing upon other types of impact assessment

Among the different types of impact assessment, we will focus on those more relevant to AI. The aim is to investigate whether they could contribute to the formation of HRIA.

i. Privacy Impact Assessment

PIA assesses the impacts of a product, service or operation in general, on the rights to privacy. The assessment of the risks extends beyond a mere compliance test. De Hert proposes a set of criteria that

[119] UN Human Rights, 'UN Human Rights Business and Human Rights in Technology Project (B-Tech) - Draft Scoping Paper for Consultation' (30 July 2019), 6.

[120] Filippo Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz and Levin Kim, 'Artificial Intelligence and Human Rights, Opportunities and Risks' (Berkman Klein Center, September 2018) 54 <http://nrs.harvard.edu/urn-3:HUL.InstRepos:38021439> accessed 20 September 2019.

need to be assessed for an *"honest"* PIA.[121] The specific technology must be in accordance with the law, serve a legitimate aim, should not violate the core essence of the right to privacy, should be necessary in a democratic society, should not give unfettered discretion, should be proportionate, appropriate and achieved with the least intrusive means and should respect other human rights besides privacy. He emphasizes that PIA's findings must be public in every stage to facilitate public debate and possibly reinforce changes in the technology so that it makes it more privacy friendly.[122] The consultation with affected stakeholders is also critical for a successful PIA. Wright and De Hert propose that effective consultation could be achieved through interviews, workshops, mediation, role-playing and monitoring and evaluation techniques.[123]

As suggested by De Hert in the context of PIA, tech companies should evaluate many factors when deciding to design or use a new technology. One that should be included in HRIA is the criterion of whether the specific technology – in our case the AI system- is the least intrusive to human rights. This criterion reflects the principle of proportionality[124] which should be achieved between the interference with the right and the legitimate aim pursued.

ii. Data Protection Impact Assessment

The DPIA is more specific than PIA, as it focuses on the protection of personal data, which is only one aspect of privacy. De Hert argues that DPIA is a mere compliance check which examines whether a particular data processing complies with the requirements set in the data protection legislation.[125]

One of the overarching principles of the General Data Protection Regulation (hereinafter 'GDPR') is the principle of accountability, ie the obligation of the data controller to take appropriate technical and organisational measures to ensure but at the same time be able to demonstrate that its data processing operations comply with the provisions of the Regulation.[126] DPIA is one of these measures, but it is not always mandatory. The data controllers must conduct a DPIA only when a data processing is likely to

[121] Paul De Hert, 'A Human Rights Perspective on Privacy and Data Protection Impact Assessments' in Paul De Hert and David Wright (eds), *Privacy Impact Assessment*, vol. 6 (Springer 2012) 43-44.
[122] ibid 75.
[123] David Wright and Paul De Hert, 'Findings and Recommendations' in Paul De Hert and David Wright (eds), *Privacy Impact Assessment*, vol. 6 (Springer 2012) 470.
[124] De Hert, 'A Human Rights Perspective on Privacy and Data Protection Impact Assessments' (n 121) 61.
[125] ibid 34.
[126] GDPR, art 24§1.

pose high risks to the rights and freedoms of individuals.[127] The Article 29 working party had emphasized that, although the "rights and freedoms of individuals" primarily refer to data protection and privacy, they may involve other rights too, such as freedom of speech, freedom of thought, freedom of movement, prohibition of discrimination, right to liberty, conscience and religion.[128] In particular, Recital 75 specifies that the data processing operation may cause physical, material or non-material damage, which could include, among others, discrimination, financial loss, identity theft, disclosure of sensitive data, reputation harm.[129] Thus, initially, it is up to the data controller to assess whether a DPIA is necessary. GDPR provides that the supervisory authorities shall designate the processing operations which will require DPIA.[130] GDPR itself has a non-exhaustive list of processing operations which mandate a DPIA, including cases where, *"a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person".*[131] The Article 29 Working Party has also redacted a list in order to facilitate data controllers. This list includes among others predicting and scoring, any automated decision-making process with legal or similar significant effect, cases where data of vulnerable data subjects, such as children, asylum seekers and elderly people are processed and when there is an innovative use of technology or a new application of a technological solution.[132]

Both PIA and DPIA are mostly centred around the right to privacy and the right to protection of personal data respectively. They are useful but they miss a point which marks the difference between AI and other data processing operations. AI can have an impact on whole groups of people and PIA and DPIA only address the individual dimension of privacy without sufficiently addressing the collective dimension and societal concerns regarding AI's impacts.[133]

[127] GDPR, art 35§1.
[128] Article 29 Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679' WP 248 rev. 01 (2017) 6.
[129] GDPR, Recital 75.
[130] GDPR, art 35§5.
[131] GDPR, art 35§3(a).
[132] Article 29 Working Party (n 128) 9-10.
[133] Mantelero (n 111) 768.

iii. Algorithmic Impact Assessment

AI Now has issued a paper which proposes a new type of impact assessment which is designed to address the special characteristics of AI. Although AI Now designed this tool for public agencies, the UN Special Rapporteur on the Promotion and protection of the right to freedom of opinion and expression (hereinafter 'UN Special Rapporteur on Freedom of expression') proposed that companies working with AI should also conduct similar impact assessments.[134]

One of the most important elements of the AIA is its public character. One of the overarching reasons for the design of this tool lied in the fact that very few companies designing, developing or using AI make public their impact assessments. Consequently, affected stakeholders have only restricted information, coming from journalists, researchers and human rights defenders.[135] The AIA requires the public agency to give a public notice about a proposed or already existing AI system and give information about its purpose, the internal policy of using this system and the expected timeline of its deployment.[136] In the context of the self-assessment, the public agency assesses whether there are issues of inaccuracy, bias and harm on affected stakeholders.[137] The public disclosure includes the self-assessment and a plan proposing ways of granting access to external researchers who will be able to review the AI system.[138] The AIA also includes a comment period during which the public can scrutinise the algorithm based on the information that has already been disclosed. AI Now emphasizes that the public must have the means to challenge the deployment of AI in case the public agency did not rectify the flaws found or failed to comply with the requirements of the assessment. Thus, it proposes that there should be a due process challenge period during which the public can make a complaint before an oversight body or court.[139]

AI Now stresses that AIA would be also beneficial to the private companies, which sell their products to public agencies, since by letting external researchers review their algorithms, they will reinforce the

[134] UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (29 August 2018) 73rd session UN Doc A/73/348, para 53.

[135] Reisman, Schultz, Crawford and Whittaker (n 108) 4, 7.

[136] ibid 13.

[137] ibid 9.

[138] ibid 9.

[139] ibid 10.

public trust and will be more competitive.[140] That way, a potential race to the top could take place and more companies could facilitate the conduct of an AIA.[141]

The model of AIA takes into account the specific characteristics of AI. AIA could inform HRIA in many respects. First, its public character enables the public to be meaningfully engaged, and second, the proposition of an oversight body which will monitor the impact assessment is useful and could be adopted in the context of HRIA.

iv. Human rights, social and ethical impact assessment

Another proposition for AI is HRESIA, which is more sophisticated version of HRIA.[142] Mantelero has proposed this new type of impact assessment which embraces human rights, social and ethical concerns in one tool. He argues that human rights are mainly protected as individual rights, while AI's impact has implications on groups.[143] He points out that HRIA does not address adequately the ethical and social concerns arising from the use of AI. HRESIA first assesses the risks generated by AI against human rights. Noting that the ethical and social concerns can be different according to the geographical and cultural context, in the second layer, HRESIA assesses the risks against the applicable ethical and social values.[144] Thus, in a sense Mantelero tries to accommodate cultural relativism in AI through integrating a local dimension in the HRIA. In a third level, HRESIA assesses the risks against specific rights and principles according to each case.[145] He also proposes the establishment of an ad-hoc Committee which will ensure that HRESIA is conducted efficiently, respects human rights and fits in the local context.[146]

HRESIA seems to be a useful tool from which HRIA can be influenced. Finding a way to assess an AI system on the basis of its risks to human rights but also to ethics would be ideal. Already, companies have started paying special attention to ethics. For example, Microsoft has established an AI and Ethics in Engineering and Research committee which examines the ethics considerations arising from a new

[140] ibid16.
[141] ibid 16.
[142] Mantelero (n 111) 757.
[143] ibid 765.
[144] ibid 769.
[145] ibid 769.
[146] ibid 771.

technology and has the power to stop new sales in case it finds that it entails risks for ethics.[147] Given that there is a "hype" towards ethics, combining ethics and human rights impact assessment could serve as a way to ensure that tech companies also assess risks to human rights and not just ethics. However, it is important that human rights are not substituted by ethics.[148]

1.2 Propositions for Human rights impact assessment in AI

Below, we will indicate some important elements that HRIA in AI systems should have.

The Consultative Committee of Convention 108+[149] of the Council of Europe (hereinafter 'CoE') issued very recently its Guidelines on AI and Human Rights.[150] Except for the States, the Committee also addresses the private sector which works with AI, and more specifically the developers, manufacturers and service providers and gives them specific guidance. It stresses that the assessment of the AI-related risks should reflect a precautionary approach in the sense that private actors should assess the risks in advance so that they can take precautionary measures and mitigate them.[151]

Furthermore, in his recent report, the UN Special Rapporteur on Freedom of expression[152] highlights that HRIA is a useful tool for addressing the impacts of AI and that it should be conducted prior to the procurement, the development and the use of AI, and should be exposed to external review.[153]

HRIA should also include meaningful consultations with *"civil society, human rights defenders, local communities and representatives of marginalised and underrepresented users".[154]* Naturally, the consultation with potentially affected groups is difficult, and therefore tech companies must engage with civil society and human rights defenders who are specialised in digital rights and can understand the

[147] Dunstan Allison-Hope and Mark Hodge, 'Artificial Intelligence: A Rights-Based Blueprint for Business- Paper 3: Implementing Human Rights Due Diligence' (Business Social Responsibility, August 2018) 10 <https://www.bsr.org/reports/BSR-Artificial-Intelligence-A-Rights-Based-Blueprint-for-Business-Paper-03.pdf> accessed 20 September 2019.

[148] Ben Wagner, 'Ethics as an Escape from Regulation - From "Ethics-Washing" To Ethics-Shopping?' in Emre Bayamlioğlu, Irina Baraliuc, Liisa Janssens and Mireille Hildebrandt (eds), *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen* (AUP 2018).

[149] Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data (adopted 18 May 2018) as amended by Protocol CETS No. 223 ('Convention 108+').

[150] Council of Europe, 'Guidelines on Artificial Intelligence and Data Protection' Consultative Committee of the Convention For The Protection Of Individuals With Regard to Automatic Processing of Personal Data, T-PD(2019)01 (25 January 2019).

[151] Ibid para II.2.

[152] UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (n 134).

[153] ibid para 53.

[154] ibid para 68.

risks and recommend solutions. For example, Privacy International, Article 19, European Digital Rights (EDRI), Access Now, None of your Business (Noyb) are examples of NGOs, that are working on the intersection of technology with human rights and could provide valuable insights in the context of HRIA.

2. Other elements of Human Rights Due Diligence in the context of AI

Given the challenges, HRDD needs to fulfil certain criteria in order to accomplish its mission.

As illustrated in Chapter II, there are certain groups which are more vulnerable to the negative impacts of AI, eg. women, immigrants, LGBT people, ethnic minorities etc. It is important that not only HRIA but also HRDD in total adopts a right-holders' perspective.[155] This means that human rights concerns should be considered during the whole process of HRDD, from the assessment, to integrating, tracking and communicating the responses of the company.

Companies should use their leverage to contribute to the respect of human rights. For example the big 5 tech companies, Apple, Alphabet, Amazon, Microsoft, and Facebook have accumulated such great power , that their saying can have a big impact not only in influencing other companies, but also in inducing a positive insight in policy making.[156] Given the scale of potential negative impacts of AI on human rights, companies which design and develop AI, should ensure that their business partners operate in a human rights compliant way and that their customers will use AI similarly. Leverage can be exercised by companies through making public their human rights impact assessments. For example, HRBDT indicated that if big tech companies publish their human rights impact assessment on AI, it can work as positive example for more companies.[157]

One way to exercise leverage and influence the behaviour of business partners is through the so-called "responsible procurement of AI".[158] A tech company can incorporate human rights clauses in its contracts with third parties, which will set human rights compliance as a condition for the continuation of business relationship.[159] It could require that the business partner adopts a human rights policy or

[155] Allison-Hope and Hodge, 'Paper 3: Implementing Human Rights Due Diligence' (n 147) 14.
[156] Pasquale, *The Black Box Society* (n 23) 140.
[157] Mark Latonero, 'Governing Artificial Intelligence: Upholding human rights & dignity' (*Data & Society*, October 2018) 18 <https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf> accessed 20 September 2019.
[158] Allison-Hope and Hodge, 'Paper 3: Implementing Human Rights Due Diligence' (n 147) 10.
[159] European Commission, 'ICT Sector Guide' (n 85) 52.

that it adheres to codes of conduct that reflect human rights standards.[160] The punishment for the third party's refusal to comply with a human rights clause can range from the refusal to sign or renew the contract to the termination of the contract.[161]

The role of tech companies' leverage can also prove to be very beneficial when it is exercised in the context of their relationships with governments. Tech companies have the required technical expertise and can inform the state policy making through giving their professional insights.[162]

Additionally, in the context of tracking the responses to the adverse human rights risks of AI, tech companies could obtain certifications. For example, a certification of the algorithm as a software object can specify the design specifications of the algorithm and the performance-based standards against which the operation of an algorithm is assessed.[163] These standards could include human rights considerations, like respecting the right to privacy, non-discrimination, freedom of expression and possibly other rights depending on the use of the specific algorithm.

Another suggestion is the use of benchmarks and indicators. The Ranking Digital Rights Project has set numerous criteria with which it ranks tech companies according to their performance based on their disclosed commitments, policies and practices.[164] Although it focuses only on privacy and freedom of expression, it is helpful for holding companies accountable. The last index was published in May 2019 and it assessed 24 tech companies according to 35 indicators.[165]

Also, tech companies are advised to become members of the multi-stakeholder platform Global Network Initiative (hereinafter 'GNI'). Companies that join the GNI, commit to adhere to its principles and implementation guidelines. Although the principles are mostly centred to the protection of the right to privacy and freedom of expression, the platform gives the opportunity to tech companies to convene

[160] ibid.

[161] Mccorquodale, Smit, Neely and Brooks, 'Human Rights Due Diligence in Law and Practice' (n 80) 215.

[162] Allison-Hope and Hodge, 'Paper 3: Implementing Human Rights Due Diligence' (n 147) 16.

[163] Lilian Edward and Michael Veale, 'Slave to the Algorithm? Why a 'Right to an Explanation' as probably not the Remedy you are looking for' (2017) 16 Duke Law & Technology Review 18, 79.

[164] Ranking Digital Rights, 'About Ranking Digital Rights' <https://rankingdigitalrights.org/about/> accessed 20 September 2019.

[165]Ranking Digital Rights, '2019 RDR Corporate Accountability Index' (May 2019) <https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf?>accessed 20 September 2019.

and devise their strategies so that is more compliant to the right to privacy and freedom of expression. One of the principles of GNI is that the member companies shall conduct HRDD.[166]

Additionally, auditing is very important in the context of AI. Auditing plays a significant role in verifying whether the AI systems raise human rights issues. The opacity of the majority of AI systems makes their public scrutiny challenging and raises various issues due to the fact that the source code is a trade secret and qualifies for proprietary protection.[167] Therefore, the audit of the algorithmic process by a small group of experts, who will have access to the source code and the input and output data and will be able to examine the efficiency of the measures taken by the tech company, could serve as way to balance the competing interests. This issue will be further analysed in the next chapter.

The Human Rights Big Data and Technology Project (hereinafter 'HRBDT') of the University of Essex has proposed that the HRIA on AI should be monitored by independent oversight.[168] Parliamentary committees, judicial or quasi-judicial bodies and special courts could monitor the human rights impact assessments as oversight bodies. In that context, the Council of Europe has recommended that the National Human Rights Commissions could be involved in the oversight.[169] HRBDT further proposes that in case of defect, the oversight bodies could have the power to force companies to redesign the algorithm or discontinue the automatic decision-making process and notify the affected groups or individuals.[170] Indeed this seems a good solution as it ensures an independent oversight which will have human rights at the centre, however it depends largely on the level of expertise of the oversight bodies members who should possess specialised technical skills and human rights education.

In the context of 'know and show' approach, companies are expected to draft formal reports when the business operation entails severe human rights risks due to its nature or due to the nature of its operating context.[171] As indicated above, AI entails numerous risks for human rights. As illustrated in Chapter II, AI can impact human rights in an ongoing manner and sometimes without people's awareness. Additionally, it can affect whole groups of individuals. Thus, I believe that tech companies

[166] GNI, 'Implementation Guidelines', 2§4 < https://globalnetworkinitiative.org/implementation-guidelines/> accessed 20 September 2019.
[167] Kamarinou, Millard and Singh (n 11) 107.
[168] McGregor, Ng, Shaheed, Abrusci, Kent, Murray and Williams (n 8) 37.
[169] CoE, 'Unboxing Artificial Intelligence: 10 steps to protect Human Rights - Recommendation' (May 2019) 7.
[170] Ibid 37-38.
[171] UNGPs, Principle 21.

which design, deploy and use AI systems which are capable of generating such risks, should conduct formal reporting and report on the identified human rights risks. The European Commission developed the Code of Practice on Disinformation in the context of the elections of the European Parliament in May 2019,[172] with a view to commit social media platforms and advertising industry to ensure transparency and respect of democracy during the pre-elections period.[173] Facebook, Google and Twitter were among the tech companies that signed the Code of Practice, and had to produce a monthly report including the measures they took to ensure transparency. Although the evaluation of the reports showed that the big tech companies should work more on fighting disinformation and ensuring transparency, the European Commission recognised the progress they made.[174] Reporting, even if it may be embellished, forces companies to show their performance and creates an incentive for them to do well in the fear of reputation loss.

One of the main drawbacks of HRDD, but also of Business and Human Rights in general, is the focus on the 'no harm principle'. The corporate responsibility to respect human rights is negatively formed leaving outside the concept of positive action in order to protect human rights.[175] I believe that the way HRDD is constructed is no longer sufficient especially in the AI context. Tech companies have taken the role of the state, since governments totally depend on the companies for the use of algorithms. Naturally, the States are the primary duty bearers and are responsible for monitoring the enforcement of the HRDD.[176] Therefore, they should be actively involved to make corporate human rights compliance real.[177]

This chapter focused on HRDD of tech companies in the context of AI. The special characteristics of AI render the conduct of HRDD very challenging. Having reviewed different types of impact assessment, we saw that HRIA could borrow some elements, such as the public nature of AIA and the incorporation of ethical considerations as proposed by Latonero in the context of HRESIA. The last part of this chapter

[172] European Commission, 'Code of practice on Disinformation' (September 2018).
[173] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Securing free and fair European elections' COM (2018) 637 final, 5.
[174] European Commission, 'Code of Practice against disinformation: Commission recognises platforms' efforts ahead of the European elections' (17 May 2019) <https://europa.eu/rapid/press-release_STATEMENT-19-2570_en.htm> accessed 20 September 2019.
[175] Ramasastry (n 81) 247.
[176] Fasciglione (n 66) 114.
[177] Justine Nolan, 'Refining the Rules of the Game: The Corporate Responsibility to Respect Human Rights' (2014) 30(78) Utrecht Journal of International and European Law 7, 21.

made recommendations with regard to what HRDDA should include in order to be sufficient and better ensure the respect of human rights in the context of AI.

IV. Human rights by design

The complicated nature of AI systems calls for ex-ante measures that will be able to ensure a higher level of human rights protection. Within this context, in order to comply with HRDD, tech companies could use the tool of "human rights by design". Human rights by design is a principle that entails a commitment of businesses to make sure that the design of products, services and technologies respects, by default, human rights.[178] More specifically it means that a technology should be designed and developed in a way that puts the human at the centre and takes into account the whole spectrum of human rights.[179]

In the context of AI, the issue of human rights by design has not been elaborated sufficiently. The Commissioner for Human Rights of CoE has stressed the urgent need to put human rights at the centre of AI technologies' design.[180] The Parliamentary Assembly of CoE, in its Recommendation on AI and human rights, stressed that the design of the algorithms must respect human dignity and human rights and give particular consideration to the rights of vulnerable people.[181]

Although there has been a lot of discussion about integration of human rights considerations in other sectors, such as mining and manufacturing, which has resulted in the production of best practices, there is no equivalent body of knowledge for the incorporation of human rights values in the design and development of new technologies, and even more so in AI particularly.

This chapter will shed light on the value of human rights by design in the context of AI. We will first focus on the role of design in technology in general and then we will draw comparisons from the notion of "privacy by design" which has been a central idea in the data protection legal framework. We will then focus on transparency, an important subset of human rights by design. We will see whether

[178] Jonathon Penney, Sarah McKune, Lex Gill, and Ronald J. Deibert, 'Advancing human-rights-by-design in the dual-use technology industry' (2018) 71 Journal of International Affairs 103, 106.

[179] Dunstan Allison-Hope, 'Human Rights by Design' (BSR, 17 February 2017) <https://www.bsr.org/en/our-insights/blog-view/human-rights-by-design> accessed 20 September 2019.

[180] Commissioner for Human Rights, 'Artificial intelligence and human rights' (High-level conference of Council of Europe, Helsinki, February 2019) 2 <https://www.coe.int/en/web/commissioner/-/we-need-to-act-now-and-put-human-rights-at-the-centre-of-artificial-intelligence-designs> accessed 20 September 2019.

[181] Council of Europe, 'Technological convergence, artificial intelligence and human rights' Parliamentary Assembly Recommendation 2102 (2017) (28 April 2017) para 9.1.5.

transparency can be meaningful through an analysis of the trade-offs that it might entail. All these elements will help us explore the content of human rights by design in the context of AI.

A. Technology as form of regulation

This subchapter will emphasise why the adoption of human rights by design could have a major impact on the effective protection of human rights in the context of AI. In order to support this thesis, we will use some arguments that relate to the use of technology as another way of constraining the human behaviour. As will be shown below, the design of technology has been considered to be a type of regulation. The inclusion of human rights considerations in the initial phase of AI design, could have a major positive impact on the protection of human rights.

It has been argued that technology has the capacity to regulate our behaviour. Lessig famously stated that "Code is law".[182] Lessig's work focused on the regulation of cyberspace. He claimed that there are four types of constraints regulating the users' activities in the cyberspace: namely the law, social norms, the market and the code.[183] What Lessig actually suggests is that the code, meaning the instructions built in software and hardware, serves as a new type of regulation; it is what he calls "the architecture".[184] The architecture of cyberspace defines which acts can take place and which acts cannot. Even though a user can choose not to abide by the legal and social norms, for example by committing acts or omissions which are forbidden under cybercrime law, there is no such option in the case of the code.[185] The difference lies in the fact that the architecture of cyberspace does not need the cooperation of the users in order to enforce its power. Thus, the code sets the rules and users have to follow these rules in order to be part of the game and enjoy the benefits of the cyberspace.

Reidenberg introduced the term *"lex informatica*", which incorporates the rules for information flows introduced by technology through network designs, standards and system configurations.[186] He said that the exponential pace of technology can undermine the efficiency of the law and adds that technology could restrict the ability of the government to regulate.[187] That is why he suggests that the

[182] Lawrence Lessig, *Code version 2.0* (Basic Books 2006) 122-123.
[183] Lawrence Lessig, *Code and other laws of cyberspace* (Basic Books 1999) 235-236; Lessig, '*Code version 2.0'* (n 182) 124.
[184] Lessig, '*Code version 2.0'* (n 182) 121.
[185] Lawrence Lessig, 'Reading the Constitution in cyberspace' (1998) 45 Emory Law Journal 869, 896, 899.
[186] Joel Reidenberg, 'Lex Informatica: The Formulation of Information Policy Rules Through Technology' (1998) 76 Texas Law Review 553, 554-555.
[187] ibid 586.

policy makers need to be aware of what technology can do and promote the adoption of technical standards for the achievement of policy goals.

Hartzog recently wrote a book specifically dedicated to the design of new technologies with regard to privacy. He supported that the design is power, since it equips the designers with the power to control how users interact with digital technologies.[188] The architects of the systems, meaning the designers have the power to determine the goals that their designs will serve. As Hartzog noted, the design of digital technology embodies the values of its creators and it should be able to empower human values instead of impairing them.[189] Like the architects in the real space who put the human in the centre of their architecture and serve the needs of people,[190] the designers of cyberspace should seek to do the same. For example, in real space, architects build roads in a way that accommodate the needs of disabled people. Similarly, in the digital space designers can design technology in a way that puts the human at the centre and accommodates the interests of vulnerable groups.

The explosion of the new digital era has rendered the abovementioned findings very topical and more relevant than ever. Companies of the ICT sector, and especially social media platforms, have taken the lead and have been informally established as co-regulators.[191] However, with great power, comes great responsibility, and as Reidenberg has stressed, the influence of private actors in regulation, entails a public responsibility.[192]

Koops attempted to set out which criteria should be fulfilled in order for the normative technology to be considered acceptable.[193] He argues that there are primary and secondary criteria. Primary criteria are basic human rights and democratic values. Secondary criteria include transparency, accountability and flexibility. Koops suggests that there is a hierarchy between the two sets of criteria, and primary criteria should have a certain priority over secondary ones.[194] It is worth pointing out that under this approach, basic human rights such as the rights to equality and non-discrimination, freedom of expression and

[188] Woodrow Hartzog, *Privacy's Blueprint: The Battle to Control the Design of New Technologies* (Harvard University Press 2018) 23, 34.

[189] ibid 43.

[190] Richard Buchanan, 'Human Dignity and Human Rights: Thoughts on the Principles of Human-Centered Design' (2001) 17 Design Issues 35, 37.

[191] Benvenisti (n 20) 56.

[192] Reidenberg (n 186) 592.

[193] Bert-Jaap Koops, 'Criteria for Normative Technology – The Acceptability of 'Code as Law' in Light of Democratice and Constitutional Values' in Karen Yeung and Roger Brownsword (eds), *Regulating technologies : legal futures, regulatory frames and technological* fixes (Hart Publishing Ltd 2008) 168.

[194] ibid 171.

privacy, which are mostly affected by the use of algorithms, belong to the primary criteria that need to be met.

As Brownsword has said, it needs to be ensured that technology is developed and applied in such a way that does not compromise human rights and human dignity.[195]

If we apply the abovementioned findings to AI, it follows that integrating human rights considerations from the beginning of designing an AI system, could serve as a tool to proactively protect human rights. Designers have the power to infuse human rights considerations in the design of the algorithms in such a way that it will be difficult to sheer off the predetermined design and contravene human rights.

B. Drawing comparisons from privacy by design

Privacy by design is a tool that has become very popular due to the application of GDPR in the EU, which introduced the data protection by design. More specifically, art. 25 GDPR prescribes that *"the data controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational and security measures, such as pseudonymisation, which are designed to implement data-protection principles"*.[196] According to GDPR, data protection considerations should be integrated in the process of developing and designing services and products which are based on processing personal data.[197] GDPR attempts to balance the protection of rights and freedoms of natural persons with the business interests. In that respect, the choice of organisational and security measures depends on a wide range of factors, such as the state of the art, the cost of implementation and the nature, scope, context and purposes of processing, which are more relevant to the business interests, as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, which are related to the protection of rights and freedoms of natural persons.[198]

Privacy by design is a principle that has been firstly introduced by Ann Cavoukian, the former Information and Privacy Commissioner of Ontario. Even though privacy by design is limited to the

[195] Roger Brownsword, 'What the World Needs Now: Techno-Regulation, Human Rights and Human Dignity' in Roger Brownsword (ed), *Global Governance and the Quest for Justice, Human Rights*, vol 4 (Hart Publishing 2004) 205

[196] GDPR, art 25§1.

[197] GDPR, Recital 78.

[198] GDPR, art 25§1.

protection of privacy, it can be helpful to explore the usefulness of human rights by design approach in AI, beyond the protection of privacy.

Cavoukian proposed that privacy should be integrated in the whole spectrum of business operations, ranging from the design processes to organisational priorities, objectives and planning operations.[199] This design-thinking perspective is what we call privacy by design. Cavoukian went further by setting out the principles that inform privacy by design. Privacy by design is proactive and not reactive, in the sense that it is to be employed before any risk materialises. Companies should adopt a culture of continuous improvement and keep implementing privacy by design techniques that are even higher than the minimum requirements set by domestic and international laws. Designers should therefore take into account privacy considerations from the very beginning of the design process, even at the stage of the conception of the idea so as to prevent any negative impact on privacy.[200]

According to Cavoukian privacy by design has the following characteristics.[201] Privacy by design should be integrated in every IT system by default, so that privacy is protected *a priori*. Additionally, the respect of the right to privacy should be embedded in the design and architecture of a system without diminishing its functionality. This can be achieved through the adoption of a holistic, integrative and creative approach. The approach is holistic because it needs to take into account all broader contexts, such as functionality, profit but also privacy concerns. It is integrative as it is necessary to engage all relevant stakeholders and finally it is creative, since some design choices may need to change in order to accommodate new challenges and at the same time continue protecting privacy. Another aspect of privacy by design is that it reflects a win-win situation, in the sense that it balances privacy with innovation, efficiency and business profit. Lastly and most importantly, privacy by design places the humans and their interests at the centre of the design.

Rachovitsa has suggested that the law needs to be complemented by technological means, such as privacy by design in order to ensure privacy. She uses as examples two standardisation bodies, the

[199] Ann Cavoukian, 'Privacy by Design The 7 Foundational Principles Implementation and Mapping of Fair Information Practices' (Internet Architecture Board, 2011) <https://iab.org/wp-content/IAB-uploads/2011/03/fred_carter.pdf> accessed 20 September 2019.

[200] Ewa Luger and Michael Golebewski, 'Towards Fostering Compliance by Design; Drawing Designers into the Regulatory Frame' in Mariarosaria Taddeo and Luciano Floridi (eds), *The Responsibilities of Online Service Providers* (Law, Governance and Technology Series 31, Springer 2017) 297.

[201] Cavoukian (n199).

Internet Architecture Board ('IAB') and the Internet Engineering Task Force ('IETF'), two bodies that have embedded privacy considerations by default into their design of the network and Internet protocols.[202] That way privacy has been established as guiding principle of the Internet design. The incorporation of technical standards into design reinforces the provisions of human rights law, since it facilitates the compliance with them.[203]

Human rights by design can build on the experiences already obtained by the use of privacy by design. Given the successful use of privacy by design, it would be very beneficial to broaden the scope and incorporate considerations of other human rights as well, such as non-discrimination, freedom of expression, right to health and right to employment. [204] Drawing comparisons from privacy by design, human rights by design should be adopted proactively by tech companies. Human rights considerations should be taken into account by default while tech companies need to make sure that stakeholders' views are integrated since the beginning of the design phase. Also, the idea that privacy by design balances system functionality with the protection privacy, would fit well in the human rights by design and could serve as an incentive for tech companies to adopt it.

C. Transparency as a subset of human rights by design in the context of AI

Transparency is generally considered to be a virtue. Public entities must provide transparent procedures in order to minimise the risks of misconduct but transparency is also relevant in the private sector. Transparency is a "*passpartou*t" principle, which can be both part of the human rights due diligence and serve as a basis for remedies. In that respect, transparency is a fundamental component of human rights by design and it constitutes a necessary safeguard for the protection of human rights. However there are several arising questions. Should tech companies design explainable algorithms? Would that have an impact on their accuracy? Apart from the design, transparency is a fundamental part of HRDD, since companies are expected to be transparent in how they address adverse human rights risks.[205] Is transparency feasible when dealing with AI? The answers to these questions are not straightforward, because the specific characteristics of AI create tensions between the commercial

[202] Adamantia Rachovitsa, 'Engineering and lawyering privacy by design: understanding online privacy both as a technical and an international human rights issue' (2016) 24 International Journal of Law and Information Technology 374, 376.
[203] ibid 376.
[204]Allison-Hope and Hodge, 'Paper 3: Implementing Human Rights Due Diligence' (n 147) 13.
[205] UNGPs, Principle 15(b) and 21.

interests of tech companies and the human rights interests. This subchapter will delve into the nature of transparency and will identify the reasons that make it a significant part of human rights by design. It will then analyse the arising trade-offs of algorithmic transparency in order to finally examine how a balance between the competing interests can be achieved.

1. Transparency: Why is it necessary?

Nowadays there is a lot of discussion about transparency. It is presented as one of the most valuable tools for making up for the information asymmetry between tech companies designing AI and the public.[206] Transparency is a valuable tool to identify the human rights risks generated by algorithms and further correct any errors in the system.[207] The opaque nature of AI as well as other characteristics, such as the dynamic nature of its operation and its complex nature, render transparency challenging and maybe sometimes unrealistic.

It is very important that transparency covers the whole life-cycle of any given AI operation. AI has to be trained on large datasets in order to be able to make correlations and ultimately produce inferences. As indicated in Chapter II, individuals often do not realise that their data are collected and are further used to feed algorithms. Thus, it is important that individuals are aware when their data are collected. However, information about the collection itself is not sufficient. One of the most challenging aspects of the data exploitation arena, is that data that were previously collected for a certain purpose, are then processed for a new purpose without the knowledge of the individual. Therefore, transparency can be a means through which data subjects will be able to be informed about the repurposing of processing of their personal data. However, this is not free from challenges, given that allegedly anonymised data can serve as proxies leading to personal data as indicated in Chapter II. In that case, given that there is no collection of personal data in the first place, providing individuals with information is impossible as data protection law does not apply.[208]

Transparency is also a means to achieve algorithmic accountability.[209] This facet of transparency refers more to the phase where an algorithm has already made a prediction and interfered with a person's

[206] Sonia K. Katyal, 'Private Accountability in the Age of Artificial Intelligence' (2019) 66 UCLA Law Review 54, 61.

[207] McGregor, Murray and Ng (n 15) 321.

[208] Sandra Wachter, 'Data protection in the age of Big Data' (2019) 2 Nature Electronics 6, 7.

[209] Diakopoulos (n 106) 58.

rights, eg. non-discrimination. In that case, the individual will need to have access to a remedy. Under GDPR, one of the remedies is the right to an explanation. Although there is lot of debate on whether GDPR actually prescribes this right,[210] this discussion is outside of the scope of the present dissertation. It suffices to say that if we accept that there is such a right under GDPR, it entails that the individuals can have access to the logic and the factors that were taken into account in the process of the automated decision taken by the algorithm.[211] The Organisation of Economic Co-operation and Development (hereinafter 'OECD') has issued a Recommendation on AI, which provides that companies should ensure that AI systems are transparent so as to facilitate the affected individuals not only to understand how they have been impacted but also to give them a leeway to challenge the interference with their rights.[212]

Last but not least, the trust of individuals can be further enhanced through strengthening transparency of AI. There is a public suspicion that algorithms are incontrollable[213] and can have irreparable impact on human rights. This suspicion is fuelled by events such as the Cambridge Analytica scandal, which caused great distrust over digital technologies.[214] This scandal refers to the exposure of data of 87 million Facebook accounts to an algorithm created by Cambridge Analytica, with the aim of identifying swing voters in the last US presidential elections and Brexit campaign and further target them with political advertisements. All this happened behind the scenes, without individuals being able to understand why they are targeted with political advertisements. Lack of transparency played an intrinsic role in this operation. Nevertheless, in the past few years, many other scandals have led individuals and society as a whole, to lose their trust in AI.[215] The embracement of transparency and the design of explainable algorithms is a key feature to re-build the trust of individuals in AI.[216]

[210]Bryce Goodman and Seth Flaxman, 'EU regulations on algorithmic decision-making and a "right to explanation"' (ICML Workshop on Human Interpretability in Machine Learning, 2016) <https://arxiv.org/pdf/1606.08813.pdf > accessed 20 September 2019; Sandra Wachter, Brent Mittelstadt and Luciano 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 76.

[211] Lilian Edwards and Michael Veale, 'Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?' (2018) IEEE Computer and Reliability Societies 46, 47.

[212] OECD, 'Recommendation of the Council on Artificial Intelligence' (22 May 2019), 1§3.

[213] Roger Taylor, 'No Privacy without Transparency' in Ronald Leenes, Rosamunde van Brakel, Serge Gutwirth and Paul De Hert (eds), *Data Protection and Privacy – The Age of Intelligent Machines* (Hart Publishing 2017) 66.

[214] Carole Cadwalladr and Emma Graham-Harrison, 'Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach' (*Guardian*, 17 March 2018) https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election accessed 20 September 2019.

[215] Trisha Ray, 'Formulating AI Norms: Intelligent Systems and Human Values' Issue Brief No. 313 (Observer Research Foundation, September 2019) 4 https://www.orfonline.org/wp-content/uploads/2019/09/ORF_Issue_Brief_313_AINorms.pdf accessed 20 September 2019.

[216] Knight (n 25) 61.

2. Trade-offs

Although transparency is indeed a virtue and a fundamental part of the human rights by design, there are some trade-offs that need to considered. If tech companies design explainable algorithms, there may be a repercussion on their accuracy and innovation. Even if algorithms are explainable, transparency would be difficult to achieve, since tech companies would claim that the algorithms constitute trade secrets and deserve proprietary protection, and thus would deny to disclose the source code of the AI system. Even if full transparency was required, the question remains as to whether that would be meaningful, since the lack of technical expertise prevents individuals from understanding how the AI system works. This subchapter will identify these trade-offs which essentially bring into the surface the eternal conflict between the competing interests of human rights and commercial interests.

Explainability refers to the design of the AI system in such a way that it enables individuals to understand how it operates and produces its output data. However, the degree of an algorithm's accuracy depends to a large extent on its complexity.[217] That is to say that the more complex the algorithm is, the more accurate results it produces. On the other hand, the more complex the algorithm is, the more difficult it is to interpret it.[218] The difficulty to explain the complexities of the algorithm becomes clearer, if we take into account that algorithms, especially in machine learning systems, are continuously changing their model as they integrate their previous predictions as new training data.[219] Thus, an attempt to make algorithms more explainable would potentially sacrifice a part of accuracy and thus efficiency of the AI system.

Another argument against explainability of algorithms, is that it may have an impact on innovation.[220] This argument is based on the fact that building simple algorithms would entail that they may be less accurate and therefore would repress innovation. For example, USA and China have been more successful in the development of AI in comparison to the European Union (hereinafter 'EU'). One of the reasons is that the EU has more restrictive data protection rules which restrain innovation. Indeed, the

[217] Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms'(2016) Big Data & Society 1, 5.

[218] Goodman and Flaxman (n 210) 29.

[219] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, 'Accountable algorithms' (2017) 165 University of Pennsylvania Law Review 633, 660.

[220] Access Now (n 6) 36.

current system provides for fines for the data controller of up to 20 million euros or the 4% of the total worldwide annual turnover of the preceding financial year, in case of data breaches.[221]

Furthermore, designing an explainable algorithm may require more financial resources than what companies are willing to spend.[222] Designing more transparent algorithms is technically demanding and thus more expensive for the company. Very often, companies need to assess whether the cost of transparency is worth it, taking into account the competitive advantage that might arise from the use of a specific algorithm.[223] Thus, explainability entails more financial expenses on behalf of the company and it is not for granted that many tech companies would be willing to spend more financial resources to meet this requirement. This point becomes more relevant in the context of small and medium size enterprises, whose budget is not as high as the budget of tech giants, like Facebook and Microsoft.

Another obstacle is presented by the fact that even if algorithms were more explainable, tech companies would oppose the disclosure of the source code based on the argument that it constitutes trade secret and it deserves proprietary protection.[224] Tech companies contend that revealing the source code would facilitate the adversaries to steal their work and optimise their own algorithms. Thus, the requirement for full transparency may serve as a disincentive for tech companies to design new algorithms. [225] This argument supports that this would have a detrimental effect to the competition among companies and would further serve again as a barrier for innovation. Similarly, it has been argued that investors may be discouraged to invest in algorithms.[226] Investing in a system that can be easily duplicated and used by another company, would minimise the competitive advantage and would discourage investors to make big investments on AI systems.

Tech companies further oppose full transparency on the ground that it could open the door to the manipulation of their AI systems.[227] Companies fear that full transparency will allow competitors or malicious actors to gain access to their algorithms and game the system.[228] Pasquale uses the example

[221] GDPR, art 83.

[222] McGregor, Murray and Ng (n 15) 323.

[223] ibid.

[224] Citron and Pasquale, 'The Scored Society' (n 36) 5.

[225] Joshua New, 'How (and how not) to fix AI' (*Tech Crunch*, 26 July 2018) <https://techcrunch.com/2018/07/26/how-and-how-not-to-fix-ai/> accessed 20 September 2019.

[226] ibid.

[227] Committee of Experts on Internet Intermediaries of Council of Europe, 'Algorithms and Human Rights' (n 39) 38.

[228] Tal Z. Zarsky, 'Transparent predictions' (2013) 4 University of Illinois Law Review 1503, 1553; Rieke, Bogen and Robinson (n 13) 24.

of the “PageRank method” used by Google for ranking the results in the search engine.[229] The more transparent this AI system became, the more websites tried to manipulate it in order to gain a higher position in Google search. Zarsky explains that the dynamic nature of the algorithm, renders it susceptible to manipulation as the interference with the system is not easy to be traced.[230]

Additionally, even if the source code was revealed to the public, the lack of expertise among the population could render transparency meaningless. The majority of the population has not received special training on AI, and consequently, people would not be in a position to understand the working method of an algorithm, even if they had access to the source code.[231] This point is further strengthened, if we consider that sometimes even the designers themselves cannot understand how these systems work.[232]

Therefore, we can see that the trade-offs of algorithmic transparency are several. There are two competing interests, human rights interests which are represented by the principle of transparency and commercial interests which lean more to secrecy. Companies need to find a balance between those two competing interests. Building explainable algorithms, is very beneficial for people, especially potentially affected groups in order to understand how they may be affected by AI. On the other hand, sacrificing accuracy, would again have an impact on tech companies’ performance but also on individuals again due to inaccurate predictions.

3. Balance between human rights interests and business interests

On the one hand, transparency is beneficial for the protection of human rights, while on the other hand it entails numerous trade-offs which ultimately pose the question of whether transparency is a realistic requirement in the context of AI systems. In this part, we will identify whether these two sides can meet at the centre through mutual compromises.

As indicated above, the quality of the training data is very fundamental for the efficient operation of AI. Having access to the historic data on which the algorithm was trained, would facilitate the examination

[229] Ibid.
[230] ibid 1553.
[231] Kamarinou, Millard and Singh (n 11) 108.
[232] Mike Ananny and Kate Crawford, ‘Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability’ (2016) New Media & Society 1, 9.

of whether there is hidden bias that can lead to discriminatory inferences. There are some new techniques in place which claim to enhance the explainability of AI. For example, the designers of the algorithm can reduce the number of the variables that are taken into account by the algorithm in order to make the inference.[233] That way, it gets easier to explain the reasoning behind a certain inference. However, as illustrated in Chapter II, even if an algorithm is not trained on biased data, there are proxies which are linked with specific characteristics, such as ethnic origin, gender, income, and can further lead to discriminatory treatment. Computer scientists have supported that in case that the proxies which are correlated to the protected groups are removed, this could have an impact on the accuracy of the algorithm.[234] Thus, removing the attributes may reduce the possibility of discrimination, but at the same time the accuracy and the efficiency of the system will be at stake. However, designers could include corrective mechanisms in the design of the algorithm, which would decrease the possibilities of a biased automated decision.[235] For example, immigrants are usually affected by the decisions of algorithms due to the biased data that sometimes the algorithms are fed with. One corrective mechanism which could be integrated in the design, would be to programme the algorithm in a way that it integrates data from the affected groups, such as data of immigrants in the given case, so that they are not disregarded and disproportionately affected.[236]

Citron asserts that the algorithms should be designed in a way that would serve transparency and accountability.[237] In that respect, she supports that vendors should disclose the source code of the algorithm to the public in order to show how the system works.

Pasquale has identified three questions that need to be addressed in order to solve the transparency riddle; (1) how much the algorithm needs to reveal, (2) who should be the recipient of such information and (3) how fast it should be revealed.[238] He argues that full transparency would have significant repercussions and thus proposes the solution of a qualified transparency. He supports that full transparency should be reserved only to a small group of experts.[239] Pasquale and Citron propose that a way to balance the competing interests of the company and human rights, would be to entrust neutral

[233] Rieke, Bogen and Robinson (n 13) 24.
[234] Barocas and Selbst, ‘Big Data's Disparate Impact’ (n 61) 721.
[235] Eyal Benvenisti (n 20) 61.
[236] Ibid 61.
[237] Danielle Keats Citron, ‘Technological Due Process’ (2008) 85 Washington University Law Review 1249, 1308.
[238] Pasquale, *The Black Box Society* (n 23) 142.
[239] Ibid 142.

experts with the power to examine the algorithm, assess whether the inferences are fair, and evaluate the variables that are taken into account.[240]

Tutt has also supported the establishment of an independent agency that would be in charge of scrutinising an AI system.[241] In particular, he emphasizes the importance of an *"ex-ante regulation"*, which entails that this agency should be in the position to scan carefully the algorithm and have the power to prohibit its entry to the market in case it is dangerous.[242] He stresses that this form of *"ex-ante regulation"* is necessary because it can identify potential harms and mitigate them already at the phase of the design and development.[243]

Also, the Special Rapporteur on Freedom of Expression proposes audits as a tool to scrutinise AI. Specifically, he recommends several ways of auditing that would not counteract with the proprietary protection of the algorithms.[244] One such way is the zero-knowledge proof which evaluates the compliance of the algorithm with certain properties without scrutinising the algorithm.[245]

Diakopoulos further stressed that full transparency of the source code is an "*overkill in many if not most cases*".[246] Instead, he argues that it would be more efficient if the companies showed aggregate results and benchmarks to the public, in order to communicate the performance of the algorithm.[247] More specifically, he contends that information to be communicated should include the human involvement, details about the quality of data and how these data were collected and edited, the model itself, statistics about the margin of error of the algorithmic inferences as well as whether an algorithm is used.[248]

Annany and Crawford contend that access to the code does not suffice.[249] They stretch their argument and support that even the access to the code and the training data would have minimal benefit, since it would offer a "snapshot" of the algorithm's functionality given the dynamic nature of the process.[250]

[240] Citron and Pasquale, 'The Scored Society' (n 36) 28.
[241] Andrew Tutt, 'An FDA for algorithms' (2017) 69 Administrative Law Review 83, 117-118.
[242] Ibid.
[243] ibid.
[244] UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (n 134) para 56.
[245] Ibid.
[246] Diakopoulos (n 106) 58.
[247] ibid 58-59.
[248] ibid 60.
[249] Mike Ananny and Kate Crawford, 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability' (2016) New Media & Society 1, 9.
[250] ibid 10.

Especially in the context of machine learning systems, the fact that the algorithm continuously learns and adapts to its environment renders the benefit of transparency relatively limited in terms of time.

The Special Rapporteur on Freedom of Expression has stressed that individuals should be informed when a decision-making process is automated, should receive information about the logic used by the algorithm and most importantly, they should be informed at the time of their data collection, whether these data will be used to feed an AI system.[251] He points out that this information should be understandable and be written in clear and intelligible manner. He stresses that instead of trying to make AI understandable to the general public which lacks technical expertise, companies should communicate to individuals information about the existence, purpose, constitution and impact of an AI system.[252] He supports that this information will serve better the role of transparency in comparison to full transparency of the source code, the training data as well as the input and output data.[253] This is what he calls radical transparency. Similarly to Diakopoulos, he proposes that companies should be in a position to provide aggregate data which will essentially include statistics about the performance of the algorithm.[254]

Following from the above analysis, qualified transparency seems to be the ideal solution to balance human rights and commercial interests. As we saw, designing explainable algorithms is beneficial for individuals but only to some extent. It can therefore be said that a way to balance the transparency need with business interests is to entrust a small group of auditors with the task to audit the AI system and examine whether it complies with human rights standards. Auditing is part of the human rights due diligence, thus it gets more imperative for tech companies designing and developing AI to enable independent experts to scrutinise the algorithms without jeopardising the proprietary protection and their trade secret, but at the same time without getting away with inscrutable algorithms. Simultaneously, tech companies should provide potentially affected stakeholders with information about the AI system, such as the human involvement, details about the quality of the data, the way they were collected, the model itself, statistics on the margin of error and whether an algorithm is used.[255]

[251] UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (n 134) para 49.
[252] ibid para 50.
[253] ibid.
[254] Ibid para 51.
[255] Diakopoulos (n 106) 60

D. Human Rights by design in AI

As analysed above, the architecture or in other words the design of technology, is considered to be a form of regulation. The design of the algorithm could serve as a constraint of peoples' behaviours by proactively protecting human rights. The algorithm reflects the choices of its designers, including potential biases,[256] therefore it is fundamental that the designers set from the beginning the frames within which the algorithm will operate. Given the opaque nature of most of AI system and taking into account that explainability and full transparency are not always meaningful, it is important that tech companies adopt a human rights by design.

The adoption of human rights by design would be beneficial particularly in the context of AI for several reasons. First, it would strengthen the trust of individuals on the companies which develop AI systems. The nearly incontrollable collection of big data and their use by AI systems have had a devastating effect on users' trust. As stated above, in the recent years, individuals have started losing their trust on tech companies which deal with their data. However, trust is an important element for sustaining the digital ecosystem. The adoption of a human rights by design approach by corporate actors would cultivate the trust of users. Nevertheless, trust is not only related to tech companies. The use of AI has expanded and has covered fundamental aspects of life, such as health, participation in political life, employment and justice. Having trust that the AI systems respects human rights can translate to having trust in major functions of society.

Additionally, the added value of human rights by design is that it has a collective dimension. As previously shown, a characteristic that differentiates AI from other technologies, is that it has an impact on whole groups, and not only on the individual. For example, an algorithm can have an impact on black people because it has been trained on historic biased data which are more related to white people. Similarly to privacy by design, human rights by design can move beyond the individual and provide a more collective protection of human rights.[257]

In order for human rights by design to be able to work and achieve its mission, it is important to bring together different disciplines; human rights specialists, computer scientists, engineers and sales and

[256] Katyal (n 206) 67.
[257] Edwards and Veale, 'Enslaving the Algorithm' (n 211) 82.

marketing teams.[258] The integration of human rights principles in the design and development process cannot but be a common effort with insights from all of these disciplines. Human rights specialists can identify the potential risks of algorithms and guide the other two disciplines through the principles of human rights, such as the requirement of prior, free, informed consent and use of unbiased data for elimination of any potential discriminatory treatment. The sales and marketing teams are the ones which can contribute to this effort by bringing into play the business interests which should also be taken into account. These considerations will ultimately help the designers and developers to design new algorithms which will put the human at the centre, but at the same time they will make sure not to minimise their functionality and harm the business objectives.

Furthermore, human rights by design entails that designers and developers of AI, as well as the auditors who will be in charge of auditing the algorithms, should receive a comprehensive human rights training. Knowledge on human rights is urgent since all these actors are entrusted with ensuring that algorithms do not impact human rights.

There is a need for a cultural shift in the approach of all stakeholders towards AI. The extensive use of AI poses many challenges that threaten the enjoyment of human rights. All stakeholders must act before the AI market and developed systems become uncontrollable as a result of lack of human rights regulation. They need to consider human rights not as an accessory but as a fundamental part of their business.

At the end though, the State is the primary duty bearer for the enjoyment of human rights. Thus, in order for States to fulfil their obligations and protect human rights, they should make tech companies more accountable through domestic legislation.[259]

This chapter focused on the importance of human rights by design in the context of AI. First, we saw how technology can serve as a way to regulate behaviour and thus, integrating human rights considerations in the design of AI would be very beneficial for human rights. Privacy by design can serve as reference point for the adoption of human rights by design. Its proactive character, its balancing

[258]Allison-Hope, 'Human Rights by Design' (n 179); Latonero (n 157) 25.

[259] Paul De Hert, 'Accountability and System Responsibility: New Concepts in Data Protection Law and Human Rights Law' in Daniel Guagnin, Leon Hempel, Carla Ilten, Inga Kroener, Daniel Neyland and Hector Postigo (eds), *Managing Privacy through Accountability* (Palgrave Macmillan 2012) 206.

character between privacy and business interests, can be a starting point from which human rights by design can be further built. Then, we focused on transparency, which is one of the most important subsets of human rights by design. The special characteristics of AI do not render algorithmic transparency always meaningful, thus the concept of qualified transparency was endorsed. Finally, in order to fully embrace human rights by design, companies need to adopt a multidisciplinary approach, ensure that designers and developers of AI systems receive human rights training and put the human at the centre of AI. Above everything, tech companies need to realise their responsibility in this era of exponential development of AI and adopt a new corporate culture, which will give more attention to human rights. Otherwise, things can turn to be uncontrollable.

V. CONCLUSION

The exponential rate of AI development has created a shift of powers. Although, unlike States, companies are not duty bearers of human rights under international human rights law, their prevalence on the digital era requires appropriate approaches protecting human rights.

This dissertation first identified the special characteristics of AI which differentiate it from other technologies in order to show why it requires special treatment in the context of the fulfilment of tech companies' corporate responsibility to respect human rights. It found that AI's dependency on massive and often inaccurate amounts of data, is one of the factors that impact human rights. Additionally, the opaque nature of algorithms entails that AI systems are inscrutable while the dynamic nature of their process makes the assessment of risks and the attribution of harm very challenging. These characteristics can lead to implications on various human rights. Here, we focused on the right to privacy, freedom of expression and right to non-discrimination. Indeed, the data-driven economy exposes individuals to threats of their privacy, as their data are collected massively and are used in novel ways. The use of AI can also have a chilling effect on the freedom of expression. Individuals censor themselves due to the constant sense of surveillance while their right to receive information is restricted as content moderation algorithms can create echo chambers. We also saw examples where the right to non-discrimination is violated by the use of AI while individuals may be discriminated against and may not be aware.

Chapter III explained how the special characteristics of AI render the HRDD of tech companies designing, developing and selling AI particularly challenging. It then explored how HRDD should be shaped in order to effectively respond to the challenges generated by AI. Among other types of impact assessment, AIA which specifically focuses on AI, proved to be helpful as it offers many opportunities of engagement with stakeholders while also 'HRESIA', which combines human rights and ethical considerations was found to accommodate better the collective dimension of human rights impacts of AI. Then the dissertation proposed other elements in which tech companies should focus in the context of HRDD, such as adopting a rights holders' perspective, exercising leverage to influence not only the supply chain but also the policy making, using benchmarks and auditing in order to track the responses to human rights impacts and engaging in reporting activities to ensure transparency. One of the propositions that came out was that an oversight independent body should be established in order to oversee the HRIAs and the responses of tech companies.

Chapter IV then focused on the importance of human rights by design as a means to ensure a human rights-centred AI. We first highlighted how beneficial the design of AI can be in the effort to make AI respectful of human rights. The idea of that argument was that designers of AI have the power to set the frames in the functions of AI and they are in a position to infuse human rights considerations ever since the beginning of AI's lifecycle. The privacy by design principle was further used as an example which can give guidance on what human rights by design should entail. Moving on, we focused on transparency, which although is an overarching principle which should be inherent in the whole spectrum of human rights due diligence, we identified several trade-offs which bring into surface the conflict between the competing human rights interests and commercial interests. More specifically, transparency is necessary for informing individuals about the further repurposing of the processing of their data, explaining potentially affected groups how an algorithm may impact their rights or explaining already affected individuals how their rights were violated. On the other hand, designing explainable algorithms may have an impact on the accuracy of predictions, may restrict innovation in general and small and medium-size companies may oppose to it due to the higher financial resources that it requires. And even if the algorithms were explainable, tech companies would use trade secret and proprietary protection claims or they would use the risk of manipulation of AI systems as a pretext to deny the disclosure of the source code. The lack of technical expertise of the population is also used as an argument against the meaningfulness of designing explainable algorithms. This dissertation argues that a way to balance the two competing interests is to enable small groups of experts to examine the operation of AI systems and their compliance with human rights but at the same time communicate information to individuals about the human involvement, the quality of data, the collection of data, the model of the AI system and the margin of error. Lastly, human rights by design should also entail that companies adopt a multidisciplinary approach through engaging different disciplines - human rights specialists, computer scientists and sales marketing teams in order to achieve the desired balance. Human rights education is also a necessary element as AI puts a lot of human rights at stake.

As a final note, I believe that AI is a cure and a curse. It can certainly bring a lot of positive impacts and improve our lives. However, the implications on human rights can be numerous and tech companies are in a position to intervene already in the phase of design and infuse human rights considerations. Although until now they do not have obligation to respect, protect and fulfil human rights, the great power that they accumulated does not leave any more space for inaction. It is time that tech companies take seriously their corporate responsibility to respect human rights and do their part.

VI. Bibliography

Legislation

Treaties

Convention for the Protection of Human Rights and Fundamental Freedoms (adopted 4.11.1950, entered into force 03.09.1953) ETS 005 ('ECHR')

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

American Convention on Human Rights (adopted 22.11.1969, entered into force 18.07.1978) OAS Treaty Series No 36 ('ACHR')

African Charter on Human and Peoples' Rights (adopted 27.06.1981, entered into force 21.10.1986) 21 ILM 58 ('African Charter')

Charter of Fundamental Rights of the European Union [2012] OJ C 326/391 ('EU Charter')

Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data (adopted 18 May 2018) as amended by Protocol CETS No. 223 ('Convention 108+)

Declarations

Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR)

UN Vienna Declaration and Program of Action (1993) UN Doc A/CONF.157/23

EU Secondary Law

European Parliament and Council Regulation 2016/679/EU 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 [2016] OJ L119/1 ('GDPR')

Books

Gillespie, T., Boczkowski, P. J., and Foot, K. A., *The relevance of algorithms, Media technologies:*

Essays on communication, materiality, and society (MIT Press 2014)

Hartzog, W., *Privacy's Blueprint: The Battle to Control the Design of New Technologies* (Harvard University Press 2018)

Lessig, L., *Code and other laws of cyberspace* (Basic Books 1999)

Lessig, L., *Code version 2.0* (Basic Books 2006)

Nadia Bernaz, *Business and Human Rights: History, law and policy- Bridging the accountability gap* (Routledge 2017)

O'neil, C., *Weapons of Math Destruction: How Big Data increases inequality and threatens democracy* (Penguin Random House UK 2016)

OECD, *Artificial Intelligence in Society* (OECD Publishing 2019) 22 <https://www.oecd-ilibrary.org/docserver/eedfee77-en.pdf?expires=1568930907&id=id&accname=oid051805&checksum=B82141F6397F76C33F77524BFB3F4F0E> accessed 20 September 2019

Pasquale, F., *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015)

Ruggie, J. G., *Just Business – Multinational Corporations and Human Rights* (W.W. Norton & Company 2013)

Solove, D. J., *Understanding Privacy* (Harvard University Press 2008)

CONTRIBUTIONS TO EDITED BOOKS

Brownsword, R., 'What the World Needs Now: Techno-Regulation, Human Rights and Human Dignity' in Brownsword, R. (ed), *Global Governance and the Quest for Justice, Human Rights*, vol 4 (Hart Publishing 2004)

De Hert, P., 'A Human Rights Perspective on Privacy and Data Protection Impact Assessments' in

De Hert, P., and Wright, D. (eds), *Privacy Impact Assessment*, vol. 6 (Springer 2012)

De Hert, P., 'Accountability and System Responsibility: New Concepts in Data Protection Law and Human Rights Law' in Guagnin, D., Hempel, L., Ilten, C., Kroener, I., Neyland, D., and Postigo, H. (eds), *Managing Privacy through Accountability* (Palgrave Macmillan 2012)

De Vries, K., and Hildebrandt, M., 'Introduction: Privacy, due process and the computational turn at a glance' in Hildebrandt, M., and De Vries, K. (eds), *Privacy, Due Process and the Computational Turn* (Routledge 2013)

Deva, S., 'Treating human rights lightly: a critique of the consensus rhetoric and the language employed by the Guiding Principles' in Deva, S., and Bilchitz, D. (eds), *Human Rights Obligations of Business: Beyond the Corporate Responsibility to Respect?* (CUP 2013)

Kamarinou, D., Millard, C., and Singh, J., 'Machine learning with personal data' in Leenes, R., Van Brakel, R., Gutwirth, S., and De Hert. P. (eds), *Data Protection and Privacy-The Age of Intelligent Machines,* vol. 10 (Hart Publishing 2017)

Kamarinou, D., Millard, C., and Singh, J., 'Machine learning with personal data' in Leenes, R., Van Brakel, R., Gutwirth, S., and De Hert, P. (eds), *Data Protection and Privacy- The Age of Intelligent Machines,* vol 10 (Hart Publishing 2017)

Koops, B., 'Criteria for Normative Technology – 'The Acceptability of 'Code as Law' in Light of Democratice and Constitutional Values' in Yeung, K., and Brownsword, R. (eds), *Regulating technologies: legal futures, regulatory frames and technological* fixes (Hart Publishing Ltd 2008)

Koops, B., 'On decision transparency, or how to enhance data protection after the computational turn' in Hildebrandt, M., and De Vries, K. (eds), *Privacy, Due Process and the Computational Turn* (Routledge 2013)

Luger, E., and Golebewski, M., 'Towards Fostering Compliance by Design; Drawing Designers into the Regulatory Frame' in Taddeo, M., and Floridi, L. (eds), *The Responsibilities of Online Service Providers* (Law, Governance and Technology Series 31, Springer 2017)

McCorquodale,R., ‘International Human Rights Law Perspectives’ in Blecher, L., Kaymar Stafford N., and Bellamy, G.C. (eds), *Corporate Responsibility for Human Rights Impacts: New Expectations and Paradigms* (American Bar Association 2014)

McGoldrick, D., ‘Thought, Expression, Association, and Assembly’ in Moeckli, D., Shah, S., and Sivakumaran, S. (eds), *International Human Rights Law* (3rd edition, Oxford University Press 2018)

Mireille Hildebrandt, ‘Defining Profiling: A New Type of Knowledge?’ in Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen:* 17 *Cross-Disciplinary Perspectives* (Springer 2008)

Nemitz, P., ‘Profiling the European citizen: why today’s democracy needs to look harder at the negative potential of new technology than at its positive potential’ in BayamlioğLu, E., IBaraliuc, I., Janssens, L., and Hildebrandt, M. (Eds), *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen* (AUP 2018)

Pieter Kleve, P., and Richard De Mulder, R., ‘Privacy protection and the right to information: in search of a new symbiosis in the information age’ in Mercado Kierkegaard, S. (ed), *Cyberlaw, Security & Privacy* (Ankara Bar Association Press 2007)

Taylor, R., ‘No Privacy without Transparency’ in Leenes, R., Van Brakel, R., Gutwirth, S., and De Hert, P (eds), *Data Protection and Privacy – The Age of Intelligent Machines* (Hart Publishing 2017)

Van Otterlo, M., ‘A machine learning view on profiling’ in Hildebrandt, M., and De Vries, K. (eds), *Privacy, Due Process and the Computational Turn* (Routledge 2013)

Vedder, A., ‘Why Data Protection and Transparency are not enough when facing social problems of Machine Learning in a Big Data Context’ in BayamlioğLu, E., Baraliuc, I., Janssens, L., and Hildebrandt, M. (eds), *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen* (AUP 2018)

Wagner, B., ‘Ethics as an Escape from Regulation - From “Ethics-Washing” To Ethics-Shopping?’ in Bayamlioğlu, E., Baraliuc, I., Janssens, L., and Hildebrandt, M. (eds), *Being Profiled: Cogitas*

Ergo Sum: 10 Years of Profiling the European Citizen (AUP 2018).

Wright, D., and De Hert, P., 'Findings and Recommendations' in De Hert, P., and Wright, D. (eds), *Privacy Impact Assessment*, vol. 6 (Springer 2012)

<u>JOURNALS</u>

Ananny, M., and Crawford, K., 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability' (2016) New Media & Society 1

Barocas, S., and Selbst, A. D., 'Big Data's Disparate Impact' (2016) 104 California Law Review 671

Benvenisti, E., 'Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance?' (2018) 19 European Journal of International Law 9

Bonnitcha, J. and McCorquodale, R., 'The Concept of 'Due Diligence' in the UN Guiding Principles on Business and Human Rights' (2017) 28 European Journal of International Law 899

Brandeis, W., 'The Right to Privacy' (1890) 4 Harvard Law Review 193

Buchanan, R., 'Human Dignity and Human Rights: Thoughts on the Principles of Human-Centered Design' (2001) 17 Design Issues 35

Buhmann, K., 'Neglecting the Proactive Aspect of Human Rights Due Diligence? A Critical Appraisal of the EU's Non-Financial Reporting Directive as a Pillar One Avenue for Promoting Pillar Two Action' (2018) 3 Business and Human Rights Journal 23

Burrell, J., 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2016) Big Data & Society 1

Citron D.K., and Pasquale, F., 'The Scored Society: Due Process for automated predictions' (2014) 89 Washington Law Review 1

Citron, D. K. and Pasquale, F., 'The Scored Society: Due Process for automated predictions' (2014) 89 Washington Law Review 1

Citron, D.K., 'Technological Due Process' (2008) 85 Washington University Law Review 1249

Datta,A., Tschantz, M. C., and Datta, A., ‘Automated Experiments on Ad Privacy Settings’ (2015) 1 Proceedings on Privacy Enhancing Technologies 92

Diakopoulos, N., ‘Accountability in Algorithmic Decision Making’ (2016) 59 Communications of the ACM 56

Edward, L., and Veale, M., ‘Slave to the Algorithm? Why a ‘Right to an Explanation’ as probably not the Remedy you are looking for’ (2017) 16 Duke Law & Technology Review 18

Edwards, L., and Veale, M., ‘Enslaving the Algorithm: From a ‘Right to an Explanation’ to a ‘Right to Better Decisions’?’ (2018) IEEE Computer and Reliability Societies 46

Fasciglione, M., ‘The enforcement of corporate human rights due diligence’ (2016) 10 Human Rights & International Legal Discourse 94

Fasterling, B. and Demuijnck, G., ‘Human Rights in the Void? Due Diligence in the UN Guiding Principles on Business and Human Rights’ (2013) 116 Journal of Business Ethics 799

Fasterling, B., ‘Human Rights Due Diligence as Risk Management: Social Risk Versus Human Rights Risk’ (2017) 2 Business and Human Rights Journal 225

Katyal, S., Private Accountability in the Age of Artificial Intelligence (2019) 66 UCLA Law Review 54

Katyal, S.K., ‘Private Accountability in the Age of Artificial Intelligence’ (2019) 66 UCLA Law Review 54

Knight, W., ‘The Dark Secret at the heart of AI’ (2017) 120 MIT Technology Review 55

Knight, W., ‘The Dark Secret at the heart of AI’ (2017) 120 MIT Technology Review 55

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H., ‘Accountable algorithms’ (2017) 165 University of Pennsylvania Law Review 633

Lehr, D., and Ohm, P., ‘Playing with the Data: What Legal Scholars Should Learn about Machine

Learning' (2017) 51 University of California Davis Law Review 653

Lessig, L., 'Reading the Constitution in cyberspace' (1998) 45 Emory Law Journal 869

Mantelero, A., 'AI and Big Data: A blueprint for a human rights, social and ethical impact assessment' (2018) 34 Computer Law & Security Review 754

Mc Corquodale, R., Smit, L., Neely, S. and Brooks, R., 'Human Rights Due Diligence in Law and Practice: Good Practices and Challenges for Business Enterprises' (2017) 2 Business and Human Rights Journal 195

McGregor, L., Murray, D., and Ng, V., 'International Human Rights Law as a Framework for Algorithmic Accountability' (2019) 68 International and Comparative Law Quarterly 309

Nolan, J., 'Refining the Rules of the Game: The Corporate Responsibility to Respect Human Rights' (2014) 30(78) Utrecht Journal of International and European Law 7

Penney, J., McKune, S., Gill, L. and Deibert, R. J., 'Advancing human-rights-by-design in the dual-use technology industry' (2018) 71 Journal of International Affairs 103

Rachovitsa, A., 'Engineering and lawyering privacy by design: understanding online privacy both as a technical and an international human rights issue' (2016) 24 International Journal of Law and Information Technology 374

Ramasastry, A., 'Corporate Social Responsibility Versus Business and Human Rights: Bridging the Gap Between Responsibility and Accountability' (2015) 14 Journal of Human Rights 237

Reidenberg, J., 'Lex Informatica: The Formulation of Information Policy Rules Through Technology' (1998) 76 Texas Law Review 553

Scherer, M., 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29 Harvard Journal of Law & Technology 353

Tutt, A., 'An FDA for algorithms' (2017) 69 Administrative Law Review 83

Wachter, S., 'Data protection in the age of Big Data' (2019) 2 Nature Electronics 6

Wachter, S., Mittelstadt B., and Floridi, L., 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 International Data Privacy Law 76

Zalnieriute, M., 'An International constitutional moment of data privacy in the times of mass-surveillance' (2015) 23 International Journal of Law and Information Technology 99

Zarsky, T.Z., 'Transparent predictions' (2013) 4 University of Illinois Law Review 1503

PAPERS

Allison-Hope, D. and Hodge, M., 'Artificial Intelligence: A Rights-Based Blueprint for Business-Paper 3: Implementing Human Rights Due Diligence' (Business Social Responsibility, August 2018) <https://www.bsr.org/reports/BSR-Artificial-Intelligence-A-Rights-Based-Blueprint-for-Business-Paper-03.pdf> accessed 20 September 2019

Allison-Hope, D., and Hodge, M., 'Artificial Intelligence: A Rights-Based Blueprint for Business-Paper 1: Why a Rights-Based Approach?' (Business Social Responsibility, August 2018) <https://www.bsr.org/reports/BSR-Artificial-Intelligence-A-Rights-Based-Blueprint-for-Business-Paper-01.pdf> accessed 20 September 2019

Business Social Responsibility, 'Applying the UN Guiding Principles on Business and Human Rights to the ICT industry' (September 2012)<http://www.bsr.org/reports/BSR_Guiding_Principles_and_ICT_2.0.pdf> accessed 20 September 2019

Cavoukian, A., Privacy by Design The 7 Foundational Principles Implementation and Mapping of Fair Information Practices (2011) <https://iab.org/wp-content/IAB-uploads/2011/03/fred_carter.pdf> accessed 20 September 2019

Goodman, B., and Flaxman, S., 'EU regulations on algorithmic decision-making and a "right to explanation"' (ICML Workshop on Human Interpretability in Machine Learning, 2016) <https://arxiv.org/pdf/1606.08813.pdf > accessed 20 September 2019

Human Rights Big Data and Technology Project, 'Background Paper on Consent Online' (June

2019) <https://hrbdt.ac.uk/wp-content/uploads/2019/06/19.06.09-Background-Paper-on-Consent-Online.pdf> accessed 20 September 2019

Raso,F., Hilligoss, H., Krishnamurthy, V., Bavitz, C., and Kim, L., 'Artificial Intelligence and Human Rights, Opportunities and Risks' (Berkman Klein Center, September 2018) <http://nrs.harvard.edu/urn-3:HUL.InstRepos:38021439> accessed 20 September 2019

Ray,T., 'Formulating AI Norms: Intelligent Systems and Human Values' Issue Brief No. 313 (Observer Research Foundation, September 2019) <https://www.orfonline.org/wp-content/uploads/2019/09/ORF_Issue_Brief_313_AINorms.pdf> accessed 20 September

Reisman, D., Schultz, J., Crawford, K., and Whittaker, M., Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability (AI Now, April 2018) <https://ainowinstitute.org/aiareport2018.pdf> accessed 20 September 2019

REPORTS

Access Now, 'Human Rights in the Age of Artificial Intelligence' (November 2018) <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf> accessed 20 September 2019

Article 19 and Privacy International, 'Privacy and Freedom of Expression In the Age of Artificial Intelligence' (April 2018) <https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf> accessed 20 September 2019

Borgesius, F., 'Discrimination, artificial intelligence, and algorithmic decision-making' (*Council of Europe*, 2018) <https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73> accessed 20 September 2019

Committee of Experts on Internet Intermediaries of Council of Europe, 'Algorithms and Human Rights - Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications' DGI(2017)12 (*Council of Europe*, March 2018) <https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5 accessed 20

September 2019> accessed 20 September 2019

European Commission, ‘ICT Sector Guide on Implementing the UN Guiding Principles on Business and Human Rights’ (June 2013) <https://www.ihrb.org/pdf/eu-sector-guidance/EC-Guides/ICT/EC-Guide_ICT.pdf> accessed 20 September 2019

Independent High-Level Expert Group on Artificial Intelligence, ‘A Definition of AI: Main Capabilities and Disciplines’ (April 2019) <https://www.aepd.es/media/docs/ai-definition.pdf> accessed 20 September 2019

Institute for Human Rights and Business, ‘Telecommunications and Human Rights: An Export Credit Perspective’ (February 2017) <https://www.ihrb.org/uploads/reports/IHRB%2C_Telecommunications_and_Human_Rights_-_An_Export_Credit_Perspective%2C_Feb_2017.pdf> accessed 20 September 2019

Institute for Human Rights and Business, ‘Telecommunications and Human Rights: An Export Credit Perspective’ (February 2017)

Latonero, M., ‘Governing Artificial Intelligence: Upholding Human Rights & Dignity’ (*Data &Society*, October 2018) <https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf> accessed 20 September 2019

McGregor, L., Ng, V., Shaheed, A., Abrusci, E., Kent, C., Murray, D. and Williams, C., ‘The Universal Declaration of Human Rights at 70 - Putting Human Rights at the heart of the Design, Development, and Deployment of Artificial Intelligence’ (*Human Rights Big Data and Technology – University of Essex*, December 2018) <https://48ba3m4eh2bf2sksp43rq8kk-wpengine.netdna-ssl.com/wp-content/uploads/2018/12/UDHR70_AI.pdf> accessed 20 September 2019

Rieke, A., Bogen, M. and Robinson, D., ‘Public Scrutiny of Automated Decisions: Early Lessons and Emerging Decisions’ (*Omidyar Network and Upturn*, February 2018) <https://www.omidyar.com/sites/default/files/file_archive/Public%20Scrutiny%20of%20Automated%20Decisions.pdf> accessed 20 September 2019

The Royal Society, 'Machine learning: the power and promise of computers that learn by example (April 2017) <https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf> accessed 20 September 2019

BLOGS

Allison-Hope, D., 'Human Rights by Design' (Business for Social Responsibility, 17 February 2017) <https://www.bsr.org/en/our-insights/blog-view/human-rights-by-design> accessed 11 September 2019

Gallagher, S., 'The fourth Industrial revolution emerges from AI and the Internet of Things' (*Ars Technica*, 18 June 2019) <https://arstechnica.com/information-technology/2019/06/the-revolution-will-be-roboticized-how-ai-is-driving-industry-4-0/> accessed 20 September 2019

ARTICLES ON THE NEWS

Angwin, J., Larson, J., Mattu, S. and Kirchner, L., 'Machine Bias' (*ProPublica*, 23 May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 20 September 2019

Cadwalladr, C. and Graham-Harrison, E., 'Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach' (*Guardian*, 17 March 2018) <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election> accessed 20 September 2019

New, J., 'How (and how not) to fix AI' (*Tech Crunch*, 26 July 2018) <https://techcrunch.com/2018/07/26/how-and-how-not-to-fix-ai/> accessed 20 September 2019

WEBSITES

European Commission, 'Code of Practice against disinformation: Commission recognises platforms' efforts ahead of the European elections' (17 May 2019) <https://europa.eu/rapid/press-release_STATEMENT-19-2570_en.htm> accessed 20 September 2019

Ranking Digital Rights, '2019 RDR Corporate Accountability Index' (May 2019)

<https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf> accessed 20 September 2019

Ranking Digital Rights, 'About Ranking Digital Rights' <https://rankingdigitalrights.org/about/> accessed 20 September 2019

Corporate Human Rights Benchmark, <https://www.corporatebenchmark.org/> accessed 20 September 2019

MISCELLANEA

Article 29 Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679' WP 248 rev. 01 (2017)

Council of Europe Commissioner for Human Rights, Human Rights Comment: Safeguarding human rights in the era of artificial intelligence (3 July 2018) <https://www.coe.int/en/web/commissioner/-/safeguarding-human-rights-in-the-era-of-artificial-intelligence>

Council of Europe, 'Guidelines on Artificial Intelligence and Data Protection' Consultative Committee of the Convention For The Protection Of Individuals With Regard to Automatic Processing of Personal Data, T-PD(2019)01 (25 January 2019)

Council of Europe, 'Technological convergence, artificial intelligence and human rights' Parliamentary Assembly Recommendation 2102 (2017) (28 April 2017)

Council of Europe, 'Unboxing Artificial Intelligence: 10 steps to protect Human Rights - Recommendation' (May 2019)

Danish Institute of Human Rights, 'Human Rights Indicators for Business platform' (2016) <https://www.business-humanrights.org/sites/default/files/HRCA_INTRODUCTION.pdf> accessed 20 September 2019

ECOSOC, 'Norms on the Responsibilities of Transnational Corporations and other Business Enterprises with Regard to Human Rights' (13 August 2003) UN Doc E/CN.4/Sub.2/2003/12/Rev.2

European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Securing free and fair European elections' COM (2018) 637 final

Global Reporting Initiative, 'GRI 102: General Disclosures' (2016)

GNI, 'Implementation Guidelines', <https://globalnetworkinitiative.org/implementation-guidelines/> accessed 20 September 2019

HRC, 'General Comment 18' (10 November 1989) UN Doc HRI/GEN/1/Rev.9 (Vol. I)

HRC, 'General Comment No. 16: Article 17 (Right to Privacy) The Right to Respect of Privacy, Family, Home and Correspondence, and Protection of Honour and Reputation' (8 April 1988) UN Doc CCPR/C/21/Add.6

Human Rights Council, 'UN Guiding Principles on Business and Human Rights' (2011) UN Doc A/HRC/17/31 ('UNGPs')

OECD, 'Recommendation of the Council on Artificial Intelligence' (22 May 2019)

Office of High Commissioner on Human Rights, 'The Corporate responsibility to respect human rights: An Interpretative Guide' (2012)

UN Human Rights, 'UN Human Rights Business and Human Rights in Technology Project (B-Tech) - Draft Scoping Paper for Consultation' (30 July 2019)

UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (29 August 2018) 73rd session UN Doc A/73/348

UNHRC, 'Business and Human Rights: Towards Operationalizing the 'Protect, Respect and Remedy' Framework' (22 April 2009) UN Doc A/HRC/11/13