CENTRAL EUROPEAN UNIVERSITY

Tackling Biases and Discrimination in the AI Regulatory Framework

– A Comparative Analysis of EU and U.S. –

By

Cezara-Alexandra Panait

LL.M in Human Rights, Capstone Thesis

Supervisor: Professor Judit Sandor

Legal Studies Department

Central European University

June 7th, 2020

Table of Contents

Acknowledgements

During these times of uncertainty, we, the students, learned to quickly adapt and we tried our best to face the challenges of finishing our degree during a pandemic. First and foremost, I am thankful to the Central European University, together with its administration, faculty and staff, for ensuring all the necessary conditions for our unhampered education.

I would like to express my deepest gratitude to Professor Judit Sandor, my capstone thesis supervisor, for her valuable contribution to my research. I am thankful for the support and advice, which helped me to advance my knowledge and to improve my writing.

I am also thankful to Professor Sejal Parmar, who has been guiding me throughout the academic year and has always encouraged me to follow my passion for Artificial Intelligence and human rights.

Nevertheless, I am grateful for the opportunity to learn many substantial notions of anti-discrimination law from Professor Mathias Möschel, which increased the value of my research.

I would also like to express my great appreciation for Professor Cameran Ashraf, who shares my genuine passion for technology and human rights, and who helped me explore this field.

Lastly, I would like to show my wholehearted gratitude to my family and my closest friends, for their unconditional support.

Abstract

One of the main human rights risks posed by Artificial Intelligence (hereinafter: AI) systems is the reinforcement of discrimination and biases on various grounds, including race, sex, gender, sexual orientation, age or poverty. The present research focuses on the main regulatory and ethical initiatives on AI, in a comparative analysis on the perspectives of the European Union (hereinafter: EU) and the United States of America (hereinafter: US). After mapping the discriminatory tendencies, the study presents the different regulatory approaches to AI and non-discrimination. Further, the legally binding framework on non-discrimination and data protection is assessed. The study continued with the analysis of a series of interviews with a whole range of stakeholders in the area of AI policy-making. The research concludes with a set of recommendations for policy-makers and stakeholders working in the AI regulatory environment. Thus, the main proposals of the study are the following: (1) conduct a comprehensive mapping on existing legal frameworks to analyze the feasibility of AI regulations; (2) advance the debate on AI ethics; (3) determine the adequate legal instrument for regulatory intervention, which can include sectorial regulations or adapting non-discrimination and data protection legislation; (4) ensure representative and high-quality datasets and (5) strengthen the cooperation between all stakeholders involved in the AI policy-making process.

List of abbreviations

AI	Artificial Intelligence
CAHAI	Ad Hoc Committee on Artificial Intelligence
CFR	Charter of Fundamental Rights of the European Union
CoE	Council of Europe
EU	European Union
GDPR	General Data Protection Regulation
HRIA	Human Rights Impact Assessment
ML	Machine Learning
UNGPs	The United Nations Guiding Principles on Business and Human Rights
U.S.	United States of America

Introduction

The impact of Artificial Intelligence (hereinafter: AI) systems is already noticed in our daily lives. AI is influencing social behaviors and it raises multiple legal questions. Thus, one consequential challenge of implementing AI tools is related to the interference of AI with fundamental rights. Until now, when analyzing the collision of human rights with AI, most of the scholars have focused on freedom of expression and the right to privacy.[1] However, I would like to address another potential risk posed by AI to the non-discrimination principle, which will be the research focus of my thesis.

AI is widely used in criminal justice system, to make predictions of the defendants' risk of reoffending, and it also provides information about in which geographical areas police forces should patrol.[2] At the same time, it is used in the healthcare field, for advanced diagnosis and treatment.[3] In assessing individuals' credit worthiness and in the employment sector, AI decision-making models are increasingly used.[4] Another application of AI is encountered in the online content display and moderation, to determine which content should be available on the Internet.[5] Nonetheless, this emergent technology is nowadays used in predicting and fighting epidemics, in self-driving cars or autonomous weapons.[6]

[1] Filippo Raso and others, 'Artificial Intelligence & Human Rights: Opportunities & Risks' [2018], p. 5, SSRN Electronic Journal <https://www.ssrn.com/abstract=3259344> accessed 4 May 2020
[2] Richardson, Rashida and Schultz, Jason and Crawford, Kate, 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data', Predictive Policing Systems, and Justice, February 2019, p. 7, <https://ssrn.com/abstract=3333423> accessed 6 June 2020
[3] Raso (n 1), p. 18
[4] Ibid, p. 18
[5] Ibid, p. 18
[6] Council of Europe, 'Discrimination, Artificial Intelligence, and Algorithmic Decision-Making', p. 7, accessed 6 June 2020

While a standard definition of AI has not been agreed on by the various stakeholders, the notion of AI is often used as an "umbrella term", encompassing broad meanings. For the purpose of this paper, the definition proposed in the European Commission Communication will be used, stating that:

> "Artificial Intelligence (AI) refers to systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals."[7]

It is important to emphasize that not all algorithmic systems and advanced digital technologies[8] represent AI. Therefore, any reference to algorithmic discrimination in this thesis will have the meaning of discrimination reinforced by AI systems.

Considering that AI can be regarded as a double-edged sword, both the beneficial and harmful potential should be carefully assessed in drafting future binding rules on AI. The existing legal framework is unable to cover all the AI implications, as the technology is evolving much faster than digital policies are implemented. Therefore, I am illustrating in my paper how the legal framework on AI could to be reformed.

1. Aims and justification of the research

This thesis aims at analyzing the impact of AI to the human rights framework. A certain emphasis will be put on tackling biases in AI, which can result in discriminatory manifestations. The purpose of AI automated decision-making is to achieve more objectivity and efficiency, that can foster

[7] European Commission, Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, 'Artificial Intelligence for Europe', Com/2018/237 Final, p. 1, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN> accessed 6 June 2020

[8] For detailed explanations of advanced digital technologies and human rights see: Karen Yeung, 'A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility within a Human Rights Framework' p. 94, accessed June 6 2020

innovation and productivity in multiple areas. However, AI can also "perpetuate and even exacerbate unfair biases". [9]

There are two main issues why such systems can lead to discrimination. First of all, the algorithmic systems and their retraining processes are perceived as "black-boxed" and cannot predict the outcome of the big data interaction.[10] Secondly, as datasets include previous biases in their training models, they consequently reproduce the societal or individual unfair value judgements.

In my paper, I have taken an in depth look at how AI is reinforcing biases and discrimination. Therefore, the research question that I address in this thesis is: How can bias and discrimination be tackled more efficiently in the AI regulatory framework, in the European Union (hereinafter: EU) and United States of America (hereinafter: US)?

The ultimate objective of the paper is to provide recommendations to decision-makers and regulatory bodies, in order to tackle the problem of AI biases and discrimination more efficiently.

I chose to pursue the analysis of two jurisdictions, respectively the EU and U.S., as they represent trend-setters in the field of AI development and deployment. Furthermore, they are amongst the first states worldwide based on the number of AI initiatives. For this reason, I comparatively evaluate their approach on regulating AI and I conclude with a set of recommendations for tackling AI biases and discrimination by regulatory intervention.

[9] Nathalie A Smuha, 'The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence', p. 2, <https://ssrn.com/abstract=3443537> accessed 6 June 2020
[10] Council of Europe (n 3), p. 10

2. Methodology and research limitations

The nature of the present study requires a qualitative research, for which I have conducted interviews with four policy advisors. Thus, I consulted one Responsible for EU Policy at a trade association representing the software industry, the Europe Policy Manager of Access Now, one Policy Advisor on Telecom and Space issues at the European Parliament and one Policy Advisor on Digital Transformation and Artificial Intelligence from Council of Europe (hereinafter: CoE). I selected the respondents in order to be able to provide different perspectives from a broad range of stakeholders involved in the policy-making process in the area of AI. Therefore, I have included the opinions of the international organization CoE, the European Parliament, involved in the co-legislation procedure in the EU, but also the private sector and the civil society.

All the four interviews were structured, including a pre-determined set of questions. The respondents received the same query, in order to ensure a possibility of analyzing the convergent or divergent directions of policy. The interviews were held between March 2020 and April 2020. The observations were audio recorded by technical means, as they were on-line video interviews. Based on the findings from the interviews and my own research of the academic literature and existing legally binding norms, I have drawn the conclusions and developed a set of recommendations.

The limitations of the research include both the fast-paced character of the AI technological advancement, which can make the findings and recommendations outdated very soon, but also the lack of sufficient academic literature on AI regulation, which can make the documentation process difficult.

3. Structure of the thesis

The first chapter draws an explanation of what AI represents. Further, it provides an overview of particular uses of AI which reinforce biases and discrimination. Afterwards, it provides an outline of the main regulatory frameworks on AI. The second chapter makes an evaluation of the most relevant legal provisions on safeguarding non-discrimination caused by AI systems, from the two jurisdictions, EU in comparison to U.S. The final chapter analyzes the challenges for implementing regulatory norms for AI, under thematic key areas, based on the illustrative interviews that I have conducted. My conclusions summarize the main key-points of the research and underline on which positions all the stakeholders I have consulted share the same opinion, but also which are the systemic differences in their approach on regulating AI. Based on my own reflections of the relevant literature and on the analysis of the interviews, I conclude by suggesting further regulatory steps that can be taken by EU and U.S. Lastly, I have included in the Annex a set of recommendations that could incentivize future policy-makers to provide more deference to human rights protection in AI regulations.

Chapter One

1. AI regulatory frameworks focusing on discrimination and biases

1.1. Introducing the concept of AI discrimination

Artificial Intelligence (hereinafter: AI) is oftentimes perceived as a broad or vague notion, but for the purpose of this paper, I will refer to AI as the capacity of some systems to collect and interpret data and to make decisions based on the selected criteria.[11]

Another concept that needs to be introduced is machine learning (hereinafter: ML), which is only a sub-set of AI, and it refers to the ability of the system to learn and to improve its performance through training and retraining processes.[12] Therefore, many times the biases and discrimination are perpetuated through ML systems.

1.2. Main discriminatory tendencies of AI systems

The focus of this research is on the multiple biases and discriminatory predispositions that occur by the use of AI in our daily lives. The issue of AI-driven discrimination represents one of the central AI critiques from a human rights standpoint and regulatory solutions were not yet implemented.

[11] European Union High Level Expert Group on Artificial Intelligence, 'A Definition of AI: Main Capabilities and Disciplines', April 2019, p. 1 <https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines>, accessed 6 June 2020

[12] Harry Surden, 'Machine Learning and Law', 2014, p. 3, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2417415>

The most serious effects of biases and discrimination led by AI systems are encountered in the criminal justice sector (especially in the U.S.), in predictive policing, facial recognition, but also in the employment sector.

Studies show how in the case of predictive policing, predominantly in the U.S., law enforcement authorities are increasingly using AI tools to determine where police forces should patrol and how to forecast criminality.[13] Using "inaccurate, skewed, or systemically biased data ('dirty data')"[14], based on historical evidence targeted against minorities or disadvantaged groups which were highly prosecuted by the police, it can lead to discretionary assessments of risk scores and legal and social inequality.[15] The AI Now research institute proves that many law enforcement agencies are using "data produced during periods of flawed, racially biased, and sometimes unlawful policing practices to train these systems". [16]

Another example from the U.S. refers to how algorithmic models are used to predict the defendants' re-offending rate, based on grounds as race, prior convictions or age. COMPAS ML tool incorrectly predicted that more black Americans would reoffend and showed the opposite predictions for the white defendants[17]. There is literature about the huge lag between Europe and

[13] Richardson et. al (n 2), p.7
[14] Ibid, p. 32
[15] Sarah Brayne, 'Big Data Surveillance: The Case of Policing' American Sociological Review 32, p. 1. <https://doi.org/10.1177/0003122417725865> accessed 7 June 2020
[16] Sarah Myers West, AI Now Institute, 'AI and the Far Right: A History We Can't Ignore', <https://medium.com/@AINowInstitute/ai-and-the-far-right-a-history-we-cant-ignore-f81375c3cc57>
[17] European Parliamentary Research Service, 'The ethics of artificial intelligence: Issues and initiatives', March 2020, p. 15

U.S. in using AI in the justice system[18], and such predictive tools are extremely rarely used in Europe, compared to U.S[19].

Other types of discrimination often occur by the use of facial recognition tools. Clearview AI[20] facial recognition system is considered to have been designed explicitly racist and is currently used by Immigration and Customs Enforcement in the U.S. [21] As well, in the past, Google image recognition software identified African-Americans as gorillas[22] and other cameras detected Asian persons as having their eyes closed, just because the filters were trained based on stereotypical appearances of only some parts of the population.[23]

AI can also discriminate on grounds as gender identity and sexual orientation. Recently, a LGBTQ group argued that YouTube's algorithm is biased against LGBTQ community and tends to restrict their content from dissemination, after suppressing their videos from the recommendation rubric.[24]

Algorithms also have sex biases, disadvantaging women: AI-based advertisements showed more well-paid jobs to males then to females.[25] As well, the Amazon AI system used for job-seekers learned to rank men higher and to downgrade female applications.[26]

[18] Serena Quattrocolo, 'An introduction to AI and criminal justice in Europe' Revista Brasileira de Direito Processual Penal. 2019;5(3):1519-1554, p. 1533, <doi:10.22197/rbdpp.v5i3.290> accessed 6 June 2020
[19] Council of Europe, European Commission for The Efficiency of Justice (CEPEJ), 'European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment', December 2018, p. 16, <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c> accessed 6 June 2020
[20] Elsevier B.V., 'Controversial Firm Clearview Gets Hacked', 2020 Biometric Technology Today 12, p. 1 <https://linkinghub.elsevier.com/retrieve/pii/S0969476520300400> accessed 7 June 2020.
[21] Myers West (n 13)
[22] Ronald Yu and Gabriele Spina Alì, 'What's Inside the Black Box? AI Challenges for Lawyers and Researchers' (2019) 19 Legal Information Management 2, p. 4 <https://www.cambridge.org/core/journals/legal-information-management/article/whats-inside-the-black-box-ai-challenges-for-lawyers-and-researchers/8A547878999427F7222C3CEFC3CE5E01> accessed 7 June 2020.
[23] European Parliamentary Research Service (n 14), p.15
[24] April Anderson and Andy Lee Roth, 'Queer Erasure: Internet Browsing Can Be Biased against LGBTQ People, New Exclusive Research Shows', 2020, 49 Index on Censorship 75, p. 1 <http://journals.sagepub.com/doi/10.1177/0306422020917088> accessed 7 June 2020.
[25] Ibid, p. 15
[26] Ibid, p. 15

Even poverty was one ground of algorithmic discrimination, as it was affirmed that poor people are under-represented in databases and the results are therefore not accurate.[27]

The research will further identify different regulatory trends to AI and non-discrimination from international organizations, EU, U.S. and private sector.

First of all, before presenting the topic, I will define specific terms which will be further mentioned. Thus, any reference to "soft law" entails that "soft law is not officially recognized (…) as legally valid law", although it "does generate rights and duties that the parties at hand".[28] By contrast, hard law is defined as "legally binding obligations that are precise and that delegate authority for interpreting and implementing the law".[29] Lastly, the notion of AI ethics refers to a "sub-field of applied ethics, focusing on the ethical issues raised by the development, deployment and use of AI".[30] Therefore, in my research I consider that AI ethics could complement hard law and could also add to soft law commitments.

1.3. Different regulatory approaches to AI and non-discrimination

1.3.1. International approach

The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems[31], a soft law agreement, provides insightful and powerful recommendations for safeguarding non-discrimination. Reaffirming the international human rights law standards, the

[27] Council of Europe (n 3), p. 12.
[28] Bart van Klink and Oliver W Lembcke, 'A Fuller Understanding of Legal Validity and Soft Law' in Pauline Westerman and others (eds), *Legal Validity and Soft Law* (Springer International Publishing 2018), p. 145 <https://doi.org/10.1007/978-3-319-77522-7_7> accessed 6 June 2020.
[29] Kenneth W. Abbott, Duncan Snidal, 'Hard and Soft Law in International Governance', p. 421-422, SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1402966> accessed 6 June 2020.
[30] EU High Level Expert Group on Artificial Intelligence, 'Ethical Guidelines to Trustworthy Artificial Intelligence', p. 11 <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419> accessed 7 June 2020
[31] The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems, May 2018,<https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/> accessed 6 June 2020

declaration calls for real solutions. Thus, designing ML systems should respect equality and non-discrimination principles and effective measures to remedy and redress should be provided.[32] Authors even consider that such "ethical regulations of AI will promote a new world order", based on the respect for human rights.[33]

Thus, the efforts of international organizations had a contribution in incentivizing global actors to address AI regulation and human rights. After the Toronto Declaration was adopted in 2018, EU issued guidelines and strategies on AI and human rights. U.S. had also similar initiatives, but not as comprehensive as those of EU.

1.3.2. The regulatory approach within the EU

The EU White Paper on AI presents the most recent strategy at EU level, dating from February 2020.[34] A human rights risk-based approach on AI is proposed, implying that some applications will be categorized as high-risk, and others as low-risk. Further, the human rights standards for such divisions will be customized accordingly. Thus, the risk-based regulatory intervention[35] could contribute to reducing AI discrimination. In assessing the possible EU Regulatory framework for AI, the White Paper discusses the possible adjustments and harmonization of existing legislation. However, particular drawbacks of the proposal include: the ambiguity in indicating the regulatory

[32] The Toronto Declaration (n 28), p. 16
[33] Bin Xu, 'Algorithm Regulation under the Framework of Human Rights Protection - From the Perspective of Toronto Declaration Academic Monograph', 2019, 18 Journal of Human Rights 495, p. 17 <https://heinonline.org/HOL/P?h=hein.journals/jrnlhmch18&i=495> accessed 7 June 2020
[34] European Commission, 'EU White paper on AI', February 2020, <https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf>
[35] Emre Kazim and Adriano Koshiyama, 'Lack of Vision: A Comment on the EU's White Paper on Artificial Intelligence' (Social Science Research Network 2020) SSRN Scholarly Paper ID 3558279 4 <https://papers.ssrn.com/abstract=3558279> accessed 7 June 2020.

approach preferred and the challenge of determining the AI applications which pose a high risk to infringe human rights.[36]

Another landmark EU document is the "Ethical Guidelines to Trustworthy Artificial Intelligence", adopted in April 2019.[37] This soft law commitment puts a certain emphasis on the non-discrimination and fairness principles. Proposed solutions for tackling discrimination and bias include implementing oversight mechanisms to overcome the systems' limitations and removing from the collection stage of the identifiable bias. Nonetheless, both input data and algorithm design should prevent the discriminatory effect of AI applications.

However, ethical guidelines lack reinforcement and oftentimes they do not influence the developers' decision-making process,[38] and thus these guidelines are not enough to safeguard fundamental rights, especially related to non-discrimination and bias.

1.3.3. The regulatory approach within the U.S.

Global inventories[39] show that U.S. issued the most initiatives on AI ethics in the world. From over 60 documents, the tech private sector represents the trend-setter, while only four instruments were adopted at national level, setting up the U.S. strategy for AI[40]. By analyzing the strategies, I the focus of U.S. on AI innovation, productivity, evolution and competitiveness, and less on the human rights centric AI.

[36] Ibid, p. 7
[37] EU High Level Expert Group on Artificial Intelligence (n 27)
[38] Thilo Hagendorff, 'The Ethics of AI Ethics: An Evaluation of Guidelines', 2020, 30 Minds and Machines 99, p. 15 <https://doi.org/10.1007/s11023-020-09517-8> accessed 7 June 2020.
[39] Yannick Meneceur, 'IA, Algorithmes, Big Data, Data Science, Robotique' <https://lestempselectriques.net/index.php/2020/05/06/ia-algorithmes-big-data-data-science-inventaire-des-cadres-ethiques-et-politiques/?fbclid=IwAR0mdPjhS1-VD4BZf7rnALut3_yrquGLGaYbjvcLrok72ByD8ztf-xtd5kg>
[40] Ibid

Nevertheless, the White House encouraged agencies to asses if regulatory intervention is needed to ensure respect for transparency and non-discrimination[41]. This assessment should consider the distinction between the regulated market and the development of new industries dominated by AI.[42]

The U.S. strategy puts emphasis on algorithms' potential bias and recommendations include the possibility of encoding "value and belief systems", in an attempt of reducing the discriminatory predisposition at the design stage of the algorithms.[43]

Interestingly, the U.S. approach seems to endorse the EU vision on "trustworthy" AI (term coined by the EU High Level Expert Group on AI), in the White House Executive Order annual report on AI[44]. Thus, the order affirms the following goal to boost AI:

> "Promote Trustworthy AI: When evaluating regulatory and non-regulatory approaches to AI, Federal agencies must consider fairness, nondiscrimination, disclosure, transparency, safety, and security."[45]

Furthermore, the Algorithmic Accountability Act of 2019[46] should be assessed in the context of algorithmic discrimination in the U.S. The law-makers proposed a bill seeking to implement

[41] White House, 'Guidance for Regulation of Artificial Intelligence Applications', January 2020, p. 11
[42] Ibid, p. 11
[43] National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee, 'The National Artificial Intelligence Research and Development Strategic Plan', October 2016, p. 26
[44] White House, Executive Order, 'American Artificial Intelligence Initiative: Year One Annual Report', February 2020,
<https://www.whitehouse.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf>
[45] Ibid, page 15
[46] US Congress, Algorithmic Accountability Act of 2019,
<https://www.congress.gov/bill/116th-congress/house-bill/2231/text>

impact assessments that would tackle bias and discrimination.[47] The justification of the initiators of this Act includes the concerns of algorithmic discrimination, biases and unfair decisions.

The high risk impact assessment of AI applications from this bill is similar to the one promoted in the EU White Paper on AI. However, there are also critiques of the proposed Act. Specifically, it only applies to automated high risk decision making, which is based on overbroad definitions, and the impact assessments are not required to be made public.[48]

However, the Algorithmic Accountability Act was neither ratified or implemented,[49] but it nevertheless addressed important aspects. From the human rights standpoint, I argue that such initiatives are needed and if passed, the Act would thus represent "the first legislative effort to regulate AI systems across industries in the U.S."[50]

1.3.4. The regulatory approach within the tech industry

The world leading tech companies have already taken a stand in developing self-regulation and codes of conduct regarding the principles of fairness and non-discrimination generated by AI systems. In the U.S., most initiatives on AI ethics are issues by the private sector[51] – Microsoft, Google, IBM, Twitter and Intel already published their approach on safeguarding human rights and how to improve their algorithmic systems to reduce bias and discrimination.

Responsibility and accountability of tech companies is needed and legal standards cannot be circumvented. Such voluntarily instruments are only complementary to legally binding norms that

[47] Margaret Jackson and Marita Shelly (eds), Legal Regulations, Implications, and Issues Surrounding Digital Data, IGI Global, 2020, p. 194 <http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-7998-3130-3> accessed 7 June 2020.
[48] Joshua New, Center for Data Innovation, "How to Fix the Algorithmic Accountability Act", September 2019, <https://www.datainnovation.org/2019/09/how-to-fix-the-algorithmic-accountability-act/>
[49] Jackson and Shelly (n 44), p. 194
[50] New (n 45)
[51] Meneceur (n 37)

should be respected. Especially as they lack enforceability and there is no oversight or control exercised by a different body on the compliance with their self-regulation, such provisions do not suffice in the absence of the already established comprehensive hard law framework.

Thus, The United Nations Guiding Principles on Business and Human Rights (UNGPs), applying to AI, aims to integrate human rights principles into business practices.[52] It thus implies a corporate responsibility to protect human rights and to address the adverse effects by prevention, mitigation and remediation.[53] Thus, placing requirements for companies to ensure adequate practices and human rights due-diligence could contribute to reducing discrimination and bias caused by AI algorithms.

However, the guiding principles have been critiqued for not offering the companies concrete guidance on how to precisely determine their human rights obligations.[54] With regard to due-diligence principle, corporations should perform human rights impact assessment (hereinafter: HRIA).[55]

To conclude with regard to the responsibility of businesses to respect human rights, UNGPs only set the ground for further development of binding rules. As a highly authoritative soft law instrument, the principles were welcomed by international community. However, I affirm that stronger commitments of the private sector on enforcement mechanisms are needed, in order to ensure fundamental rights safeguards.

[52] Mathias Risse, 'Human Rights and Artificial Intelligence: An Urgently Needed Agenda',2019, 41 Human Rights Quarterly 1., p. 9
[53] UN Human Rights Council, 'Protect, respect and remedy: a framework for business and human rights: report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises', John Ruggie, 7 April 2008, A/HRC/8/5, p. 13 <https://www.refworld.org/docid/484d2d5f2.html> accessed 7 June 2020
[54] Surya Deva, 'Guiding Principles on Business and Human Rights: Implications for Companies', p. 7.
[55] Ibid, p. 8

Chapter Two

2. Applicable legally binding framework for non-discrimination

In this chapter, I underline that there is already established a comprehensive legally binding framework to safeguard the principle of non-discrimination at a global scale. Specifically, non-discrimination and data protection legislation contain the most important provisions that are applicable to AI discrimination and biases. As this paper is centered on EU and U.S., I will further explain what are the binding norms within these two jurisdictions.

2.1. Protection against algorithmic discrimination within the EU

2.1.1. EU non-discrimination law

Enshrined as a fundamental right in the Article 21 of the "Charter of Fundamental Rights of the European Union" (hereinafter: CFR), the principle of non-discrimination benefits from a special protection under EU law.

Secondary EU law safeguards non-discrimination especially within the framework of four EU directives, as follows[56]: the Race Equality Directive 2000/43/EC protects the grounds of racial and ethnic origin, the Framework Directive 2000/78/EC prohibits discrimination on the grounds of religion or belief, disability, age and sexual orientation, the Goods and Services Directive 2004/113/EC prohibits gender discrimination and lastly, the Gender Equality Directive 2006/54/EC allows a redress mechanism in employment cases.[57]

With regard to the scope and interpretation, the Article 52 of CFR provides, inter alia, that:

[56] Philipp Philipp Hacker, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law', Social Science Research Network, 2018, p. 8, <https://papers.ssrn.com/abstract=3164973> accessed 11 May 2020

[57] Ibid, p. 9

> Rights recognised by this Charter for which provision is made in the Treaties shall be exercised under the conditions and within the limits defined by those Treaties.[58]

Thus, it is very important to acknowledge the different scope and application of the aforementioned EU Directives. For instance, the Directive 2000/78/EC only applies in the employment context, and AI discrimination not falling under this provision would not benefit from the protection of this Directive. The difference in the scope of the various EU Directives on non-discrimination may also imply consequences on different levels of protection. For example, AI discrimination based on race and sex grounds would extend the legal protection to goods and services as well.

Therefore, one possible solution could be the long standing proposal to extend the scope also for sexual orientation, age, religious and disability discrimination, which will enhance the legal protection against AI discrimination as well. Such an extension of the scope was already affirmed previously in the European Commission Proposal for a Council Directive on Implementing the Principle of Equal Treatment between Persons Irrespective of Religion or Belief Disability, Age or Sexual Orientation, but it still has not been adopted yet.[59]

Further, a crucial distinction is made in the EU law between different types of discrimination, but for the scope of the paper I will only briefly address direct and indirect discrimination. Direct discrimination is defined in the Race Equality Directive 2000/43/EC as occurring "where one person is treated less favorably than another is, has been or would be treated in a comparable situation on grounds of racial or ethnic origin"[60] and similar definitions are included in the other

[58] European Union, Charter of Fundamental Rights of the European Union, 26 October 2012, 2012/C 326/02, <www.refworld.org/docid/3ae6b3b70.html> accessed 5 June 2020

[59] European Commission, 'Proposal for a Council Directive on Implementing the Principle of Equal Treatment between Persons Irrespective of Religion or Belief Disability, Age or Sexual Orientation', COM (2008) 426 final <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52008PC0426> accessed 7 June 2020

[60] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, Article 2, para 2 (a) <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32000L0043>

three directives mentioned previously. Indirect discrimination represents that "a practice which seems neutral at first glance ends up discriminating against people of a certain ethnic origin, or another protected characteristic".[61]

Algorithmic discrimination, can be either direct, if for example, employers introduce in the system criteria to rank lower candidates of specific races, ethnicities or gender identities, but it can also be indirect, when facially neutral rules have the outcome of discriminating people on certain rules.[62] The indirect algorithmic discrimination is more problematic and harder to prove.[63] For example, when banks decide on credit worthiness and refuse a person from receiving a loan by using algorithmic systems, there is no redress or explanation about the actual reason of loan denial.[64] Therefore, it cannot be assessed if the system is denying more credits for people of certain race, ethnic origin or sex.

Another drawback of indirect discrimination is that the alleged discriminator can successfully oppose the objective justification[65] to circumvent the prohibition.

Concluding, although acknowledging the beneficial provisions set forth by EU non-discrimination law, they lack precision in all the algorithmic discrimination instances and case-law can be controversial. As algorithmic processes are not entirely transparent, their outcome decision could be discriminatory. However, if the discrimination cannot be proved, or when the objective justification is successful, the victim may not receive redress in the end. Hence, I argue that EU

[61] Frederik J Zuiderveen Borgesius, 'Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence', 2020, The International Journal of Human Rights, p. 6, <https://www.tandfonline.com/doi/abs/10.1080/13642987.2020.1743976> accessed 11 May 2020
[62] Ibid, p. 7
[63] Ibid, p. 7
[64] Ibid, p. 7
[65] Ibid, p. 7

non-discrimination law includes several critical gaps with regard to the application of AI systems, and therefore it precludes a thorough legal protection against AI discrimination.

Therefore, I support the opinion that there is a "critical incompatibility between European notions of discrimination and existing work on algorithmic and automated fairness".[66] For this reason, I conclude that there is a need to update and reform the EU non-discrimination legislation, in order to ensure legal protection for the cases of AI discrimination, which are not covered by the current EU Directives.

2.1.2. EU data protection law

Protection of personal data is a fundamental right in the EU, enshrined in the CFR and also in the comprehensive mechanism set by the EU, the General Data Protection Regulation 2016/679 (hereinafter: GDPR). The GDPR entails provisions applying to "automated decision-making", with the purpose of preventing illegal or unfair discrimination.[67]

As some authors mention, "Article 22 prohibits certain fully automated decisions with legal or similar significant effects", but the critiques are that many algorithmic decisions are not bound by GDPR or that the it does not entail a right to explanation of such processes.[68] The right to explanation would solicit to the authority or the body using AI systems to show to the alleged victim of discrimination how the decision was reached.[69] However, oftentimes either it can be hardly assessed or even it cannot be an effective safeguard during litigation.[70]

[66] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI', 2020, p. 1, <https://www.ssrn.com/abstract=3547922> accessed 17 April 2020.
[67] Borgesius (n 59), p. 9
[68] Ibid, p. 9
[69] Ibid, p. 9
[70] Ibid, p. 10

Other downside argument regarding the protection offered by GDPR to non-discrimination principle is that GDPR cannot be applied to predictive models, as it is not applied to an individual person.[71] However, the discrimination persists and it can be targeting larger groups of population.

Furthermore, GDPR is still a quite recent tool, and compliance and enforcement has not been widely assessed until now.[72] There is more need for academic research and case-law to be developed on the application of GDPR to algorithmic decisions leading to discrimination, in order to adequately conclude on the suitability of the legal instrument in mitigating bias and discrimination caused by AI algorithms.

Therefore, the EU laws on non-discrimination and data protection provide a comprehensive framework, that aims at mitigating the negative effects of algorithmic decision-making, leading to bias and discrimination. However, I argue that there are still legislative gaps that could be covered by further legally binding rules.

In order to ascertain in which sector there is a need for regulatory intervention, there is proposed a model of analysis which firstly determine the applicable binding rules, then it evaluates the risk of infringing human rights and lastly, assesses if and how the particular regulations should be adapted.[73]

2.2. Protection against algorithmic discrimination within the U.S.

The U.S. federal law prohibits discrimination caused intentionally (similar concept of direct discrimination under EU law) or through "disparate impact" (the corollary of EU indirect discrimination). The main federal anti-discrimination laws include: Civil Rights Act of 1964,

[71] Ibid, p. 11
[72] Ibid, p. 11
[73] Ibid, p. 14

Pregnancy Discrimination Act, Age Discrimination in Employment Act, Americans with Disabilities Act, Equal Pay Act or the Immigration Reform and Control Act.

There are not yet in force specific legally binding norms to mitigate bias and discrimination caused by algorithmic processes, although there have been initiatives as the proposal of Algorithmic Accountability Act of 2019, which was analyzed in the previous chapter.

Although these binding norms and the protection offered is not explicitly tailored for the use of algorithmic systems, any discrimination that occurred and was caused by AI is nevertheless protected and there are available redress mechanisms. What is actually an advantage for the targets of intentional discrimination is that when algorithms are responsible for it, the source of discrimination can be traced back, identified and proved.[74]

On the contrary, in the case of disparate, it is much harder to prove that AI systems led to discrimination or biases.

[74] Jon Kleinberg and others, 'Discrimination in the Age of Algorithms', 2018, 10 Journal of Legal Analysis, p. 2 <https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086> accessed 16 April 2020.

Chapter Three

3. Thematic analysis of interviews

For the purpose of the research, I have conducted a series of four interviews with different stakeholders involved in the process of policy-making on Artificial Intelligence (hereinafter: AI). Thus, considering that Council of Europe (hereinafter: CoE) established The Ad Hoc Committee on Artificial Intelligence (hereinafter: CAHAI) in order to assess the opportunity of regulating AI, I have interviewed Yannick Meneceur, Policy Advisor on Digital Transformation and Artificial Intelligence at CoE. Furthermore, taking into account the role of the European Parliament in the co-decision procedure and the interest of EU to regulate AI, I have interviewed Cristian Bulumc, Policy Advisor on Telecom and Space issues at the European Parliament. Nonetheless, I have included the perspective of the software industry by interviewing a Responsible for EU Policy at a trade association representing the software industry. Lastly, I consider that the contribution of the civil society in the policy-making process is crucial. Thus, the activity of the international non-governmental organization Access Now is very relevant, as it was also one of the initiators of the aforementioned Toronto Declaration and it is very involved in advocating for digital rights protection. For this reason, I have interviewed Fanny Hidvegi, the European Policy Manager of Access Now.

I have divided the responses under thematic areas of interest and I provided my own reflections interlinked with the assessment of the responses from the interviews. As mentioned previously, the interviews are used as illustrative perspectives of different stakeholders in the policy process, and they only provide additional lines of argumentation to support the claims I made based on the literature review and analysis of main regulatory frameworks.

3.1. Perspectives on establishing a regulatory framework for AI

There is an increasing concern worldwide from regulators, companies and international organizations related to the necessity of implementing a regulatory framework for AI. Therefore, the CoE set up CAHAI on September 2019, which has the clear mandate of studying the feasibility of a legal framework on AI.[75] CAHAI already started an in-depth mapping and analysis of existing legal framework, in order to identify possible gaps.[76] Thus, the activity of CAHAI will contribute to determining what could be the appropriate and suitable legal measures, which might have the form of a Convention, framework Convention, but also recommendations or guidelines.[77]

With regard to the contribution of other international bodies, the European Parliament, involved in the co-decision procedure in the EU, has recently issued several resolutions on AI.[78] The European Commission did not initiate the legislative procedure on this issue, and the opinion of experts seem to head towards a non-legally binding approach on AI.[79]

At the same time, it is important to emphasize that there are already in place protection mechanisms for fundamental rights, which are fully applicable to AI systems as well.[80] Regulating AI through a horizontal approach is not feasible or legislatively possible.[81] However, there are good practices of mapping current legal provisions on specific fields. For example, the Annex of the EU White

[75] Interview with Yannick Meneceur, Policy Advisor on Digital Transformation and Artificial Intelligence, the Council of Europe, video call, 25th of March 2020
[76] Ibid
[77] Ibid
[78] Interview with Cristian Bulumac, Policy Advisor on Telecom and Space issues, the European Parliament, video call, 24th of March 2020
[79] Ibid
[80] Interview with Responsible for EU Policy at a trade association representing the software industry, video call, 27th of April 2020. *Note: The name of the source is not publicly available, for confidentiality purposes related to commercial interest*.
[81] Ibid

Paper on AI outlines the civil liability legal standards and acknowledges the level and manner of further regulatory intervention in the case of AI.[82]

The perspective of the civil society fully supports the idea of implementing laws that are specific on a particular issue, rather than horizontal laws on AI.[83] In the process of drafting legislation, the scope of the law could not technically refer to the broad notion of AI.[84]

Therefore, there is agreement from the whole range of stakeholders that I have consulted that at this point, there is necessary to conduct comprehensive mapping and assessments of what is the existing legal framework applicable to AI and what are the legislative gaps which would require clarifications or further regulatory intervention. I believe that such approach of a moderate regulatory intervention referring to AI systems, which would only adapt or modify certain legal provisions, is the most adequate way that we should follow in the future.

Thus, I acknowledge that at this point, CoE presents the most holistic approach on AI regulation, while the tech industry and the civil society are rather in favor of a potential regulation only for specific legal areas.

3.2. The contribution of AI ethics and soft law in reducing bias and discrimination – is it sufficient?

Currently, the most numerous initiatives of introducing AI ethics principles are issued by actors representing either the private tech sector or the civil society. The illustrative responses from my interviews converge with my overall reflections that in the realm of AI applications, ethical

[82] Ibid
[83] Interview with Fanny Hidvegi, European Policy Manager, Access Now, video call, 2nd of April 2020
[84] Ibid

standards are very welcomed, and soft law principles are advancing the discussions in the policy-making process, leading to a more human rights oriented vision on regulating AI.

All the interviewees perceive the soft law mechanisms as able to implement more flexible regulation, "capable of adapting to rapidly changing environments"[85] as AI is, but warn about the possible "AI ethics washing"[86], in which such attempts would slow down the process of implementing legal frameworks.

My own conclusions fully support the idea that ethical principles of tech companies should never be used as means to override or circumvent legally binding rules. At the same time, the increasing concept of "ethics bashing"[87] (the trivialization of applying ethical principles) should not constrain attempts of establishing ethical, non-binding standards.

Unanimously, the respondents consider that the main deficiency of AI soft law is the "lack of enforcement and control".[88] Another approach suggested by all the respondents is that soft law tools could be considered to fill in the existing legislative gaps, which are not covered by hard law. One argument in the detriment of hard law is that, although the legally binding norms are enforceable and sanctions can be imposed in case of non-compliance, there is also a potential chilling effect on innovation and economic flexibility.[89]

[85] Interview with Yannick Meneceur (n 75)
[86] Ibid
[87] Elettra Bietti, 'From Ethics Washing to Ethics Bashing. A View on Tech Ethics from Within Moral Philosophy'. In Proceedings of ACM FAT* Conference (FAT* 2020). ACM, New York, NY, USA, 2019, p. 1 <https://doi.org/10.1145/3351095.3372860>
[88] Interview with Cristian Bulumac (n 78)
[89] Ibid

3.3. Role of datasets in AI discrimination

The main risk of algorithmic biases and discrimination is stemming from the data used in the system. This risk can be further reinforced by the way the program is trained or programmed[90] and "the absence of a data governance policy creates a major risk"[91] to human rights infringements. The datasets can lead to discrimination and biases through various ways: by replicating conscious or unconscious societal biases, by being under-representative in terms of population, geography etc., or by being poorly selected. Other threats are posed by "cross-referencing databases", which can lead to "making again meaningful even anonymized data", or by using such data for advertising targeting or political purposes (see Cambridge Analytica issue[92]).[93]

The private industry representative supports the principle of data minimization and being extremely careful about data quality used in algorithmic systems. When developing and deploying AI applications, it is essential to assess the quality of data introduced in the system, to ensure it is representative and not excluding any categories.[94]

Another important aspect in solving the "black-box" argument referring to the general unknown processes in which the algorithms are trained and retrained and the way in which big data sets interact, is increasing transparency.[95] This position was agreed by all respondents, affirming, among others, that also the general competition market will require high quality standards for the applications deployed.[96]

[90] Interview with Yannick Meneceur (n 75)
[91] Ibid
[92] See more about Cambridge Analytica case at: Ellen Emilie Henriksen, 'Big Data, Microtargeting, and Governmentality in Cyber-Times. The Case of the Facebook-Cambridge Analytica Data Scandal'.
[93] Interview with Yannick Meneceur (n 75)
[94] Interview with Responsible for EU Policy at a trade association representing the software industry (n 80)
[95] Interview with Fanny Hidvegi (83)
[96] Interview with Responsible for EU Policy at a trade association representing the software industry (n 80)

Although human rights error corrections can be introduced at a later stage after the system is already operational, it is crucial that developers would include from the design stage principles of transparency, non-discrimination and bias reduction.[97] However, with regard to the notion of developing applications which are "ethical by design", this is currently a technological challenge. It is instrumental to understand that although AI has the potential of accelerating bias, it can as well accelerate the fight against bias and discrimination.[98]

Concluding on this key-theme, I believe that accurate, reliable and representative datasets could have a remarkable contribution on reducing biases and discrimination reinforced by AI systems. All the respondents share the same vision on the importance of datasets. Although they proposed in some cases different solutions on this topic, their approach is not divergent. On the contrary, the various proposals are complementary and they can be successfully implemented simultaneously.

3.4. Cooperation between the public and private sector in mitigating algorithmic biases and discrimination

I analyzed the issue of cooperation between the private sector and the decision-makers in the area of AI regulation. The big tech companies are very influent in developing self-regulations related to AI applications, and thus I believe that a constant cooperation between the tech industry and the decision-makers should be promoted. The co-regulatory approach, implying that the applicable legal standards are designed and agreed on together by the industry stakeholders and regulators has multiple advantages. Thus, it can ensure liability and accountability mechanisms, and it can hamper innovation at the same time.[99]

[97] Interview with Cristian Bulumac (n 78)
[98] Ibid
[99] Ibid

The respondents unanimously agreed that a permanent consultation between the stakeholders could successfully contribute to the objective of achieving adequate legislation.

3.5. Could the EU influence the U.S. in the approach of regulating AI?

The major difference of the two jurisdictions on human rights infringements caused by AI systems is that the EU has a stronger grounding of international human rights norms than the U.S., including the European Convention on Human Rights (hereinafter: ECHR) and also the Charter of Fundamental Rights of the European Union (hereinafter: CFR).[100] Therefore, the mechanism of having a human rights court system and the clear enforceability structure are capable of providing stronger protection for human rights.[101] Consequently, I believe that the cases of algorithmic discrimination can be tackled more efficiently within the EU legal system.

While there are some divergent trends in setting up a legal framework for AI, both EU and the U.S. made a few steps in developing standards for algorithmic systems, thus aiming to minimize the human rights breaches.

Therefore, considering its human rights-centric approach on AI and the robust human rights legal system, the EU could potentially influence its competitor to raise up the human rights standards as well.[102] The first signs of influence have already been ascertained, when the U.S. Administration unofficially endorsed the term of "trustworthy AI", previously coined by the EU.[103] The AI trustworthiness argument encompasses three pillars, namely the AI system should be ethical, legal and technically robust.[104] Thus, one question that would need to be solved in the future is how to achieve the legal compliance of AI applications, in both EU and U.S.

[100] Interview with Fanny Hidvegi (83)
[101] Ibid
[102] Interview with Yannick Meneceur (n 75)
[103] Interview with Fanny Hidvegi (n 83)
[104] Ibid

All the respondents consulted agreed on the opinion that EU is promoting a human rights based approach to AI, developing as well ethical principles for AI, while U.S. is not currently providing the same level of protection, and there is also a lack of regulation on AI. Therefore, the opinions converge to the idea that EU could potentially influence and inspire the U.S. to raise human rights standards.

Concluding on the comparative approach of EU and U.S. on regulating AI, I share the same view of the interviewees, and I consider that U.S. should have in place more mechanisms to protect against algorithmic biases and discrimination.

3.6. Opportunity of setting up further regulatory frameworks

With regard to the opportunity of developing future legally binding regulations for AI, the perspective of the private sector is that regulating AI in a horizontal manner is almost impossible. Instead, we should focus on adapting the existing legislation to the technological advancement and to adopt sectorial laws suited for the technological solutions. [105] This position is as well endorsed and supported by the civil society.[106]

Consequently, considering the relevant literature and the position of experts, I argue that EU seems to promote a more interventionist approach compared to U.S. In the latter jurisdiction, the market players are dominating the AI regulatory environment by self-regulation, codes of conduct and co-regulation. This aspect raises several concerns, as now companies use self-regulation rather to address the behavior of others, instead of their own behavior.[107] At the same time, the voluntarily

[105] Interview with Yannick Meneceur (n 75)
[106] Interview with Fanny Hidvegi (n 83)
[107] Ibid

commitments of the private sector lack enforceability and accountability, and they cannot provide an increased protection for human rights.[108]

From a human rights standpoint, the EU vision for regulating AI is deeply rooted in the international human rights commitments, it imposes stricter rules for AI developers and strives to mitigate the negative effects on fundamental rights. Given the fact that U.S. already endorsed the "trustworthy" view on AI as coined by the EU, we can already perceive a sign of positive influence.

However, the competition market is often shaped by the industry players who are providing the latest and most sophisticated technological advancements. Needless to say, ensuring stronger protection for human rights leads to higher production cost, lengthier processes of development and deployment on the market, while competitors which are not bound by the same regulations would be able to launch their technology sooner, at a lesser market price and without oversight or direct repercussions in case of human rights infringements.[109]

Such concerns require also an assessment of the hypothetical global regulatory framework on AI. While the paper only focuses on EU and U.S. regulatory frameworks, there are also other states having a huge contribution on AI development. In this regard, China, another global leader on AI development, has already been critiqued for infringing fundamental rights through their AI technology. There are also voices assuming that China's future AI systems "have little inclination to solve the value alignment problem in a human rights spirit".[110]

With regard to the possibility of establishing a global regulatory framework for AI to mitigate biases and discrimination, the opinion of all interviewees was unanimous – such objective seems

[108] Interview with Cristian Bulumac (n 78)
[109] Interview with Responsible for EU Policy at a trade association representing the software industry (n 80)
[110] Risse (n 47), p. 10

to be overreaching, but there are in place already commonalities in the approach of regulating AI, at least at the EU and U.S. levels.

In this sense, the representatives of the European Parliament, CoE and private sector consider that a global framework for AI might be achieved. However, such agreements would have a low legal impact and they would only support the non-controversial aspects of AI regulation regarding human rights. In practice, this could not be an effective tool of enforcing human rights protection, but rather it could only represent an international commitment on future AI developments.

On this issue, the civil society representative had a distinct opinion, considering that a common global framework for AI is unlikely to be achieved. Instead, an appropriate objective should be to apply the framework of international human rights worldwide.[111]

[111] Interview with Fanny Hidvegi (n 83)

Conclusion

I have analyzed in my research the feasibility of implementing a regulatory framework for Artificial Intelligence (hereinafter: AI) in two jurisdictions, the European Union (hereinafter: EU) and the United States of America (hereinafter: U.S.). The final objective was to ascertain which measures could be appropriate for tackling biases and discrimination caused by the interference of AI systems. For this purpose, I have conducted a mapping of the most common discriminatory tendencies and I acknowledged the multiple grounds on which algorithmic discrimination can occur, including: race, ethnicity, sex, gender, sexual orientation, age or poverty.

The two main reasons of perpetuating biases and discrimination in algorithmic systems are the biased, flawed or non-representative datasets used by the algorithm, and also the opaque models of training and retraining of the algorithms, which can replicate conscious or unconscious human biases.

While AI algorithms can perpetuate and reproduce societal biases and lead to discrimination, they can also be used to minimize human biases in decision-making processes. Moreover, AI has the capacity to provide more objective, accurate and scalable assessments in various areas of our life, from criminal justice, to healthcare and employment.

Further, I have outlined the regulatory approaches for AI within the EU and the U.S., but I have also included the initiatives undertook by international organizations and by the tech industry. Thus, I concluded that EU has already set up the grounds for further regulatory intervention, through adopting the EU White Paper on AI and also the Ethical Guidelines to Trustworthy AI. With regard to the U.S., there is more need to implement legal strategies at federal level for AI and human rights. At the international scale, I highlighted the relevance of the Toronto Declaration on

algorithmic discrimination, and the need to follow the recommendation enshrined in it. Lastly, I drew the attention on the practices of companies which are developing AI systems. Thus, the business sector has influence in embedding principles of non-discrimination and transparency in building their products. Nonetheless, I stress the desideratum that companies should respect The United Nations Guiding Principles on Business and Human Rights (hereinafter: UNGPs).

Moreover, I have assessed the existing legal framework providing protection against AI discrimination and biases. Thus, the most comprehensive safeguards are offered through the provisions of non-discrimination law and data protection law. However, I outlined several flaws in both frameworks of non-discrimination and data protection, which allows AI to circumvent the legal compliance in certain cases. Therefore, the scope of non-discrimination legislation could be extended to cover algorithmic discrimination as well, or new legal instruments can be adopted. Similarly, data protection laws could be adjusted to extend their protection to the issues that are now not bound to respect data protection legislation.

Moreover, a central part of my research has focused on conducting interviews with four professionals involved in the area of AI policy-making. I have selected the respondents in order to maintain a fair balance between all the stakeholders involved, that would ensure that my findings are reliable and accurate. I have analyzed the interviews thematically and I have outlined the key-points from their opinions. Hence, I am able to present the takeaways based on the consultations that I have pursued.

First of all, all the respondents agreed that there is more need for undertaking a comprehensive mapping of the impact of AI to human rights, in order to acknowledge what would be the level of regulatory intervention needed and what would be the most suitable legal instrument to tackle

biases and discrimination led by AI. Although there is ongoing debate on the certain legal form that could be adopted in EU and U.S., the position of the respondents converges to the idea that it would not be possible to have a horizontal law on AI, but rather sectorial laws, applied for specific legal areas. Thus, I endorse the view that sectorial regulation could mitigate AI biases and discrimination.

Another topic where agreement was reached by the respondents is related to the benefits of implementing AI ethics and soft law commitments. However, as their shared vision emphasizes, although such initiatives provide more flexibility, the pitfalls include their lack of enforceability and control in cases of non-compliance. Thus, the efficiency and effectiveness of AI ethics and soft law are not sufficient.

Further, I acknowledged a shared opinion from the interviewees on the importance of the cooperation of all the stakeholders involved in the AI policy-making process, in order to ensure that discrimination and biases are mitigated in further regulations, without the risk of hampering innovation and productivity.

Nevertheless, a very important recommendation for reducing AI discrimination and biases is to establish that datasets used by algorithms are reliable, accurate and representative and that they are not biased. The respondents had distinct proposals for achieving this objective of fair and adequate datasets, including data selection, error corrections, minimization of data or promoting a data governance system. However, these concrete solutions are not divergent. On the contrary, I believe they are complementary and the simultaneous application of all measures is not excluded.

With regard to the concrete challenges and differences on the approaches on regulating AI in EU and U.S., the essential takeaway agreed by all respondents is that EU has set in place a more

comprehensive framework for human rights, which already provides stronger safeguards against algorithmic discrimination. Therefore, the recommendations made by the consulted stakeholders include the way in which U.S. can be inspired by the human rights centric approach of EU in further regulatory developments on AI systems.

Lastly, on the hypothetical opportunity of having a global framework for AI, the majority of respondents considered that such an objective is not necessarily impossible to achieve, but it would have not a significant legal impact. The distinct view came from the representative of the civil society, which believes that a global framework should not be an objective in itself, but rather international human rights law should be applicable worldwide.

It is extremely important to emphasize that technology, including AI, is in itself ethically neutral. Its malicious or beneficial applications are entirely dependent on the way in which AI is applied and implemented in practice. AI might contribute to another technological revolution, implying a set of opportunities and challenges at the same time. As we live in a time of technological transformation, legal scholars should work together with policy-makers, technical community and civil society, to agree on how and to what extent AI should be regulated in the future, in the EU and in the U.S., but also worldwide.

Annex

Legal policy recommendations

1. Conduct a comprehensive mapping of the legal frameworks impacted by AI

This research has led to the conclusion that, at this moment, there is no global or regional consensus on which aspects of AI would require adopting further regulations and would be the most adequate manner of the intervention. Moreover, a horizontal legally binding framework for all AI applications is not appropriate or legislatively feasible, since the various AI systems cannot be incorporated under a single unitary framework.

Therefore, I strongly support the opinion indicated as well by the respondents of the interviews, stating that currently, there is a crucial need for acknowledging the interference of AI with human rights, based on the existing legal framework. Only after a thorough assessment of particular legal sectors impacted by AI, regulators could decide to take further legislative actions.

2. Advance the debate on AI ethics

The debate on AI ethics ought to be grounded in the international human rights law. The principles of non-discrimination and equality and are already enshrined in the international human rights law, as well as in the EU and the U.S. legislations. Therefore, I believe that AI ethics complements hard law and add to soft law commitments, aiming to ensure safeguards mechanisms against algorithmic discrimination.

This research concludes that is desirable from the international actors to advance debates and initiatives promoting AI ethics, as this could meaningfully incentivize regulators to take into consideration these ethical principles when developing further hard law in the area of AI.

Therefore, regulators will have the difficult task of maintaining and constantly adjusting the digital equilibrium, in order to adapt the legal provisions for safeguarding human rights and for offering equal chances to the market players.

3. Determine the adequate legal instruments for AI regulation

The pitfalls of a horizontal regulatory framework for AI have been discussed in the previous chapters, leading to the conclusion that in practice, a single legislative instrument could not tackle the human rights infringements caused by AI applications. Moreover, AI overregulation would be detrimental because the imposition of excessively severe conditions on the private sector will hinder innovation and the evolution of digital economy.

While the EU published its strategy on regulating AI in the EU White Paper in February 2020, the U.S. has still not adopted a similar strategy on regulating AI, to prevent human rights infringements.

Therefore, the decision-makers could determine what is the most suitable legal instrument to tackle the issue they would like to solve. However, this step could only take place after fulfilling the first recommendation of conducting a comprehensive legal mapping on AI interferences.

3.1. Implement sectorial regulations, especially for high-risk AI applications

Based on my research and the interviews that I have conducted, I conclude that one possible regulatory intervention could be to adopt sectorial regulations, aimed at filling the gaps of existing hard law on non-discrimination and data protection vis-à-vis AI. Thus, in the cases of AI applications that pose a high risk of reinforcing discrimination, regulators could set specific rules on liability and accountability. For this reason, a thorough assessment of which AI applications should fall under the high risk category should be conducted.

For example, in areas as criminal justice, predictive policing, and facial recognition, it is possible to conceive of tailored regulations to ensure an increased transparency and fairness of the systems, in order to reduce potential biases and discrimination. Other sectoral laws could be adopted in employment, healthcare, and credit worthiness sectors.

3.2. Fill in the gaps of non-discrimination law

Another recommendation could be to update the current legislation prohibiting discrimination and to customize its application in cases of algorithmic discrimination. One viable option could be to extend the scope of the protection provided through non-discrimination laws, to cover the infringements caused by AI.

In the EU, such an extension could be operated either by adopting a distinct EU directive on algorithmic discrimination, for example, or to adjust the application of the EU anti-discrimination directives, in order to provide legal safeguards against the novel type of AI discrimination. Similarly, the U.S. could modify its own legislation, to tackle the particular cases of algorithmic discrimination.

3.3. Increase safeguards provided by data protection law against AI discrimination

As indicated previously in the analysis, there are several flaws in the data protection legislation, that lead to the circumvention of the application of data protection legal framework in several cases of AI decision-making processes. Consequently, there are instances when AI systems reinforce biases or discrimination, but the victims cannot benefit from redress mechanisms provided by data protection laws.

Therefore, regulators could opt to address in the future legislative interventions the situations which are currently not covered by data protection framework.

4. Ensure representative and high-quality datasets

There was a shared opinion of all the respondents consulted in the interviews that one of the main discriminatory tendencies occurs when algorithms are trained using "unrepresentative, flawed, or biased data". [112]

There should also be set in place legal policies to ensure the use of representative datasets, respecting the principles of "accuracy, consistency and validity"[113] in selecting the input data. From the design and programming stage, developers should follow legal provisions that state the necessity of including representative, accurate and unbiased datasets.

Hence, based on the proposals of all the interviewees, I conclude that, in order to reduce the risks of discrimination and biases in algorithmic decision-making, further regulation could focus on ensuring that datasets are representative and high-qualitative.

5. Strengthen the cooperation between all the stakeholders involved in policy-making act

Regulation in the area of technology is oftentimes leading not only to legal and policy implications, but also to economical and ethical ones.

I consider that private companies should be incentivized to embed principles of transparency, fairness and non-discrimination while developing their AI products on the market, not only in the

[112] Fjeld et al., 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI', January 2020, p. 26, <https://ssrn.com/abstract=3518482> accessed 6 June 2020
[113] Ibid, p. 27

EU and the U.S., but at a global scale. As well, efforts should be made to bridge the gap between technical community and regulators, to ensure their permanent collaboration.

Therefore, I conclude that a strong cooperation between all the stakeholders involved in the policy-making process, including the public sector and the tech industry, but also the academia and the civil society, could achieve the final objective of implementing legislation able to surpass the challenges posed by rapid digital transformation.

Bibliography

Abbott, K and Snidal, D, 'Hard and Soft Law in International Governance SSRN' <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1402966> accessed 6 June 2020

Anderson A and Roth AL, 'Queer Erasure: Internet Browsing Can Be Biased against LGBTQ People, New Exclusive Research Shows' (2020) 49 Index on Censorship 75 <http://journals.sagepub.com/doi/10.1177/0306422020917088> accessed 7 June 2020

Bietti E, 'From Ethics Washing to Ethics Bashing. A View on Tech Ethics from Within Moral Philosophy'. In Proceedings of ACM FAT* Conference (FAT* 2020). ACM, New York, NY, USA, 2019, p. 1 <https://doi.org/10.1145/3351095.3372860>

Bulumac C, Policy Advisor on Telecom and Space issues, the European Parliament, Interview, video call, 24th of March 2020

Brayne S, 'Big Data Surveillance: The Case of Policing' American Sociological Review 32

Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32000L0043>

Council of Europe a, 'Discrimination, Artificial Intelligence, and Algorithmic Decision-Making'

Council of Europe b, European Commission for The Efficiency of Justice (CEPEJ), 'European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment', December 2018 <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c> accessed 6 June 2020

Deva, S, 'Guiding Principles on Business and Human Rights: Implications for Companies SSRN' <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2028785> accessed 9 May 2020

European Commission, Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, 'Artificial Intelligence for Europe', Com/2018/237 Final, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN> accessed 6 June 2020

European Commission, 'EU White paper on AI', February 2020, <https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf>

European Commission, 'Proposal for a Council Directive on Implementing the Principle of Equal Treatment between Persons Irrespective of Religion or Belief Disability, Age or Sexual Orientation', COM (2008) 426 final <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52008PC0426> accessed 7 June 2020

European Parliamentary Research Service, 'The ethics of artificial intelligence: Issues and initiatives', March 2020

European Union, Charter of Fundamental Rights of the European Union, 26 October 2012, 2012/C326/02, <www.refworld.org/docid/3ae6b3b70.html> accessed 5 June 2020

European Union High Level Expert Group on Artificial Intelligence, 'Ethical Guidelines to Trustworthy Artificial Intelligence' <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419>

European Union High Level Expert Group on Artificial Intelligence, 'A Definition of AI: Main Capabilities and Disciplines', April 2019, <https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines>, accessed 6 June 2020

Elsevier B.V. 'Controversial Firm Clearview Gets Hacked' (2020) 2020 Biometric Technology Today 12 <https://linkinghub.elsevier.com/retrieve/pii/S0969476520300400> accessed 7 June 2020

Fjeld et al., 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI', January 2020, <https://ssrn.com/abstract=3518482> accessed 6 June 2020

Hidvegi F, European Policy Manager, Access Now, Interview, video call, 2nd of April 2020

Hacker P, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law' (Social Science Research Network 2018) SSRN Scholarly Paper ID 3164973 <https://papers.ssrn.com/abstract=3164973> accessed 11 May 2020

Hagendorff T, 'The Ethics of AI Ethics: An Evaluation of Guidelines' (2020) 30 Minds and Machines 99 <https://doi.org/10.1007/s11023-020-09517-8> accessed 7 June 2020

Henriksen EE, 'Big Data, Microtargeting, and Governmentality in Cyber-Times. The Case of the Facebook-Cambridge Analytica Data Scandal'

Jackson M and Shelly M (eds), Legal Regulations, Implications, and Issues Surrounding Digital Data: (IGI Global 2020)

<http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-7998-3130-3> accessed 7 June 2020

Kazim E and Koshiyama A, 'Lack of Vision: A Comment on the EU's White Paper on Artificial Intelligence' (Social Science Research Network 2020) SSRN Scholarly Paper ID 3558279 <https://papers.ssrn.com/abstract=3558279> accessed 7 June 2020

Kleinberg J and others, 'Discrimination in the Age of Algorithms' (2018) 10 Journal of Legal Analysis 113 <https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086> accessed 16 April 2020

Klink B and Lembcke OW, 'A Fuller Understanding of Legal Validity and Soft Law' in Pauline Westerman and others (eds), Legal Validity and Soft Law (Springer International Publishing 2018) <https://doi.org/10.1007/978-3-319-77522-7_7> accessed 6 June 2020

Myers West S, AI Now Institute, 'AI and the Far Right: A History We Can't Ignore', <https://medium.com/@AINowInstitute/ai-and-the-far-right-a-history-we-cant-ignore-f81375c3cc57

Meneceur Y a, 'IA, Algorithmes, Big Data, Data Science, Robotique : inventaire des cadres éthiques et politiques' (*Les Temps électriques*, 6 May 2020) <https://lestempselectriques.net/index.php/2020/05/06/ia-algorithmes-big-data-data-science-inventaire-des-cadres-ethiques-et-politiques/> accessed 9 May 2020

Meneceur Y b, Policy Advisor on Digital Transformation and Artificial Intelligence, the Council of Europe, Interview, video call, 25th of March 2020

National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee, 'The National Artificial Intelligence Research and Development Strategic Plan', October 2016

New J, Center for Data Innovation, "How to Fix the Algorithmic Accountability Act", September 2019, <https://www.datainnovation.org/2019/09/how-to-fix-the-algorithmic-accountability-act/>

Quattrocolo S, 'An introduction to AI and criminal justice in Europe' Revista Brasileira de Direito Processual Penal. 2019;5(3):1519-1554 <doi:10.22197/rbdpp.v5i3.290> accessed 6 June 2020

Raso F and others, 'Artificial Intelligence & Human Rights: Opportunities & Risks' [2018] SSRN Electronic Journal <https://www.ssrn.com/abstract=3259344> accessed 4 May 2020

Responsible for EU Policy at a trade association representing the software industry, Interview, video call, 27th of April 2020

Richardson R, Schultz J and Crawford K, 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data', Predictive Policing Systems, and Justice, February 2019 <https://ssrn.com/abstract=3333423> accessed 6 June 2020

Risse M, 'Human Rights and Artificial Intelligence: An Urgently Needed Agenda' (2019) 41 Human Rights Quarterly 1

Smuha NA, 'The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence' 10

Surden H, 'Machine Learning and Law', 2014, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2417415>

The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems, May 2018, <https://www.accessnow.org/the-toronto-declaration-

protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/> accessed 6 June 2020

UN Human Rights Council, 'Protect, respect and remedy: a framework for business and human rights: report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises', John Ruggie, 7 April 2008, A/HRC/8/5, <https://www.refworld.org/docid/484d2d5f2.html>

US Congress, Algorithmic Accountability Act of 2019, <https://www.congress.gov/bill/116th-congress/house-bill/2231/text>

Wachter S, Mittelstadt B and Russell C, 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI', 2020, SSRN Electronic Journal <https://www.ssrn.com/abstract=3547922> accessed 17 April 2020

White House, Executive Order, 'American Artificial Intelligence Initiative: Year One Annual Report', February 2020, <https://www.whitehouse.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf>

White House, 'Guidance for Regulation of Artificial Intelligence Applications', January 2020

Xu B, 'Algorithm Regulation under the Framework of Human Rights Protection - From the Perspective of Toronto Declaration Academic Monograph' (2019) 18 Journal of Human Rights 495 <https://heinonline.org/HOL/P?h=hein.journals/jrnlhmch18&i=495> accessed 7 June 2020

Yeung K, 'A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility within a Human Rights Framework' 94

Yu R and Alì GS, 'What's Inside the Black Box? AI Challenges for Lawyers and Researchers', 2019, 19 Legal Information Management 2 <https://www.cambridge.org/core/journals/legal-information-management/article/whats-inside-the-black-box-ai-challenges-for-lawyers-and-researchers/8A547878999427F7222C3CEFC3CE5E01> accessed 7 June 2020

Zuiderveen B, 'Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence', 2020, The International Journal of Human Rights 1 <https://www.tandfonline.com/doi/full/10.1080/13642987.2020.1743976> accessed 11 May 2020