UNIVERSITÀ DEGLI STUDI DI PADOVA

DEPARTMENT OF POLITICAL SCIENCE, LAW,
AND INTERNATIONAL STUDIES

Master's degree in
Human Rights and Multi-level Governance

Artificial intelligence governance in international relations:

a human rights perspective

Supervisor: Prof. Pietro de Perini

Candidate: Katarina Stanisavljevic

Matriculation No.: 2046767

A.Y. 2023/2024

"Thus every organic body of a living being is a kind of divine machine or natural automaton, infinitely surpassing all artificial automatons. Because a man-made machine is not a machine in each of its parts; for instance, the teeth of a brass wheel have parts or bits which to us are not artificial products and contain nothing in themselves to show the use to which the wheel was destined in the machine. However, the machines of nature, that is to say, living bodies, are still machines in their smallest parts *ad infinitum*. Such is the difference between nature and art, that is, between divine art and ours."

Gottfried Wilhelm Leibniz

Acknowledgments

I would like to express my sincere gratitude to Prof. Pietro de Perini for his patience, invaluable advice, and professionalism, which have greatly contributed to this process. Your way of teaching has been a source of inspiration for many students, including myself.

I would also like to extend my heartfelt gratitude to my parents, my brother, my family and my loved ones, who, despite being hundreds of kilometers away, have unconditionally supported me and believed in me every step of the way on this adventure and encouraged me to surpass my own expectations. Without you, none of this would be possible.

Additionally, I am deeply thankful to the incredible people I met on this journey, from whom I learned so much and who will forever be my inspiration and part of the most cherished memories of my life. Thank you for your infinite love, warmth and kindness. You made me feel at home.

Further, I want to thank the University of Padova for providing me with an opportunity that has profoundly impacted both my life and academic career.

Finally, I dedicate this accomplishment to my mother, who went above and beyond to provide me with this opportunity and supported me every step of the way. Thanks to you, I've turned a dream into an achievement.

Table of Contents

Abstract

Advancements in artificial intelligence (AI) technology are rapidly transforming global governance structures and institutions, including key bodies responsible for promoting and protecting human rights and putting AI as one of the key topics on the international agenda. However, the implications of AI for the effectiveness of international human rights mechanisms remain unclear since the development of AI regulations is still in its initial phase. This thesis aims to investigate the role of AI in shaping global governance and its potential impact on the efficacy of human rights institutions worldwide. The thesis will explore international efforts towards human rights-centric AI governance and current ethical and human rights considerations. This thesis framework can be intended as a starting point in recognizing the importance of international cooperation in AI governance and its significance for the protection of human rights.

INTRODUCTION

In 1950, the British mathematician Alan Turing, influenced by the increasing significance of machines, and his own experience during World War II in cracking the German encryption device Enigma, asked the question: "Can machines think?" Today, the answer to this question still remains "no", since a machine hasn't yet passed the Turing test (originally known as the imitation game), as it still doesn't possess the capability to display intelligent behaviour that is equivalent to, or indistinguishable from, that of a human (Bussler 2020).

Today, the world has come a very long way and the rapid development of technology in the last decades has had a tremendous impact on international relations and policy-making for protecting human rights. Technology now allows international actors and governments to process large volumes of data to face contemporary challenges such as pandemics and climate change, it enables real-time reporting of human rights abuses, predicting global economic trends, and even assisting in conflict resolution. Data analysis and forecasting algorithms enable governments and international organizations to respond more efficiently to complex situations. However, implementing technology in international relations and policy-making, especially artificial intelligence, can also raise questions about transparency, discrimination, fairness, and possible consequences for the global society.

The main concerns around AI from a fundamental rights perspective, are about transparency, bias, accountability, discrimination and the right to privacy and data protection (Mazzini, 2019). AI is developing in a manner that is hard to catch up with, especially when law is in question. The long administrative processes take a lot of time while the development of AI comes with almost daily updates, making it hard to regulate its ever-changing possibilities. The evolution of technology, mainly in the area of artificial intelligence, has attracted the attention of the public mostly because of its work-related benefits. Programs such as ChatGPT and similar AI tools have quickly been incorporated into various programs and applications and changed the traditional way of working. Although there are many concerns about its ethical aspect, big technological companies are using the "slowness" of IR actors in regulating policy around AI and gaps in law provisions since many issues are regulated on an ad-hoc basis.

The objective of this thesis is to detect the complex interplay of international actors and understand AI in the context of protecting human rights in the international arena, highlighting the necessity of a comprehensive global governance approach. Therefore, the main research question of this thesis is: *"How are international actors addressing the regulation of AI in a way that does not unnecessarily restrict its development, while simultaneously upholding core values of international humanitarian law and fundamental rights protection?"*. To shed more light on this question it's important to approach it from different angles, particularly, by analyzing how technology is affecting human rights, how international actors tackle this issue, in which areas it is used and what are the risks and benefits for human rights imposed by AI. To analyze these interactions and impacts, the research will examine various ways in which human rights intersect with AI through case studies and real-life examples. This includes detecting the main principles of human rights and examples of how they are positively or negatively affected by AI technology. Further, by examining cases in which AI technology has been deployed to enhance human rights protection mechanisms, such as human rights monitoring, conflict prediction, and climate change and its societal impact. Additionally, the research will explore the initiatives taken by different international actors, such as EU bodies and international organisations, to regulate AI. To do that, different acts, policy documents and debates will be assessed to include a comparative analysis of different approaches and strategies, with special attention paid to the examples of challenges and successes of international cooperation.

The academic significance of this research is that it contributes to the expanding body of literature in a relatively new academic intersection – between artificial intelligence and human rights. Currently, the literature on this topic is limited, which highlights the need for further exploration of best practices and areas needing additional attention. This research offers a comprehensive analysis of existing international AI policies, providing insights into current practices, gaps, and opportunities for cooperation between international actors and contributing to a better understanding of how this constantly evolving field is governed globally. In light of international efforts for global governance of AI, this research draws attention to aligning AI technology development with human rights standards.

This research also carries societal and practical significance, by raising awareness of the fragility of human rights in emerging AI technologies and raising awareness on the importance of keeping

international actors and technology companies accountable. Further, it emphasizes the need for a balanced and joint approach to technological innovation and regulation.

In this study, the methodology includes a mixed-method approach by analyzing the data comprising both primary and secondary sources. The primary sources include documents issued by European authorities and other international actors, such as international organisations and advisory bodies, which include regulations, recommendations, guidelines, white papers, international declarations and resolutions. More specifically, for the analysis of the European AI policy, the documents analyzed were the General Data Protection Regulation, The EU White Paper on Artificial Intelligence, the EU Artificial Intelligence Act and several other initiatives. In terms of other international actors' work, the United Nations' reports and resolutions on AI were examined, along with other policy documents and initiatives of the Organisation for Economic Cooperation and Development (OECD), UNESCO and Council of Europe. These primary documents provide an overview of the existing regulatory frameworks and policies. Additionally, to complement the primary sources, secondary data was gathered from academic research papers, news articles and reports by international organisations, human rights organisations and prominent news organizations specializing in European Union affairs and politics such as Politico and Euractiv. The motivation behind this mixed approach is to ensure a comprehensive examination of AI regulations, their impact on human rights and the international actors' positions, providing a strong basis for the analysis.

The first chapter of the thesis provides a review of the biggest challenges to human rights by artificial intelligence systems such as which is crucial for understanding the challenges and risks imposed on human rights by AI technologies. This chapter approaches the principles of transparency, oversight, accountability, responsibility, privacy and data protection. Further, the second part of the chapter examines the possible misuse of AI technology by shedding light on autonomous weapons systems (AWS) and deepfakes. Throughout the chapter, these principles and applications were examined by defining them within international regulatory frameworks, showing their applicability through examples of their use and providing recommendations for their enhanced regulation.

The second chapter of the thesis focuses on the overview and analysis of existing policy frameworks that are formulated by international actors. Special attention is paid to highlight the

way these policies protect human rights through AI regulation. Additionally, where sources allow, it provides examination of the political positions of different international actors during the processes of negotiation and the impact on their practice. The chapter starts with an analysis of the policies of the EU and continues with international actors, with the goal of providing a comprehensive comparison of different regulatory approaches across the world and their effectiveness in the protection of human rights.

In the third chapter, the aim is shifting to the opposite side of the understanding of Artificial Intelligence in the protection of human rights – the use of AI for good. This chapter focuses on the positive effects AI can have on human rights and how that is achieved in the international arena. This refers to various human rights protection mechanisms, human rights monitoring and reporting, violent conflict prediction, early warning systems and climate change. These are all areas in which human rights need to be protected and that has shown to be more efficient with the help of AI tools. This chapter also includes numerous case studies that show the benefits AI can provide and how international actors have cooperated in order to combine human rights expertise with technological expertise in the protection of human rights. This chapter ends with highlighting all the opportunities that AI is providing to the global society and what are how AI can help the global society to strive.

Finally, the conclusion chapter will offer a comprehensive response to the main research question by highlighting the main findings of this research and suggesting avenues for future research on the topic.

CHAPTER I: HUMAN RIGHTS UNDER THREAT - CATEGORIZING AI'S CHALLENGES

This chapter aims to delve into the primary challenges artificial intelligence is posing on human rights and how the international community is addressing those issues. Considering that there are numerous risks and challenges to human rights imposed by artificial intelligence, in the selection of the ones that would be analyzed I drew upon the world's first binding treaty on AI: *the Framework Convention on Artificial Intelligence and human rights, democracy, and the rule of law*, which was adopted on May 17, 2024 by the Council of Europe (Council of Europe, 2024). In the treaty the principles of transparency and oversight, accountability and responsibility and privacy and personal data protection were mentioned as the most important principles necessary for safeguarding human rights in AI systems (Council of Europe, 2024). Further, in the United Nations Common Agenda Policy Brief 5 - *A Global Digital Compact — an Open, Free and Secure Digital Future for All*, UN Secretary-General António Guterres has said that "the international community should make transparency, fairness and accountability the core of AI governance and consider the adoption of a declaration on data rights that enshrines transparency" (United Nations, 2023).

Besides the examination of the challenges imposed on human rights, two examples of misuse of the AI system are included, specifically the Automated Weapons Systems (AWS) and Deepfakes, which show extreme ways in which the violations of human rights can be conducted through the use of AI, and can represent threats to all fundamental human rights, including the right to live.

In order to have a basic understanding of AI necessary for this study, it's important to acknowledge that because of its complexity, the unique definition of it still doesn't exist. Therefore, a couple of different definitions will be provided in this chapter, but also in the succeeding one, where the regulation of AI will be analyzed based on the approach of each international actor. Most of the definitions can be simplified to define AI as the ability of a machine to act like a human. For example, Artificial Intelligence can be defined as "the capability of a machine to imitate intelligent human behavior, including learning, problem-solving, and decision-making" (Russell and Norvig 2016). Further, according to Organisation for Economic Co-operation and Development, Artificial Intelligence refers to "systems that display intelligent behavior by analyzing their environment and taking actions – with autonomy to

achieve specific goals" (OECD 2019b). Further, UNESCO defines Artificial Intelligence as “the ability of machines to exhibit human-like intelligence and perform tasks that traditionally require human cognitive function”. (UNESCO 2021). Moreover, while the EU has included many different definitions of AI in its communications, which will be discussed in the next chapter, in the latest EU AI ACT the EU defines AI systems as: “a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments” (European Commission, 2021).

The chapter is structured as follows: firstly the analysis of the principles of transparency and oversight, accountability and responsibility, and privacy and personal data protection are assessed. Further, the examination delves into the risks of AI misuse, particularly AWS and Deepfakes.

1. Challenges

The aim is to examine the primary challenges imposed on these principles, highlighting the legal and ethical challenges posed by AI, discussing the necessity of holding entities responsible for decisions and actions done by AI, and emphasizing the risks the misuse can pose on human rights. These principles can be considered crucial for several reasons. Firstly, during the literature review, these principles were often depicted as interdependent principles, often described as unable to exist in isolation from one another. Transparency is considered a prerequisite for accountability, privacy is saved through the existence of transparency and accountability, transparent processes allow users to have insight into how their data is protected and in violation of those principles accountability ensures that violations are addressed and the responsible ones are held accountable.

This interdependence is also thoroughly explained in The AI Risk Management Framework from the U.S. National Institute for Standards and Technology which says that "Trustworthy AI greatly depends upon accountability. Accountability presupposes transparency. Therefore, transparency reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system. Meaningful transparency provides access information based on the stage of the AI lifecycle and tailored to the role of AI actors or individuals

interacting with or using the AI system" (U.S. National Institute for Standards and Technology, 2023).

1.1 Transparency and oversight

The lack of algorithmic transparency and oversight in Artificial Intelligence systems represents a significant issue from a legal perspective (Rodrigues, 2020). Today, AI plays a major role in our lives whether we want that or not, imposing on us both numerous benefits and risks. Therefore, opacity poses a significant challenge to human rights. As an example of how lack of transparency affects people's lives, Rodrigues highlights "examples of people who were denied jobs, refused loans, were put on no-fly lists or denied benefits without knowing why that happened other than the decision was processed through some software" (Rodrigues, 2020). Artificial intelligence is also becoming an integral part of our economies, and yet the algorithms used are neither easily understandable nor accessible to the public. AI systems should be understandable so we can grasp how decisions are made and ensure they are carried out in a fair, unbiased and ethical way. Although transparency already has its limitations, information about the functionality of algorithms is frequently deliberately obscured (Rodrigues, 2020). Therefore, one of the main challenges regarding transparency is its opacity in the decision-making algorithms (Rodrigues, 2020).

In April this year, an open letter demanding transparency in AI was signed by the current and former employees of the biggest technology companies, including OpenAI, Google DeepMind and Anthropic (Hilton et al. 2024). This group warned about the dangers of opacity and lack of oversight, emphasizing that these companies receive "strong financial incentives" in order to avoid transparency (Hilton et al. 2024). Further, they warned about the risks of deepening already existing inequalities, the high risk of manipulation and misinformation and even "the loss of control of autonomous AI systems potentially resulting in human extinction " (Hilton et al. 2024). They claim that as long as there is no impactful government oversight, only the current and former employees are the ones who can hold these companies accountable to the public (Hilton et al. 2024).

A study conducted in 2023 by researchers from Stanford University, Massachusetts Institute of Technology, and Princeton University, titled "The Foundation Model Transparency Index," has shown alarming deficiencies of transparency from 10 major technology companies such as

OpenAI, Google, and Meta (Bommasani et al. 2023). The results of the research were based on a transparency scoring model in which the researchers came to the conclusion that there is a widespread lack of transparency in the AI ecosystem. Specifically, based on the *Foundation Model Transparency Index* created by the researchers, none of the companies had a score higher than 60% out of the possible 100% of transparency, proving that these companies do not share enough information about the functioning of the use and development of their models (Bommasani et al. 2023).

Further, even though the governments that are home to these companies are taking measures to promote transparency, the research concluded that none of the companies shares appropriate information on copyrighted data it uses, on how the model might be abused or on the environmental impact AI systems produce (Bommasani et al. 2023). For example, many civil society organisations have advocated for more comprehensive transparency provisions in the EU AI ACT, among whom the AccessNow, Algorithm Watch, and the European Center for Not-for-Profit Law (Bommasani et al. 2023). Further, Freedom House warned that AI allowed governments to enhance their online censorship which poses a risk to a global decline in internet freedom, which is only further enabled by these non-transparent AI models (Bommasani et al. 2023). Additionally, they proved that on more than 80 out of 100 measures of transparency, at least one company provides sufficient information. This means that those examined companies could mutually share their practices (Bommasani et al. 2023).

This study also suggested recommendations to developers, deployers, and policymakers, some of them are (Bommasani et al. 2023):

(1) Lawmakers should make sure that the AI policy interventions are based on solid empirical evidence – therefore, as it enables governments to gather the necessary evidence for which transparency is crucial;
(2) Transparency needs to be increased across the supply chain;
(3) Companies that deploy foundational models from other companies can't have the appropriate transparency if developers don't share the information about their own products with them, which represents a great risk that needs to be assessed.

A European Parliament study on *A governance framework for algorithmic accountability and transparency* (further reffered as EP STOA study) defines that the primary role of transparency is

to enable accountability (European Parliament 2019). This definition refers to the transparency of the information on data, algorithms, outcomes, usage of automated decision-making systems, etc. (European Parliament 2019). For an AI application to be considered trustworthy it must be transparent, so individuals could be able to understand the decisions made by the AI system. A lack of transparency could raise questions about accountability, bias, discrimination and fairness. To show the importance of transparency the EP study has defined 7 areas of machine learning systems in which transparency should be demanded in order to respect fundamental rights (European Parliament 2019).

(1) Data – it refers to raw data, sources of data, data processing, verification of unbiasedness and representation
(2) Algorithms – it refers to testing, reducing variables to the most significant ones therefore choosing which one to validate, testing the system to check for prejudicial data, inspection of bug reports, algorithms analysis etc.
(3) Goals – it refers to transparency in the system's goals and priorities, for example transparency about goals and priorities should be required from the manufacturers of AI products.
(4) Outcomes – it refers to the transparency about the outcomes of the deployment of certain algorithmic systems, internal states of the systems, the effect on external systems etc.
(5) Compliance – it refers to the manufacturer's compliance with transparency requirements that have been imposed upon them, even requiring from them proofs that can be inspectable by regulators or the general public
(6) Influence – refers to transparency about whether any element of the AI process was bent to favour a particular outcome on purpose; for example, a trusted search platform has to flag to its users if it's boosting some specific search results because they were paid to do that.
(7) Usage – refers to the transparency about which personal data is used by the system to personalize outcomes, this might lead to users wanting to control the usage of data, whether for further personalization of the outcomes or on the other side, stopping the use of their data if it violates their privacy (European Parliament 2019).

The importance of ensuring transparency in AI systems is further proved by the number of international documents that regulate AI, including the question of transparency. For example, the General Data Protection Regulation (GDPR) from 2016, mentions transparency as a core principle of data protection and privacy and highlights that any processing of personal data should be lawful, fair, and transparent (European Commission 2016). Further, due to the global response to the GDPR in which many states are adopting data privacy legislation inspired by the GDPR, the EU is proving itself as the international coordination for algorithmic accountability and transparency (European Parliament 2019).

Moreover, the ethical guidelines published by the EU Commission's High-Level Expert Group on AI (AI HLEG) in April 2019 highlight transparency and oversight as two of its seven key requirements for the realization of trustworthy AI, among technical robustness and safety, privacy and data governance, diversity, non-discrimination and fairness, societal and environmental wellbeing, and accountability (High-Level Expert Group on Artificial Intelligence 2019). The importance of transparency as a key requirement was again confirmed in the EU White Paper on Artificial Intelligence (European Commission 2020).

In the international arena, in May 2019, the OECD Member States adopted common ethical principles on AI - OECD Council Recommendation on Artificial Intelligence. These ethical principles include a provision on boosting the transparency on the creation of AI systems and greater checks on how these models access people's data, which international companies and OECD governments agreed on, together with 6 partner countries (OECD 2019a).

Transparency in AI has a significant role in the general aim of developing a more trustworthy AI. The reason for that is the fact that transparency is closely connected to other important aspects of the protection of human rights in AI systems, such as data protection, non-discrimination, consumer protection, product safety, privacy and rules of liability (Larsson and Heintz 2020). In order to address algorithmic transparency and accountability the EP STOA study proposes a set of four policy options in order to address algorithmic transparency and accountability (European Parliament 2019):

(1) raising awareness through education, watchdogs and whistleblowers;
(2) ensuring accountability in public-sector use of algorithmic decision-making;
(3) regulatory oversight and legal liability in the private sector; and

(4) the global dimension of algorithmic governance (European Parliament 2019).

1.2. Accountability and responsibility

Transparency is often considered as a crucial first step toward accountability (Bommasani et al. 2023). Accountability is considered as one of the cornerstones of the governance of artificial intelligence (Novelli et al., 2023). Therefore, one of the most commonly raised concerns regarding AI is about keeping someone accountable while using what AI systems offer us. Accountability, same as transparency and oversight, are crucial for good public and private governance, for that reason, someone must be held responsible in case these AI systems make errors. The issue with the definition of accountability is that it doesn't have a unique descriprion and is often defined quite broadly. Accountability can be defined as "a relation of answerability requiring authority recognition, interrogation and limitation" (Novelli et al., 2023). In some of the major European documents on AI, accountability is often defined differently and. For example, in the High-Level Expert Group (HLEG) reports, accountability is defined "both as a principle that ensures compliance with the key requirements for a trustworthy AI and as a set of practices and measures" (Novelli et al., 2023).

Further, GDPR defines accountability as "a principle which requires that organisations put in place appropriate technical and organisational measures and be able to demonstrate what they did and its effectiveness when requested" (European Commission 2016). The previously mentioned EP study on *A governance framework for algorithmic accountability and transparency* defines accountability as "primarily a legal and ethical obligation on an individual or organisation to account for its activities, accept responsibility for them, and to disclose the results in a transparent manner" (EP 2019). EU AI Act defines it as "providers and implementers of AIs are accountable for different reasons and in different ways depending on the risk level of the respective AIs" (European Commission 2021). Further, the Act in this way introduces the concept of 'provider accountability', which means that "the individuals or organizations are held responsible for their actions when developing, employing or operating AI systems" (European Commission 2021). Accountability has many definitions but at its core, is an obligation to to disclose and explain one's actions to an authority.

To grasp the essence of accountability, it's important to identify the goals that accountability is supposed to serve. Novelli et al. (2023) identified 4 main goals which are widely recognized to

influence how policymakers conceive accountability accountability regimes within governance frameworks (Novelli et al. 2023) :

(1) Compliance – represents a goal to bind the agent to act according to agreed ethical and legal standards. In this way, accountability is understood similar to the concept of having a sense of responsibility, which would refer to acting in a transparent and fair way.
(2) Report – refers to proper reporting of the agent's conduct, in order for the relevant information to be preserved to be explainable and justifiable. This enables challenging and disapproving of the agent's conduct.
(3) Oversight – refers to examining information, evidence and conduct.
(4) Enforcement – refers to determination of what consequences the agent has to bear in case of misconduct, such as prohibitions, sanctions and authorizations.

1.3 Privacy and data protection

AI is already invading our personal lives, through applications and social networks such as WhatsApp and Facebook; through the use of AI technology to select and profile individuals and groups by police and security agencies or to prevent threats such as terrorism (Van den Hoven van Genderen, 2017). The main risks posed to privacy and data protection by AI are risks regarding integration with surveillance technologies, informed consent, bias and discrimination, infringement of data protection rights of individuals such as right to prevent processing of personal data, right of access to personal data, loss of control over personal data, right not to be a subject to an automated processing decision, and many more (Rodrigues, 2020). Consequently, there is a risk that mass government surveillance could be a greater challenge to democracy than the security it aims to safeguard through these measures (Van den Hoven van Genderen, 2017).

Legal systems need to go hand in hand with the development of AI technology to make the processing of personal data by AI transparent before AI becomes too autonomous. If the autonomy of AI systems becomes greater, the challenge of ensuring the transparency of personal data significantly increases (Van den Hoven van Genderen, 2017). The right to privacy is a fundamental right that is protected both by international and national law, ensuring protection against government interference (Van den Hoven van Genderen, 2017). According to the General

Data Protection Regulation, individuals are given the right to empower them to contest and demand a reassessment of automated decision-making that significantly impacts their rights or legitimate interests (European Commission 2016). Data subjects have the right to raise objections, on grounds relating to their specific situation, at any time to the processing of personal data that relies on tasks performed in the public interest or legitimate interests (Rodrigues, 2020). GDPR also provisions that data controllers have to implement suitable measures in order to safeguard rights and freedoms (Rodrigues, 2020).

The fundamental aspects of privacy and data protection are based on the non-interference principle outlined in Article 8 of the European Convention on Human Rights (ECHR) which specifies the protection and control over personal data (European Convention on Human Rights, art. 8):

(1) Everyone has the right to respect for his private and family life and his home
(2) There shall be no interference by a public authority with the exercise of this right except such as is by the law and is necessary in order to ensure security, public safety, for the prevention of crimes, protection of health, freedom, morals and rights

In order to ensure the protection of privacy and personal data, the EU Committee on Civil Liberties, Justice and Home Affairs is giving guidance for future regulations which also highlights the importance of cooperation between private and public sectors and academia and uniform and horizontal approach in the Union (European Parliament Committee on Civil Liberties, Justice and Home Affairs 2016):

(1) by highlighting the responsibility of AI developers by ensuring "that the right to the protection of private life and the right to the protection of personal data as enshrined in Article 7 and 8 ECHR and Article 16 TFEU apply to all areas of robotics and artificial intelligence and that the Union legal framework for data protection must be fully complied with; underlines the responsibility of designers of robotics and artificial intelligence to develop products in a way that they are safe, secure and fit for purpose and follow procedures for data processing that's compliant with existing legislation, confidentiality, anonymity, fair treatment and due process"

(2) Committee calls on the Commission to ensure that any Union legislation on robotics and AI will take into account the rapid technological development to ensure that Union legislation does not lag behind the curve of technological development and deployement when making the rules on privacy and data protection
(3) Committee advocates for a uniform, horizontal approach to robotics and artificial intelligence in the Union regulatory framework
(4) Committee calls on the Commission and the Member States to promote strong and transparent cooperation between the private and public sectors and academia which would reinforce knowledge sharing and promote education and training in the field of fundamental human rights.

2. Risks of misuse

Besides the important challenges that were previously discussed, advancements in artificial intelligence make space for significant risks of misuse. This particularly shows through technologies such as Autonomous Weapons Systems (AWS) and deepfakes, which will be assessed in further analysis. Through these examples, the urgent need for global approach to AI regulatory framework and deeper international cooperation is highlighted, in order to prevent and control the misuse of AI and ensure security, democratic integrity and protection of human rights worldwide.

2.1 Risks of misuse -Autonomous weapons systems (AWS)

Among other challenges, the law and policies also have to catch up with technological and scientific advances in the area of weaponry. As in other fields that intertwine with artificial intelligence in modern times, the same problem exists in weaponry: ”The incalculable resources poured into technological advances for the development of warfare have far outweighed the resources invested into the constantly aging principles and rules that are supposed to govern them” (Acquaviva, 2020). There is still no universally agreed-upon definition, but autonomous weapons systems are defined broadly as “systems that are capable of selecting and attacking targets without human intervention or control” (Acquaviva, 2022). The International Committee of the Red Cross defined AWS as “weapons that can independently select and attack targets, i.e. with autonomy in the ‘critical functions’ of acquiring, tracking, selecting and attacking

targets"(ICRC, 2014). Human Rights Watch defines AWS as weapons that "would identify and fire on targets without meaningful human control" (Human Rights Watch, 2014). Those weapons are considered autonomous because AI becomes involved due to the absence of human intervention or control, therefore it substitutes the human input typically present in a traditional weapons system (Acquaviva, 2022).

The report published in 2012. by Human Rights Watch and the Harvard Law School International Human Rights Clinic - *Losing Humanity: The Case Against Killer Robots* is considered as one of the pioneering reports to extensively address the ethical and legal implications of AWS. As the primary concern it highlights the impact fully autonomous weapons would have on the protection of civilians during wartime (Human Rights Watch, 2012). This report analyzes whether the emerging technology would comply with international humanitarian law and maintain other safeguards against civilian casualties, and it also finds that fully autonomous weapons would not only fail to meet legal standards but would also weaken crucial non-legal safeguards for civilians. Finally, the research concludes that fully autonomous weapons should be banned and it urges governments to take immediate action toward this goal. Regarding the definition, in this report, the terms "robot" and "robotic weapons" include all three categories of unmanned weapons, which can be everything from remote-controlled drones to fully autonomous weapons. Further significance of the Losing Humanity report is the division of robotic weapons (unmanned weapons) into three categories, based on the autonomy they have:

Human-in-the-Loop Weapons	Weapons can operate only with a human command
Human-on-the-Loop Weapons	Weapons that can select targets and deliver force under human oversight with override ability
Human-out-of-the-Loop Weapons	Weapons that can decide and act on their own

The international community has since shown great interest in regulating autonomous weapon systems, including lethal autonomous weapons systems (LAWS). In public debates and policies the attribute "lethal" is sometimes added to the term AWS, highlighting the possible severity of the consequences this technology imposes. It remains unclear which exact technologies are considered LAWS since most of the definitions include many different aspects such as humanoid robot soldiers, landmines, combat drones, close-in weapon systems or purely virtual cyber

weapons (Bächle and Bareis 2022). In 2013, the framework of action was developed on the international level, when a Meeting of High Contracting Parties to the Convention on Certain Conventional Weapons (CCW) resolved that the Chairperson would assemble an informal meeting of experts, known as the Group of Governmental Experts (GGE) to discuss issues related to emerging technologies in the area of lethal AWS (LAWS). In this way, the CCW serves as a platform for international cooperation, dedicating itself to the regulation of LAWS. In 2019, this Group of Governmental Experts adopted 11 guiding principles as follows (Acquaviva, 2022):

(1) International humanitarian law applies fully to all weapons systems, including LAWS;
(2) Accountability cannot be transferred to machines, therefore humans will always be the ones having the responsibility for the decisions;
(3) Human-machine interaction has to comply with international humanitarian law;
(4) Accountability for the development, deployment, and use of any emerging weapons system must be ensured following applicable international law. This includes operating such systems under a responsible chain of human command and control;
(5) States must assess if the use of a new weapon violates international law before studying, developing, acquiring, or adopting it;
(6) The principles of physical security, appropriate non-physical safeguards (including cyber-security against hacking or data spoofing), the risk of acquisition by terrorist groups and the risk of proliferation should be considered when developing or acquiring new weapons systems;
(7) All weapons systems need to include risk assessments and mitigation measures.
(8) How emerging technologies in LAWS can be used while still protecting human rights;
(9) This means that when developing potential policy measures, emerging technologies in LAWS systems should not be anthropomorphized;
(10) Progress or access to peaceful uses of intelligent autonomous technologies should not be hampered;
(11) The CCW offers an appropriate framework for dealing with the issue of emerging technologies in the area of lethal autonomous weapons systems, which finds a balance between military necessity and humanitarian considerations.

In 2015 an open letter signed by AI and Robotics researchers called for "a ban on offensive autonomous weapons beyond meaningful human control" (AI & Robotics Researchers 2015). The Autonomous Weapons Open Letter: AI & Robotics Researchers is one of the most significant developments in the international discourse on AWS regulation representing a collective unified protest. To this day (6th of June 2024) this open letter was signed by 4985 AI/Robotics researchers and 27800 others including one of the most prominent AI and Robotics researchers, such as Stephen Hawking, Elon Musk, Steve Wozniak, Noam Chomsky and Stephen Goose (AI & Robotics Researchers 2015).

The international community has not yet adopted a unified approach to whether AWS is properly governed by the existing international law (Chengeta, 2022). Most of the international actors, including the International Committee of the Red Cross, take the view that existing law is inadequate and that it needs further regulation, but there is still a small number of states, including the United States, Russia, Israel, Australia and the United Kingdom, who argue that international humanitarian law (IHL) still adequately addresses the challenges posed by AWS (Chengeta, 2022). On the other hand, many nations are aiming for military advantage and therefore want to be included in the global regulation of AI, especially the United States, Russia, Israel, South Korea, the UK, Australia, Germany and France (Bächle and Bareis 2022). These countries also host companies that are leading in technological innovation, including robotic military innovation, most probably because their governments are involved in geopolitical tensions and conflicts in which they tend to use AWS (Bächle and Bareis 2022). Therefore, it's in great interest for them to be included in the process of regulating AI and shaping it based on their needs.

For example, the worsening of the international security situation has prompted the USA to reduce its standards of human control over AWS, increasing the likelihood of employment of AWS (Bächle and Bareis 2022). The Department of Defense of the USA has published the DoD Directive 3000.09 on Autonomy in Weapons Systems which went into effect in January 2023, appears to be moving toward further distancing the human element from the deployment of AWS, contrary to the definitions mentioned above by the ICRC and Human Rights Watch, which advocate for meaningful human control (Barbosa, 2023). More precisely, the new Directive definition of AWS replaces the term "human" with "operator", who may not necessarily be a

human (Barbosa, 2023). These changes in definitions demonstrate the dominance of US military interests over the protection of human rights. Therefore, such actions are undermining international humanitarian efforts in the task of establishing global governance of AWS with rules that would be binding, supranational and human-rights centered (Bächle and Bareis 2022).

In December 2023. the rising concern over the use of autonomous weapons was highlighted by UN General Assembly Resolution 78/241, which has been voted on to call for a rigorous study of the AWS topic. 152 states voted in favor of the resolution, with only 4 states voting against it (Belarus, India, Mali, and Russia) and 12 states staying abstained (United Nations General Assembly 2023). This resolution stressed the urgent need for international actors to address the challenges and concerns raised by AWS. It's expected from the international community to do this through the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. Additionally, this resolution obliges the UN Secretary-General to aks for and take into account the positions of member states and observer states on LAWS in order to have insight into challenges from different humanitarian, legal, security, technological, and ethical perspectives (United Nations General Assembly 2023). It also asks the Secretary-General to take into account the views of international and regional organizations, the International Committee of the Red Cross, industry, civil society, and the scientific community (United Nations General Assembly 2023).

Even though fully automated LAWS are not yet used on the field, it's of great concern that this can happen in the near future based on the fast development of tech tools in the private sector, such as the development of drones and self-driving cars (Reeves, Alcala, & McCarthy 2020). The international community is struggling with regulating the products of emerging technologies which becomes even harder as soon as they become widely available and cheap, and therefore within the reach of most people (Reeves, Alcala, & McCarthy 2020). Therefore, the limited time in which states need to react to the fast development of AWS with new regulations is making this field even harder to regulate. Another reason is the high cost of AI research and development, which remains a possibility only for those states and large corporations that can afford it (Reeves, Alcala, & McCarthy 2020). At the moment, there are numerous obstacles that stand in the way of reaching international consensus on LAWS on a global level: global and individual state-specific political approaches, questions of taxonomy and differences in legal approach and

similar. In addition to that, even though AI research and development predominantly remains at the state level, opinions created in public by prominent figures and corporations may also shape the trajectory of the AWS regulation. All these challenges prompted states to call upon the international community to approach the AWS regulation on a global level in order to stay in line with its advancements (Reeves, Alcala, & McCarthy 2020).

2.2 Risk of misuse - Deepfakes

Another problem that emerged with the fast development of digital technology is that it has become significantly more challenging to distinguish between real and fake media and information. What contributes to this problem is the emergence of deepfakes which are videos that apply artificial intelligence and are hyper-realistic in depicting things that never happened. Europol in its report *Facing reality? Law enforcement and the challenge of deepfakes* define deepfakes as a "technology that uses Artificial Intelligence to audio and audio-visual content. Deepfake technology can produce content that convincingly shows people saying or doing things they never did, or create personas that never existed in the first place" (Europol, 2022). Deepfakes can further be defined as "the product of AI applications that merge, combine, replace, and superimpose images and video clips in order to create fake videos that appear authentic" (Maras & Alexandrou, 2018). Many scholars predict that deepfakes in the future will be used in various malicious and misinformative activities, including bullying, producing fake video evidence in courts, revenge porn, political sabotage, terrorist propaganda, blackmail, market manipulation, and creating and spreading fake news (Maras & Alexandrou, 2018). This represents a huge threat to public discourse and democracy since false information can be disseminated very quickly through social media, especially through videos, which have become increasingly popular as a source of information. Disinformation campaigns and deepfake content aim to misinform the public and influence their opinion (Europol, 2022). The increasing distribution of disinformation and deepfakes can change the public perception of authority and trust in information networks, which can be used to escalate existing conflicts or provoke new ones, erode trust in institutions and undermine political opponents (Europol, 2022).

Many studies rank deepfake technology as one of the biggest threats to human rights today, even bigger than identity theft, primarily to privacy, autonomy and democracy which present one of the principal values of liberal democracy (Europol 2022). As deepfakes flood the systems for

knowledge and screen human rights abuses, trust in authorities and officials erode, for this reason, experts fear that this could create a world in which citizens live in different fictional realities, a situation sometimes referred to as 'information apocalypse' or 'reality apathy (Europol, 2022). Since deepfakes first appeared on the Internet in 2017, academic literature has been sparse on this topic, the public seems relatively uninformed about the dangers of deepfakes and the international community is still trying to find a way to regulate it properly (Europol, 2022).

As for the trends in the regulation of deepfakes, in the EU case, it's noted that European law has been lagging behind the advancement of technology and the redefinition of crime, both at the national and regional levels (Europol, 2022). The creation of new policies should, of course, be mindful of current digital law enforcement needs in light of changing ethical norms since many others see the digital domain becoming more heavily regulated in the next ten years (Europol 2022). According to the European Parliament report, *Tackling Deepfakes in European Policy*, the COVID-19 pandemic shed more light on this topic but has also at the same time increased the use of video conferencing tools that often use adjustable backgrounds, which is one example of using manipulation of digital realities in our daily routine (European Parliament Research Service, 2021). The report also highlights that the regulatory landscape in the European Union that's related to deepfakes "comprises a complex web of constitutional norms, as well as hard and soft regulations on both the EU and the Member State level" (European Parliament Research Service, 2021). EU AI ACT takes a risk-based approach to the regulation of AI and its applications. Deepfakes are not completely banned but they have to adhere to certain minimum requirements (European Commission 2021). Under Article 52(3) of the Act, transparency is required from the creators, meaning that in case of creating a deepfake its artificial origin must be disclosed together with the information about the techniques used to make it (European Commission 2021). By making this information public, it's expected that consumers will be less prone to manipulation. However, it's still uncertain if these requirements are enough to tackle the challenge of deepfakes, transparency alone might not be sufficient to address the malicious potential of deepfakes, particularly if creators discover new ways to bypass the disclosure requirements (Sgueo, 2024).

Further, in February 2024 the EU AI Office was established, with one of its key roles in promoting and facilitating the development of codes of practice at the Union level to ensure the effective implementation of responsibilities related to the detection and labeling of artificially generated or manipulated content (Sgueo, 2024). Under this mandate, the Commission is authorized to enact implementing acts to endorse these codes of practice, thereby ensuring they meet specific standards and effectively tackle the challenges posed by artificially generated or manipulated content (Sgueo, 2024). In general, EU legislation adopts a proactive approach to addressing deepfakes and AI-generated text, but the Act fails to establish a clear framework for the legal liability of developers involved in deepfake technology and has issues regarding clarity and specificity remain in the definitions of "deepfake" and "artistic/creative work" (Sgueo, 2024). Furthermore, while deepfakes are right now categorized as "limited risk" AI systems under the AI Act, their increasing potential for harm shows they will potentially need more regulations and be classified as “high-risk” so the effective enforcement mechanisms could address them (Sgueo, 2024).

A recent example in the EU that shows the importance of international cooperation in tackling deepfakes and misinformation is the joint effort between the EU, EU political parties and technology companies. In March 2024 the European Commission published guidelines for Very Large Online Platforms and Search Engines before, during, and after electoral events (European Commission, 2024). With these guidelines, the European Commission urged social platforms of major technology companies, among whom Meta platforms, to take action against artificial intelligence deepfakes ahead of the European elections in June. In the past years, technology companies have also been putting effort into tackling the challenge of deepfakes and misinformation, for example, in 2020, Meta adopted a new policy banning deepfakes from their platforms and claiming that it would remove content that was AI-edited and could mislead people, but at the same time still permitting satire/parodies made by AI (Europol, 2022). Therefore, in April 2024 Meta announced that it will start applying "Made with AI" labels to AI-edited videos, images and audio posted on its platforms (Facebook, Instagram and Threads) which officially started in May (Reuters 2024). The example of Meta's policy on misinformation shows us their determination to have control over deepfakes, and other AI-edited media, where “manipulation isn’t apparent and could mislead, particularly in the case of video content” (Meta, 2024).

Moreover, in April 2024, the majority of the EU's political parties – among whom European People's Party (EPP), Party of European Socialists (PES), European Conservatives and Reformists Party (ECR) - signed a voluntary code of conduct in preparation for the European elections taking place from June 6-9 2024, includes a commitment to refrain from creating, using, or distributing "any form of deceptive content" (Politico, 2024). This code of conduct arises from growing worries about foreign interference through disinformation campaigns and cyberattacks. Videos, photos, and audio content edited with artificial intelligence tools to impersonate public figures, including politicians, have begun spreading in Slovakia, the United Kingdom, the United States, France and Poland (Politico, 2024). In July 2023, The UN Secretary-General Antonio Guterres addressed the Security Council over the the human development potential of artificial intelligence but also the need to guard against its malicious uses (Guterres 2023). Guterres highlighted that AI can present serious threats for global peace and security if used for generating deepfakes and spreading disinformation and hate speech, and therefore urging the need for a universal global approach (Guterres 2023).

The further impact of deepfakes hinges on the strategies that will be adopted by international actors in order to address this issue. This challenge has to be approached on a global level since deepfakes and misinformation pose a dual threat, to both individuals and states by manipulating domestic politics and public opinion. Other than states and international organisations, a big role in regulating deepfakes will be in the hands of technology companies by further advancing AI tools for preserving the authenticity and aiding in the detection of deepfakes. The joint effort between the EU and technology companies ahead of the EU elections represents a big step forward in the global regulation of this issue and shows that this challenge cannot be tackled on an individual level. Since deepfakes are not completely banned, as seen in EU AI ACT, they still have to adhere to certain minimum requirements (European Commission 2021). Therefore, policymakers and individuals at both domestic and international levels still have the ability to use them. It’s up to the international actors to further regulate this technology to make sure that the use of deepfakes doesn’t lead to misinformation, manipulation of public opinion or discreditation of political opponents by developing and enforcing ethical guidelines.

Conclusion

This chapter focused on detecting the major challenges and risks imposed on human rights by Artificial Intelligence. Transparency, oversight, accountability and responsibility, privacy and data protection have been identified as interdependent principles that are crucial in order to ensure the protection of human rights in AI systems. Additionally, examples such as Automated Weapons Systems and Deepfakes, the potential implications of AI misuse on human rights were shown. In all of these examples, the importance of global governance has been noticed while handling the challenges that could be imposed on human rights. Even though the choice of the principles was based on the principles outlined in international declarations and policy documents to ensure global relevance, it can be noticed that the analysis of the legal and policy framework that was referred to in this chapter mostly draws from European-level frameworks and declarations. The reason for this is that the European frameworks have gone a long way in setting a precedent and having a pioneering role in AI global governance by enshrining these principles in many of its AI regulations, which will be accessed more thoroughly in the succeeding chapter. This European leadership in regulating the challenges imposed by AI highlights the importance of the need for a global AI governance, in order for this transformative technology to uphold human rights universally.

CHAPTER II: GLOBAL ACTORS SHAPING AI POLICY

To demonstrate the connection between artificial intelligence and human rights, it's essential to show the influence that technology has on international policy framework development. This innovative correlation between two disciplines—human rights and technology— helps us to gain a deeper understanding of the complexity of protecting human rights on a global level. Although at first look it may seem like these disciplines have nothing in common, the evolution of technology has made them inseparable. The multidimensional nature of technology has made it present in every aspect of people's lives through communication channels, data collection and analysis, surveillance mechanisms, digital activism, social networks, governance structures and even electoral processes. Consequently, human rights couldn't be left out of the impact of this evolution. In the context of human rights, technology can present both opportunities and challenges. Complex topics such as education, health system, military, democracy and freedom of speech could be enhanced and improved with AI but could also present an opportunity for manipulation, surveillance, control and abuse by various international actors, including governments, private companies, or even non-state actors such as extremist groups or hackers.

In recent years, with the fast-paced development of technology, the international community has made its efforts to find a balance between technological advancements and the expansion of the human rights framework. Governments worldwide are increasingly recognizing the fact that developments and the effects of AI on social life, the economy, or national politics can be unpredictable (Franke & Sartori 2019). Based on the complexity of AI, a single regulation is most likely not suitable; instead, a system of AI global governance is needed - which would be built on the existing legal framework, composed of specific and general regulations (Mazzini, 2018).

This chapter will elaborate on the current regulations and practices of different international actors providing insights into addressing the research question. It also aims to show existing motivations and strategies, if any, for the internationalization of cooperation in AI regulation. The chapter starts with an examination of regulations and practices within the European Union. The fundamental documents for the regulation of AI in the EU are examined, particularly: the Ethics Guidelines for Trustworthy AI, the General Data Protection Regulation (GDPR), The White Paper on Artificial Intelligence and the EU Artificial Intelligence Act (EU AI Act). The

analysis of these documents focuses on the provisions regarding the regulation of AI from a human rights perspective, but also on showing how these regulations might have a positive impact on global governance. Where sources permit, different attitudes of the Member States are shown. Furthermore, in the same spirit, the focus is shifted to the regulations and actions undertaken in the international arena, mainly through the work of international organizations. The goal of this chapter is to show how the integration of human rights principles can shape the regulation of AI by various international actors. A particular emphasis is put on how these actors approach the global governance of AI and how the regulations on the international level improve the protection of human rights.

1. European Union and AI

In June 2018, the European Commission established an independent High-Level Expert Group on Artificial Intelligence (AIHLEG) which created Ethics Guidelines for Trustworthy AI in April 2019 and with it made a first big step towards AI regulation (Smuha, 2019, 1). According to the guidelines, trustworthy AI should fulfill three requirements, it should be lawful – adhering to all relevant laws and regulations, ethical – ensuring compliance with ethical principles and values, and robust – both from a technical and social perspective, as AI systems can cause unintentional harm despite good intentions (High-Level Expert Group on Artificial Intelligence 2019). The Guidelines also include four ethical principles which should be considered as ethical imperatives in the context of AI: respect for human autonomy, prevention of harm, fairness and explicability (Smuha, 2019, 1). The Guidelines also propose 7 key requirements that AI systems should meet to be considered trustworthy

(1) human agency and oversight,
(2) technical robustness and safety,
(3) privacy and data governance,
(4) transparency,
(5) diversity,
(6) non-discrimination
(7) fairness, societal and environmental well-being and accountability (High-Level Expert Group on Artificial Intelligence, 2019).

With these guidelines for the creation of AI systems, Europe made a significant advance toward establishing a unified ethical framework for AI, showing three key aspects in which international cooperation is evident (Smuha, 2019, 9):

- Inclusivity: the group ensured that the process of creation of the Framework for the trusty AI was inclusive by including experts from different countries, disciplines and stakeholder groups, by also ensuring the inclusion of society on a large scale by including a multi-stakeholder platform European AI Alliance.
- Agility: Acknowledging the importance of regulating the right issues in proper way and at the right time. Highlighting the importance of not unnecessarily blocking AI innovations while setting up new regulations
- Globality: the European Commission has put its focus on international cooperation and global governance solutions through communication.

Just as there is still no consensus on the definition of artificial intelligence at the global level, different definitions of AI systems can also be found at the EU level. For example, European Commission in its communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee, and the Committee of the Regions has defined AI as the "systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals" (European Commission, 2018). Another definition of AI systems by the High-Level Expert Group on Artificial Intelligence systems defines AI as: "software (and possibly also hardware) systems designed by humans 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" (High-Level Expert Group on Artificial Intelligence 2019). Finally, the latest EU AI ACT defines AI systems as: "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (European Commission, 2021).

Polities must adapt to technological advancements, to harness the potential of innovation, while also minimizing the associated risks.

The EU faces a challenge: becoming a global leader in AI and at the same time extensively regulating AI to guarantee "ethical, responsible and sustainable outcomes" (High-Level Expert Group on Artificial Intelligence 2019). Achieving this balance is challenging considering the fact that the EU has lagged behind the US and China in AI adoption and innovation. (Franke & Sartori 2019). Currently, the American and Chinese technological companies appear to be moving in different directions, and they possess a considerable head start in terms of empirically measurable capability (Franke & Sartori 2019).

When it comes to the question of a common European AI approach, it mainly depends on the agreement among its member states. One of the ways to achieve this is having a unified European approach and priorities other than the pursuit of national priorities. The start of this European approach can be seen with the signing of the Declaration of Cooperation on Artificial Intelligence in April 2018 when member states agreed to collaborate in addressing AI questions (Franke, 2020). This cooperation aimed to create a coordinating entity in tackling the AI social, economical, legal and ethical issues (Franke, 2020). Several member EU countries have already had their national AI strategies, among whom the European "big two" – France and Germany, in which they both highlight both bilateral and multilateral cooperation, alongside the simultaneous pursuit of both national and European goals. Even thou similar, the motivation behind these strategy provisions are different for each: Germany focuses on its cooperation with France and European cooperation in general, while France aims to also focus on its cooperation with Germany but selectively supports European cooperation in areas it finds beneficial, because of European's geopolitical power and capacity to confront other players, specifically the US and China (Franke, 2020).

Other than the Ethic Guidelines for Trustworthy AI, this thesis has identified other 3 important documents when it comes to EU AI regulation and its effect on human rights: General Data Protection Regulation (2016), the White Paper on Artificial Intelligence (2020) and the EU AI Act (2021).

1.1 General Data Protection Regulation (GDPR)

The European Data Protection Regulation is applicable as of May 25th, 2018 in all member states intending to harmonize data privacy laws across Europe (European Union 2016). Since GDPR is mostly focused on the protection of personal data, it is of great importance to analyze it since AI is most likely to pose risks to fundamental rights such as rights of privacy, personal data protection, and non-discrimination, when processing personal data (European Union 2016). Moreover, GDPR is specifically designed to apply to AI systems that process personal data whether partially or fully automated (European Union 2016). Further, it is specifically important to include GDPR in the analysis since it constitutes a great example of a complex but flexible piece of legislation that is therefore particularly well-suited to contribute to a system of AI governance (Hacker, 2018). The complexity of GDPR combines

(1) general rules, encompassing provisions that apply equally to the processing of personal data by both humans and by automated means;
(2) specific rules including the provisions related to processing by automated means; and
(3) co-regulatory rules, which require data controllers to independently analyze and mitigate the risks associated with the processing methods they use, thus giving them the discretion to self-regulate within the general protection standards established by the GDPR (Wrigley, 2018).

The GDPR protects the personal data of all EU residents, regardless of the location of the processing and by personal data it considers "information that, directly or indirectly, can identify an individual, and specifically includes online identifiers such as IP addresses, cookies and digital fingerprinting, and location data that could identify individuals" (Goddard, 2017). Its significance lies in the wide territorial scope and expanded definition of personal data which enhances the protection of personal data (Goddard, 2017). The GDPR has six general data protection principles

(1) fairness and lawfulness;
(2) purpose limitation;
(3) data minimisation;
(4) accuracy;
(5) storage limitation; and

(6) integrity and confidentiality (General Data Protection Regulation 2016).

The core of GDPR is data protection by design and default which is ensured by transparency, through providing full information to individuals in an accessible style and manner and accountability, by ensuring that all organisations are taking responsibility when using personal data (Goddard, 2017). The most important provision of this document is a change in the standard required for consent which by GDPR needs to be freely and explicitly given, verifiable, specific, informed and evidenced by clear affirmative action, which shifts the balance of power between organisations and individuals, elevating individuals' right to access and control the use of their personal data. (Goddard, 2017).

GDPR is overall considered to have the potential to address actual or potential undesirable uses and applications of AI systems because its provisions address all challenges that AI poses to privacy, personal data protection, and the prohibition of discrimination (Mazzini, 2018). Still, several gaps could be defined and show the need for further regulation of AI (Ufert, 2020):

(1) The first gap is that often the required consent for the process of personal data is being disrespected by "a simple click on the "yes" box under several pages of Terms and Conditions and/or the reduced explainability of certain AI systems" (Ufert, 2020)
(2) the concept of personal data lacks an exhaustive definition, resulting in the scope of the right to information under the GDPR being disputed
(3) Although there are several different definitions within EU institutions and bodies, the EU's definition of AI by High-Level Expert Group on Artificial Intelligence "AI systems are software (and possibly also hardware) systems designed by humans 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" (High-Level Expert Group on Artificial Intelligence 2019) limits AI systems to be "designed by humans" and the GDPR reflects this aspect by requiring meaningful human oversight for the use of automated means.

Nevertheless, the development of AI is a process that will continue for a long time and it will eventually work on filling the initial gaps, like the issues surrounding the concept of specific

consent and similar, using more specific provisions tailored to AI and its distinct characteristics that differ from human actions. Only by further development and use, AI can fully adhere to fundamental rights.

1.2 The EU White Paper on Artificial Intelligence

The EU White Paper on Artificial Intelligence was published in 2020 by the European Commission as a part of the EU's broader digital strategy. Preceding the white paper, the EU faced significant pressure due to the uncertainties around Brexit (Berger 2018). This was also a time when other competing actors, in rapid succession to one another, formulated ambitious AI goals with comprehensive investment plans, raising concerns about the EU's potential lagging behind. During this time China loudly proclaimed that it would be the world leader in AI by 2030 (Franke 2020). It was quickly realized that achieving competitiveness in AI cannot rely solely on individual member states which is, for example, echoed by for example France's national AI strategy expressing concerns about Europe potentially becoming a "cyber colony" (Franke 2020). Additionally, EU policymakers are alarmed by the emergence of a multipolar world order, in particular the rise of China which was even further exacerbated by the UK leaving the EU together with its resources in AI (Lee 2018).

According to the European Commission (2020), the White Paper on Artificial Intelligence outlines as its main goal the importance of a solid European approach to the regulation and development of AI under European values such as human dignity, privacy protection, and preventing fragmentation by creating national initiatives (European Commission 2020). According to the White Paper, the international cooperation approach on AI matters must be based on the respect of fundamental rights, including human dignity, pluralism, inclusion, non-discrimination and protection of privacy and personal data. This document was based on Commission's President Ursula von der Leyen's announcement in her Political Guidelines from 2020 about a coordinated European approach to the human and ethical implications of AI as well as a reflection on the better use of big data for innovation (Von der Leyen, 2020).

In this paper, AI is defined as "a collection of technologies that combine data, algorithms and computing power" (European Commission, 2020). The White Paper highlights the areas of benefits of AI for citizens and society such as improved health care, safer and cleaner transport systems, better public services, reduction of the costs of providing services, improving the

sustainability of products and ensuring the security of citizens through supporting law enforcement (such as crime and acts of terrorism), as a mean of reaching Sustainable Development Goals and as a support of democratic processes. As a key element of the future regulatory framework for AI in Europe, it introduces the 'ecosystem of trust' which is defined as "a policy objective in itself which should give citizens the confidence to take up AI applications and give companies and public organisations the legal certainty to innovate using AI" (European Commission, 2020). And for that to be achieved, the development of AI must ensure compliance with EU rules, including the rules protecting fundamental rights and consumers' rights.

The White Paper highlighted the fact that Member States already started to get more involved on a national level in regulating AI due to the lack of a unified European approach which presents a risk for fragmentation in the internal market. Before the presentation of the White Paper, the European Commission was tasked by the German Data Ethics Commission to include a five-level risk-based regulation system ranging from no regulation for the least harmful AI systems to a complete prohibition for the most dangerous AI systems (European Commission, 2020).

The White Paper shows examples of a tendency for a unified approach to tackling this topic by emphasizing the cooperation between EU bodies and other international actors, both other like-minded countries and global players. For example, the EU High-Level Expert Group on Artificial Intelligence (AI HLEG) involved non-EU organisations and several governmental observers in the creation of its ethical guidelines. Further, the EU was involved in the creation of the OECD's ethical principles of AI, and the follow-up of the report of the United Nations High-Level Panel on Digital Cooperation. The paper highlights the 7 requirements already identified in the Guidelines of the High-Level Expert Group which are human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing, and accountability (European Commission, 2020). The White Paper also highlights that developers and deployers of AI are already subject to European legislation on fundamental rights (e.g. data protection, privacy, non-discrimination), consumer protection, and product safety and liability rules) through directives such as:

- Race Equality Directive (Directive 2000/43/EC) – which ensures that AI systems do not perpetuate racial discrimination and by that promoting fairness and equality.

- Directive on Equal Treatment in Employment and Occupation (Directive 2000/78/EC) - crucial when AI applications are used in recruitment or employment, by preventing discrimination based on age, religion, belief, disability or sexual orientation.
- Directives on Equal Treatment between Men and Women about employment and access to goods and services(Directive 2004/113/EC; Directive 2006/54/EC) – crucial for preventing the perpetuation of gender biases or stereotypes by AI system.
- Data Protection Law Enforcement Directive (Directive (EU) 2016/680 of the European Parliament and of the Council) – vital for the protection of personal data during the collection, storage and usage of data by AI systems in law enforcement.

On the other hand, the White Paper mentions possible challenges and risks to human rights which could stem from deficiencies in the overall design of AI systems or even from utilizing data without correcting potential biases. AI makes it easier to track people's actions daily, which could be a capability exploited by governments through mass surveillance, limiting freedom of expression and privacy. With its ability to realize links between different data, AI could also de-anonymize data about persons and therefore create data protection breaches. Another problem is that it's hard to make sure that AI will always function in compliance with the rules of existing EU law because of its complex and often unpredictable operation. Also, it's hard to access justice and prove how some decision made directly by AI or with the involvement of AI was taken and thus whether relevant regulations were adhered to.

The twopartite solution that the Commission proposes through the White Paper is two building blocks:

- ecosystem of excellence: which represents the mobilisation of money, support of research and innovation, and creation of incentives for adopting AI systems with the goal of the promotion and development of AI.
- ecosystem of trust: related to compliance with EU rules, providing citizens the confidence to use AI applications, promoting ethical and human-centric approach to technologies with the goal of addressing the risks AI can impose.

Most importantly, in order to address the challenges imposed on fundamental rights, the White Paper suggests implementing a risk-oriented strategy to regulate AI and to adapt the current legislation in order to go in line with new advancements in technology. By this, it focuses on

high-risk AI applications for which the regulation would be needed. The Paper mentions that two cumulative criteria, where is it used and how is it used, define high risk: (1) Public sector, such as healthcare, energy, transport and other are defined as critical sectors (2) Critical use (how it is used), AI systems are employed in a manner where substantial risks are likely to occur, such as the risk of injury, fatality or significant material or immaterial damage.

With this said, the Paper also provides new legal requirements in accordance with the high-risk AI. Among those requirements, there are some especially relevant for the protection of human rights, which are in the areas of:

- Training data: data sets need to be sufficiently broad and cover all possible scenarios needed to avoid dangerous situations, data needs to be non-discriminatory and representative, it must be ensured that data privacy and personal data are protected
- Keeping of records and data: the records, documentation and, where applicable, data sets should be preserved for a defined and reasonable period to ensure the effective enforcement of the legislation.
- Robustness and accuracy: requirements that ensure the AI systems are resillient, robust and accurate, with reproducible outcomes
- Human oversight: before any output of the AI system becomes effective it has to be previously reviewed and validated by a human, or otherwise, human intervention needs to exist as an ensured option afterward, there needs to be monitoring of the AI systems and possibility for the system to be stopped by a human at any time, by imposing operational constraints on the AI system
- Specific requirements for remote biometric identification: AI can only be used for remote biometric identification in situations where that use is duly justified, proportionate and subject to adequate safeguards. However, there are no details about what such additional safeguards might be as The Paper mentions this question will be regulated in further debates.
- Compliance and enforcement: The relevant legal requirements must be practically adhered to and effectively enforced both by competent national and European authorities and by affected parties. Competent authorities should be in a position to investigate individual cases and also evaluate their societal impact.

- Governance: A European AI governance framework, comprising collaboration among national competent authorities is necessary to avoid fragmentation of responsibilities, increase capacity in Member States, and make sure that Europe equips itself progressively with the capacity for testing and certifying AI-enabled products and services.

1.3 EU Artificial Intelligence Act

In April 2021, the European Commission proposed a Regulation on Artificial Intelligence known as the AI Act which was adopted by The European Parliament in March 2024, thus becoming the world's first comprehensive AI law (European Commission, 2021). The significant difference between the EU AI Act and all previous regulations of AI is that the Act provides a comprehensive risk-based approach specifically for AI technologies, which mixes the reduction of trade barriers with fundamental rights concerns. Its significance also lies in the fact that this regulation will have the power to limit the Member States' actions. This regulation is based on Article 114 of the Treaty on the Functioning of the European Union (TFEU) which signifies that It aims to achieve four specific objectives: safe AI systems that respect fundamental rights and EU values, enhancement of governance and effective enforcement of the existing AI legislation, Enabling the establishment of a single market for lawful, safe, and trustworthy AI and preventing market fragmentation (Yordanova, 2022). The wide scope that the AI Act has together with the broad extraterritorial effect can be compared with the approach of the General Data Protection Regulation (GDPR), in which the EU shows its tendency to regulate global markets (Yordanova, 2022).

Comparing the processes of the creation and adoption of this resolution, the stakes were much higher this time for the Member States. The reason for that is that the AI Act rules would put significant limits on the actions some states might want to take, especially when it comes to innovation since the balance between regulation and innovation in AI can therefore significantly influence the economic growth of Member States. This was one of the reasons why only a couple of weeks before the planned adoption of the AI Act countries such as Germany, France and Austria made hints that they might oppose the proposed text in the final voting (Politico, 2024a). Each of them had their own reasons for this attitude. Firstly, Austria had issues with the proposed

data protection provisions, on the other hand, French and German concerns were about the impact that some of the Act's provisions could have on innovation, in particular about the possible constraints for their companies Mistral and Aleph Alpha, that are considered as "Europe's budding AI champions" (Politico, 2024a). The stakes were quite high at the pre-vote period since there was a risk of permanently halting the law if enough number of countries opposed it. The Belgian Council presidency was put in an uncomfortable position since the French Economy Minister, Bruno Le Maire, even called for further rounds of negotiations. On the other side, Germany's coalition government told POLITICO that it's important that the European Commission clarifies that this AI Act doesn't refer to the use of AI in medical devices (Politico, 2024a). Further, it was important that the Act is adopted before the European elections scheduled for June 2024, so the members of the European Parliament had to make sure that the adopted regulation justified the expectations of the public (Politico, 2024a). Another disagreement concerned biometric surveillance (including facial recognition) in which EU lawmakers want to ban the use of AI, described in Article 5 which provoked heated discussions.

Differently from the EP, the individual governments wanted to keep that possibility as an exception for law enforcement - national security, defense and military purposes (Reuters, 2024). On the other hand, many civil society organizations demanded for a broader ban than what was envisioned in Article 5, which would mean a full prohibition of remote biometric identification (that are considered high-risk AI systems by the text of the Act), which was backed up by serious lobbying efforts (Yordanova, 2022). Another concern was the effect the AI Act would have on the EU's position in global competitiveness, which is one of the main goals of many of the EU AI strategies (Mügge, 2024). The concern regards high regulation standards which might make it more appealing for talent and investors to leave for Silicon Valley rather than the highly regulated European Union (Politico, 2023).

When it comes to the Act itself, it adopts a risk-based strategy to regulate AI systems, categorizing them into different risk tiers depending on the degree of risk for public interest and EU fundamental rights (European Commission 2021). By this, the Commission sets apart different risk levels concerning AI practices and separates them into four categories :

(1) unacceptable risks,
(2) high risks,

(3) limited risks
(4) minimal risks

An important note with this classification is that it's not static, which means that any AI system could change its type if necessary (Yordanova, 2022).

When it comes to unacceptable risks, they are described in Article 5 of the AI Act, and they refer to prohibited AI systems that could cause serious harm (European Commission 2021) :

(1) Using subliminal, manipulative, or deceptive techniques to alter behavior and undermine informed decision-making.
(2) Taking advantage of vulnerabilities related to age, disability, or socio-economic circumstances to distort behaviour.
(3) Biometric categorisation systems that infer sensitive attributes such as race, political opinions, trade union membership, religious or philosophical beliefs, sex life, or sexual orientation, except when used for labeling or filtering lawfully acquired biometric datasets or for categorizing biometric data in law enforcement.
(4) Social scoring which involves categorization based on social behaviour or personal traits, leading to harmful or unfair treatment of those individuals or groups.
(5) Evaluating the risk of an individual committing criminal offenses solely based on profiling or personality traits, unless it is used to supplement human assessments with objective, verifiable facts directly linked to criminal activity.
(6) Creating facial recognition databases by indiscriminately scraping of facial images from the internet or CCTV footage.
(7) Detecting or interpreting emotions in workplaces or educational institutions, unless its necessary for medical or safety reasons.
(8) 'real-time' remote biometric identification (RBI) in publicly accessible areas for law enforcement purposes, except under the following conditions: searching for a missing person, abducted, human trafficked or sexually exploited victims; preventing threats to life, or terrorist attack; or identifying suspects involved in serious criminal activities.

Based on the Act, high-risk AI systems include AI used in biometrics, critical infrastructure, education, employment, essential services, law enforcement, migration, and justice. But, there are some exceptions that apply to the use of AI for biometric identity verification, detecting

financial fraud, or organising political campaigns (European Commission 2021). Probably the ones most discussed are AI systems intended to be used for the "real-time" and "post" biometric identification of natural persons, as it was already mentioned that it caused fierce debates since many actors advocated for the complete ban on the use of AI in this case (Yordanova, 2022). Additional types of high-risk AI systems that are significant for human rights are those that refer to (European Commission 2021):

(1) AI systems which are "AI systems intended to be used to control or as safety components of digital infrastructure"
(2) "AI systems used to control emissions and pollution"
(3) AI systems for recruitment purposes or for making decisions regarding promotions or terminations
(4) AI systems used for access to private and public services and benefits that could be used by public authorities in order to check someone's eligibility for benefits or to decide on accession on educational and vocational training institutions
(5) AI systems used by law enforcement for detecting the emotional state of someone for lie-detection purposes
(6) AI systems used for processes of migration, asylum and border control management
(7) AI systems used by a judicial authority, for interpreting facts or law

Next, limited-risk AI systems are regulated in Article 52 of the AI Act and they include chatbots, emotion recognition systems and systems that generate deepfakes. The aim in order to protect human rights, in this case, is to make it clear to individuals that they are interacting with a machine in the case of chatbots or that content has been edited or generated by an AI system. The important feature of these systems is that they have the potential to become high-risk if they are misused (Yordanova, 2022).

The fourth type of AI systems is the minimal risk and general purpose AI systems, for which the Commission suggests a voluntary approach through self-regulation.

The EU AI Act can serve as an example of the potential for a regional regulation with a global scope in terms of its ethical norms in the field of AI. The complexity of the Act and the extensive work standing behind it will have an influence much further from the borders of the EU, establishing a framework for international cooperation in responsible AI development. This Act

serves as proof of the EU's commitment to fostering AI development while simultaneously preserving European ethical and societal values. The new risk-based classification system could have a tremendous impact on transparency and accountability, which gives the European approach to AI both a step forward in the protection of human rights but also a question of whether the investors and technological companies will rather invest in AI in other countries where they would have less regulation, such as USA or China.

The reputation of the EU as a "regulator rather than the inventor" often disregards the positive influence regulations have on human rights, particularly in fields like AI, which could have enormous negative consequences for human rights worldwide. Therefore, the Act's potential to inspire global ethical norms in AI is rooted in its principles' alignment with universal ethical values such as transparency, accountability and the protection of fundamental rights, making these regulations adaptable and applicable in different jurisdictions. The Act's journey shows the importance of regional initiatives, even though it's currently only implemented on the regional level, its influence will certainly bypass geopolitical boundaries and create a global impact.

1.4 A critique - Eurocentric approach towards the AI regulation

Based on the analysis of different European Commission strategies for the regulation of artificial intelligence crafted until 2021 done by Mügge, the Commission consistently highlights that AI is marked by fierce global competition, repeated references to the term "AI made in Europe" and mentions of the establishment of its own globally competitive AI sector (Mügge, 2024). Mügge further highlights that in one of its communications in 2018, the European Commission highlighted its worry that EU AI uptake is too slow and that a risk of brain-draining and consumption of solutions developed elsewhere is possible if major efforts don't take place.

Further, he continues proving the European competitive rhetoric in AI by highlighting the Annexes to the 2021 update of the Commission's AI strategy *Fostering a Common Approach*, and its 17 chapters, out of which only 3 chapters talk about AI for people, AI as a force for good and global promotion of the EU's vision of sustainable and trustworthy AI and the remaining 14 chapters highlight the EU leadership in AI. Moreover, he highlights that even though many EU AI strategies include statements about leaving no one behind and benefiting people and society as a whole, no concrete measures or proposals have been made (Mügge, 2024).

Another mention from this study is that the EU in its strategies focuses too much on the financial risks and benefits rather than questioning and offering measures to tackle worker replacement and disruption of labour market segments. Moreover, by rarely mentioning in its AI strategies countries other than the ones who are considered the main powers and main competitors in AI, the EU shows little to no regard for its impact outside of the Global North (Mügge, 2024). As one of its goals, EU mentions the enabling European businesses to benefit from a competitive advantage and increase consumer trust which will be the competitive edge that sets the EU apart and will make consumers want to use European AI systems rather than 'untrustworthy' systems from other countries: "current ML algorithms display some of the inputs and outputs but do not understand fully what happens in-between, and how certain outputs, including decisions and actions, are derived" (European Commission, 2018b). Mügge concludes that the EU AI policy goal, in the beginning, was not about doing good on a global level but about gaining a competitive advantage in AI. Mentions of the global impact of EU AI strategies mostly refer to the ethical superiority of the EU AI regulations.

The EU is deeply torn between two divergent sets of principles, on one side the global technological progress that with itself brings the benefit of economic growth and on the other side the preservation and safekeeping of European values through the creation of its own European-centered regulation. The EU's deep concern lies in securing its place in the global competitive environment and it aims to achieve that goal through the regulations that ensure trustworthy AI, which could bring more users to European AI systems based on consumer trust. But even though EU cooperation is central to the EU's vision of AI, still, many examples of the effort for a global collaboration are evident in the previous regulations and practices analysis.

Further, even though Mügge suggests that the primary goal for the EU AI governance has been acquiring a competitive edge, Mügge highlights that after 2021 there are gradual changes in EU AI approaches, primarily with the entering of more member states and the European Parliament into AI negotiations, within the EU AI Act discussions. When it comes to the European Parliament, in its amendments the EP stresses the global environmental and societal impacts of AI, the importance of human autonomy and the global benefits of AI even though still mentioning the need for "AI made in Europe" (Mügge, 2024).

2. Regulation of AI on the international level

Besides the enormous effort made by the European Union in order to regulate this emerging technology, it's important to shed light on the efforts made by the international community, more specifically organizations like the United Nations, UNESCO, Organisation for Economic Co-operation and Development, the Council of Europe and other international initiatives. Even though many national initiatives on the regulation of AI have been established, the rapid development of AI and its presence all around the world have shown to the international community that this technology needs to be regulated on an international level, since its use surpasses all the physical borders. International organisations have also emerged as pivotal actors in the global governance of AI. Even though, faced with the challenges of having to find a mutual solution between dozens or even hundreds of member states, international organisations have done a lot of work in the regulation of AI by creating ethical frameworks, standards, regulations, joint initiatives and others.

Therefore, the second part of this chapter aims to examine how these different international actors managed to address the issue of AI until now, what debates and negotiation processes stood behind today's results and how they motivated and influenced each other in the process of creating global norms and establishing guidelines for the responsible development of AI. The examination of these multifaced initiatives shows the role of international actors in addressing key challenges and creating initiatives which have resulted in the creation of advisory bodies which are shaping AI policies, legally binding treaties, norms on safeguarding human rights and democracy and much more.

2.1. United Nations

In October 2023, the United Nations formed a High-Level Advisory Body on Artificial Intelligence intending to foster a globally inclusive approach to AI governance (United Nations Advisory Body on Artificial Intelligence 2023). This advisory body consists of 39 members that include government officials, and academics from different countries, including Japan, Saudi Arabia, Spain, USA, and China, and tech company executives (among whom representatives from Microsoft and Sony), with the main task of addressing issues in the international governance of artificial intelligence (United Nations Advisory Body on Artificial Intelligence 2023).

In December 2023 the AI Advisory Body launched its *Interim Report: Governing AI for Humanity* which advocates for a closer alignment between international norms and the development and deployment of AI (United Nations Advisory Body on Artificial Intelligence 2023). The main part of the report is a proposal to enhance the international governance of AI by implementing seven essential functions in the international governance regime for AI (United Nations Advisory Body on Artificial Intelligence 2023) :

(1) Regular evaluation of the future directions and implications of AI
(2) Enhance the interoperability of governance efforts emerging around the world and ensure their grounding in international norms through a Global AI Governance Framework endorsed within a universal setting, such as the UN.
(3) Creating and harmonizing standards, risk management frameworks and safety protocols.
(4) Promote deployment, development and utilization of AI through international multistakeholder cooperation in order to gain economic and social benefits.
(5) Encourage international collaboration on talent development, access to computing infrastructure, the creation of diverse datasets, responsibly sharing open-source models, and utilizing AI for public benefits to achieve the Sustainable Development Goals.
(6) Incident reporting, risk monitoring and emergency response coordination
(7) Compliance and accountability based on norms

In March 2024 United Nations adopted a resolution aimed at promoting "safe, secure and trustworthy" AI systems, which also have a goal ofcreating benefits for sustainable development for all (United Nations General Assembly 2024). This marks the first time the UN General Assembly has adopted a resolution concerning the regulation of this emerging field (UN News 2024). US Ambassador and Permanent Representative to the UN, Linda Thomas-Greenfield, harbored hope that the dialogue that preceded this resolution would serve as an example for future conversations on AI challenges, such as those concerning peace and security and the responsible military deployment of AI autonomy (UN News 2024). It can be noted that the text of the resolution mentions the need for international cooperation in the regulation of AI various times (United Nations General Assembly 2024):

(1) Recognizing the potential of AI in the achievement of the Sustainable Development Goals, the resolution emphasizes the urgent need to achieve global consensus on

artificial intelligence systems that have to be trustworthy, safe and secure. Further, it stresses the need for inclusive international cooperation through the creation of safeguards, standards and practices in order to prevent the fragmentation of the governance of artificial intelligence systems and promote innovation. Moreover, the resolution highlights different levels of technological development between and within countries that need to be reduced and the challenges that developing countries are facing in keeping pace with technological development. Therefore, in order to eliminate digital gaps between countries, the resolution emphasizes the urgency of strengthening capacity building and the need for technical and financial assistance for those developing countries to close digital divides between and within countries' representation in international processes and forums on AI governance.

(2) The Resolution further asks Member States and multi-stakeholders worldwide including international and regional organizations, the private sector, civil society, academia, research and technical institutions and communities and also individuals to participate in the regulation and governance of AI systems by creating partnerships and cooperations.

(3) The Resolution also encourages international research and cooperation to comprehend the potential benefits and risks that AI systems could have in closing digital gaps and achieving Sustainable Development Goals and expanding digital solutions, such as open-source artificial intelligence systems

(4) The Resolution also encourages international research and cooperation to identify the effects of AI systems on labour markets and offer assistance to mitigate potential negative impacts on workforces, especially in developing countries. It also highlights the need of programmes for digital training, innovation and enhancement of benefits of AI systems.

(5) The Resolution also highlights the importance of responsible, fair and inclusive cooperation on international data governance for the development and operation of AI systems. It also urges Member States to promote and share their best practices on data governance to advance trusted cross-border data flows.

(6) The Resolution notes that the United Nations system is contributing to reaching global consensus on the regulation of AI systems by promoting inclusive

international cooperation which is aligned with international law, precisely with the Charter of the United Nations; the Universal Declaration of Human Rights and the 2030 Agenda for Sustainable Development

The Resolution also calls for the private sector, public sector, specialized agencies, funds, and related organizations of the United Nations system, programs and other entities and bodies, civil society, academia and research institutions to address the challenges jointly (United Nations General Assembly 2024).

Throughout the resolution, it can be noticed that there was a tendency to commend the good practices of other international actors and organisations. For example, it was acknowledged in the resolution the importance of the efforts made by the International Telecommunication Union, in collaboration with 40 United Nations bodies, to convene the Artificial Intelligence for Good platform. It also mentioned its annual AI for Good Global Summit and the launch of the International Telecommunication Union's Artificial Intelligence Repository, whose goal is to identify responsible and practical AI applications to advance the Sustainable Development Goals.

It also highlighted the adoption of the Recommendation on the Ethics of Artificial Intelligence by the General Conference of the United Nations Educational, Scientific and Cultural Organization on 23 November 2021. Further, it also mentions the adoption by the General Conference of the United Nations Educational, Scientific and Cultural Organization of its Recommendation on the Ethics of Artificial Intelligence of 23 November 2021 (United Nations General Assembly 2024).

2.2 Organisation for Economic Cooperation and Development (OECD)

In May 2019, the OECD Member States together with six partner countries adopted common ethics principles on AI - OECD Council Recommendation on Artificial Intelligence. The recommendations were drafted by an expert group together with representatives of Member States, and constituted 'the first set of intergovernmental policy guidelines on AI, agreeing to uphold international standards that aim to ensure AI systems are designed to be robust, safe, fair and trustworthy' (OECD 2019a). In its Recommendation, OECD highlighted the following principles:

(1) inclusive growth, sustainable development and well-being;

(2) Respect for the rule of law, human rights and democratic values, including fairness and privacy;
(3) Transparency and explainability;
(4) Robustness, security and safety;
(5) Accountability.

This document also serves as a representation of the cooperation between the European Union and OECD since the European Commission was one of the expert group's members, which explains how the OECD guidelines heavily draw upon the concept of 'Trustworthy AI' as developed by the European Commission's High-Level Expert Group on AI (Smuha 2021). The Recommendation also highlights the importance of international cooperation for trustworthy AI by:

(1) active cooperation on the advancement of these principles between governments, including developing countries and stakeholders;
(2) Governments should work together within the OECD and other global and regional forums to promote the sharing of AI knowledge, and by this encourage international, cross-sectoral, and open multi-stakeholder initiatives to develop long-term AI expertise;
(3) promotion of the development of the global technical standards;
(4) Governments should promote the development and their own use of internationally comparable indicators to measure AI research, development, and deployment and gather the evidence needed for the implementation of these principles. In June 2019, a set of ethical principles for AI was endorsed by the G20, based on the work of the OECD (Smuha, 2021).

2.2.UNESCO

In November of 2021, UNESCO published the first-ever global standard on AI ethics – the 'Recommendation on the Ethics of Artificial Intelligence' which was adopted by all 193 Member States (UNESCO, 2021). The main focus of this Recommendation is on the protection of human rights and dignity and the importance of human oversight of AI systems. This Recommendation is addressed to Member States but it also gives ethical guidance for all AI actors, including the ones in public and private sectors.

As its aims, the Recommendation mentions bringing a globally accepted normative instrument that focuses on values and principles but also on their practical implementation through concrete policy recommendations, with a significant focus on gender equality, environmental protection, and ecosystem preservation.

As its objectives, the Recommendation on the Ethics of Artificial Intelligence lists the following:

(1) to provide a universal framework of values, principles and actions to guide States in the formulation of their legislation, policies or other instruments regarding AI, consistent with international law;
(2) to guide the actions of individuals, groups, communities, institutions and private sector companies to ensure the embedding of ethics in all stages of the AI system life cycle;
(3) to protect, promote and respect human rights and fundamental freedoms, gender equality, human dignity and equality, including gender equality; preservance of the environment;and cultural diversity;
(4) to foster multi-stakeholder, multidisciplinary and pluralistic dialogue and consensus building about ethical issues relating to AI systems;
(5) to promote equitable access to developments and knowledge in the field of AI and the sharing of benefits, with particular attention to the needs and contributions of LMICs, including Least Developed Countries, Landlocked Developing Countries and Small Island Developing States (UNESCO, 2021).

When it comes to the attitudes of the United States and China, it's important to mention that the US is not part of UNESCO and is not a signatory of this Recommendation (Politico, 2021a). China, on the other hand, did sign off on this pledge which includes the banning of the use of AI for "social scoring" which was created and put into practice by the Chinese government. According to Politico, this is also the first time China signed up to principles which include the end of AI mass surveillance (Politico, 2021b).

2.3. Council of Europe (CoE)

In September 2019 the Council of Europe has established an ad hoc committee on AI, called CAHAI, which had a task to examine the possibility of creating "a legal framework for the development, design and application of artificial intelligence, based on the Council of Europe's

standards on human rights, democracy and the rule of law" (Council of Europe, 2019). With this, the Council became the first multilateral organization to announce its intention to examine the adoption of binding rules for AI (Smuha, 2021). This Committee has been succeeded by the Committee on Artificial Intelligence (CAI).

On 17 May 2024, the CoE adopted the first ever international legally binding treaty at the 133rd session of the Committee of Ministers at Strasbourg - Council of Europe *Framework Convention on artificial intelligence and human rights, democracy, and the rule of law* (Council of Europe, 2024). The importance of this treaty in the global governance of AI is reflected in the fact that, unlike the EU AI Act, the treaty is accessible to non-European countries as well, establishing a legal framework encompassing the entire lifespan of AI systems and tackling potential risks, while promoting responsible development and innovation, with the respect to human dignity, democracy and rule of law (Council of Europe, 2024).

During the process of negotiation, there was a risk of a diplomatic blockage led by the United States, Canada, Japan and the United Kingdom (Politico, 2024b). According to Politico, they intended to exclude their leading private companies from the scope of this treaty. These countries are home to one of the world's most powerful technological companies, especially the US which has companies like Google, Microsoft and OpenAI. Even the US government failed to agree on binding AI regulations, because of their consequent demand for an opt-in model for their companies. Even if the US, Canada and Japan are non-voting observers at the CoE, without their ratification this initiative would have quite limited influence on a global level.

According to Politico, the EU (whose Member States at that period had just agreed on the EU AI Act) said that excluding these private companies from the scope of the treaty would represent "diminishing its value and sending a wrong political message that human rights in the private field do not merit the same protection" (Politico, 2024b). The negotiations on the regulation of the private sector have made the text of the treaty significantly weaker in comparison to its original version, in the end it was agreed that the Convention covers both the public and private sectors, but still allows states to choose how they will implement regulations to the private sector (Council of Europe, 2024).

3. Conclusion

This chapter clarified the intricate relationship between human rights and AI within the international regulatory framework through the analysis of the regulations and policies both on the European and international levels. Examining how AI impacts policymaking on the international level has made it evident that instead of creating an AI race to achieve economic and technological superiority, it's crucial for human rights that the international actors approach the regulation on a more cooperative level since many of the challenges cannot be solved either on a national or regional level.

The challenges to achieving this global governance of AI are numerous. They can go from different regulatory approaches and priorities between different countries and regions to complex ethical considerations that require countless negotiation rounds, over geopolitical landscape and competitiveness to the rapid pace of technology development and its unpredictability. Despite all this, with the creation of the European Union Artificial Intelligence Act – the first regulation on Artificial Intelligence, and the Council's Framework Convention on artificial intelligence and human rights, democracy, and the rule of law – the first legally binding international treaty on Artificial Intelligence, it became evident that the international community has the capacity for the global governance of AI.

CHAPTER III: AI FOR GOOD

Often when Artificial Intelligence is mentioned the connotation behind it is negative and it usually refers to the dangers and challenges AI can produce and the catastrophic scenarios depicting the "end of humanity." While being justified to some extent, these assumptions overlook the other side of AI — one that can bring benefits both to individuals and society as a whole. So, rather than focusing on the negative impact AI can have on human rights, this chapter will delve into examples of AI used as a tool in the protection of human rights through human rights monitoring and reporting, violent conflict prediction, early warning systems and climate change. In situations when human rights are being violated or are possibly in danger, it's necessary to use all tools that would benefit their protection, one of which could be Artificial Intelligence. As AI becomes more present and integrated into various sectors, it will inevitably extend into the realm of human rights, and potentially offer transformative potential in governing human rights on the international level. Therefore it's crucial to know how and when AI can be deployed as an option and what actions international actors are taking in seizing this opportunity.

Therefore, after analyzing the risks and challenges that AI can impose on human rights in the first chapter and the global governance of AI in the second chapter, the third chapter will delve into possible benefits AI can provide when implemented for the protection of human rights by international actors. The goal of this chapter is to change the negative one-sided understanding of AI by shedding light on its potential as a tool for protecting and promoting human rights. This chapter aims to prove how AI can be effectively implemented into existing practices that tend to protect human rights, such as human rights monitoring and reporting, forecasting of violent conflicts, early warning systems and climate change. This was done through various analyses, reports and case studies which show the current use of AI for good. By showing these positive examples, this chapter seeks to promote the appreciation of AI as a potential in the protection of human rights and not as a threat to them.

AI can be described as a "technology focused on automating specific tasks that normally require or involve human intelligence when being performed" When these functions are applied to human rights monitoring, reporting, conflict prediction, danger detection, etc. AI is usually used for data collection and analysis, computational capabilities, forecasting, satellite imaging, decision-making, sound processing, etc. (Dulka, 2022).

When it comes to the international actors, among numerous that already started using AI, a couple of them stand out with their initiatives and achievements. Amnesty International is engaged in multiple endeavors and was pivotal in launching the first-ever initiative which combined machine learning and human rights monitoring through a collaboration between NGOs and technical experts (Cornebise et al., 2018). In this chapter, the involvement of Amnesty International is shown through 2 case studies, firstly in quantifying village destruction in Darfur in collaboration with academics from the University College London and the University of Amsterdam and with the help of volunteers all around the world, which finally resulted in the publishing of a whitepaper promoting the advantages and potential of integrating satellite imagery and machine learning (Marin et al. 2020). The next case presented in this chapter is on media monitoring used for tracking death penalty cases together with AI company Element AI, funded by the Canadian government (Amnesty International & Element AI, 2019). And a third case on using machine learning to track abuse against women on Twitter, in which Amnesty collaborated with Element AI by using machine learning to examine digital harassment targeting women in the United States and United Kingdom (Amnesty International, n.d.).

Besides these cases, Amnesty has also combined machine learning and geospatial analysis in Mexico in order to assist local organizations in their search to find missing people (Panic 2024). Amnesty was also engaged in reconstructing events in Gaza after the kidnapping of an Israeli soldier together with Forensic Architecture and SITU Research. Amnesty had a task to create a 3D model of Rafah by using collected video and image proofs and testimonies, for the visual reconstruction they used machine learning but at that time most of the work was still done manually and required a lot of time and resources (Panic 2024).

Further examples come from the Center for Human Rights Science at Carnegie Mellon University, which developed a tool Event Labelling Through Analytic Media Processing (E-LAMP) that was a combination of machine learning and computer vision with a function to analyzed large volumes of videos and conducted speech recognition (Panic 2024). This tool was also used in Ukraine in 2013 and 2914 for the reconstruction of the protests in which members of police forces were hurt and killed (Panic 2024).

The chapter starts with defining how AI can contribute to more efficient human rights monitoring and reporting, followed by FIVE case studies that showcase the hands-on implementation and

evident benefits of AI in this area. By focusing on these case studies, the research shows how AI is currently being deployed by civil society organisations and in what way are the mechanisms of monitoring and reporting being improved by AI. Subsequently, the focus switches to the use of AI in violent conflict prediction, specifically through the example of the Violence & Impacts Early-Warning System (ViEWS) and the detection of danger through the early warnings system. Next, the attention shifts to the use of AI in tackling the challenges of climate change. Last but not least, the chapter ends by shedding light on the opportunities of AI for society and the conclusion of the chapter. During the research done on this topic, most of the cases referred to the seminal work by Dulka (Dulka, 2020), due to the fact that scholarly work on this topic is still underdeveloped due to the recent emergence of these examples.

1. Human rights monitoring and reporting

In order to advance mechanisms of human rights monitoring and reporting, international actors have started using the help of technology, specifically AI models. As mentioned before, there are numerous benefits that AI can provide to more successful human rights protection. There are numerous areas in which AI can be helpful and contribute to more efficient monitoring and reporting on human rights:

- To compile information and draft reports (Dulka 2022)
- It can notice and predict trends useful for the strategy planning of different international actors and states (Dulka 2022)
- AI can enhance organizations or states' capacity to comprehend information and data by providing deeper insights across demographic markers (Dulka 2022)
- The AI can better track information over time and do it in a timely manner (Dulka 2022)
- It can analyze data in order to find out how rights are enjoyed or how are they violated (Dulka 2022)
- Instead of being limited to small and unrepresentable sources of data, AI models and machine learning models could analyze big open-source databases in order to make decisions on more representative data and keep better track of human rights violations (Littman et al. 2021) The complexity and the amount of data sometimes can be overwhelming for humans or impossible to examine in a timely manner, therefore, the

help of technology is more than needed (Littman et al. 2021) AI can examine and select needed data and notice trends that otherwise might not have been noticeable to human operators (Littman et al. 2021)

- While analyzing data, AI algorithms can find and reveal gaps and point out inconsistencies in the examined datasets, which helps initiate possible areas that need improving and critical information that's missing which is crucial for decision-making (Littman et al. 2021). This would be greatly beneficial for human operators to make informed decisions and see the connections between different data sets more quickly thanks to the AI (Horowitz et al. 2018).
- By cross-verifying big data sets and reports, AI can further improve human rights reporting. In this way, the connections between different reports can be made, but also it can notice inconsistencies or differences between them. In this way, the reliability of the results of findings is increased and therefore contributes to more informed decision-making (PricewaterhouseCoopers Consulting 2020)

Therefore, there are many reasons and benefits that show why AI should be used in human rights monitoring and reporting both by civil society organizations and by states. The impact of these reports is not only in informing about human rights violations but also in holding states accountable for the violation of the same rights (Dulka 2022). AI technology in human rights reporting has mostly been used by civil society organizations in order to monitor, track and report on human rights, to hold states accountable for the violation of those same rights and to submit reports to public or international human rights treaty bodies (Dulka 2022). They use AI for immediate capture of human rights violations, investigation and detection effors, for remote sensing data, for small and big data analysis, and for noticing and examining patterns and trends that otherwise might not have been detected by humans and their traditional ways of data collecting and processing (Panic 2024).

Technological advancements have allowed for data to be gathered through satellite and drone images for which the "data collector" doesn't even need to be anywhere close to the area of the possible or ongoing human rights violations, which creates less danger compared to the case when the human data collectors have to be physically present. Thanks to these images, it's possible to detect human rights violations such as labor campuses, destruction of living areas,

forced displacement etc. (Panic 2024). Combined with the data gathered by the witnesses of these situations a better understanding of the situation could be provided since the analysts would work with proofs from different sources and different types of technology.

Integrating AI in human rights monitoring and reporting could present an evolutionary advancement towards the efficacy of these mechanisms. If human rights organisations could without limits use AI for data gathering and data analysis, this could enable much bigger coverage of human rights violations and much more probability that the actors responsible for these violations will be held accountable. The use of AI in strengthening accountability efforts, quicker response to violations and broadening of data sources can be promising and impactful for the future protection of human rights. As this technology continues to evolve, its integration in human rights monitoring and reporting represents a transformative potential to uphold fundamental rights worldwide.

1.1 Quantifying Village Destruction in Darfur

For example, a groundbreaking initiative started in 2018 that for the first time combined machine learning and human rights monitoring by NGO partnership between civil society and technical experts (Cornebise et al., 2018). This initiative intended to quantify the village destruction in Darfur and was organized and conducted by Amnesty International together with academics from the University College London and the University of Amsterdam (Cornebise et al., 2018). The destruction in the village was detected by the use of an AI algorithm designed to analyze satellite imagery and classify the extent of damage by the use of multi-task binary classification (Cornebise et al., 2018). The importance of this initiative is also in the inclusion of the public. For the algorithm to be trained by AI experts, Amnesty provided open-source data which was collected during their campaign Eyes on Darfur and consisted of 2.6 million satellite images and labeled by 28,600 volunteers (Cornebise et al., 2018).

This model has proven to be very successful in this task, firstly by reducing the time and resources but also by providing higher accuracy (Dulka 2022). This machine-learning model was later used to identify destruction in other habitats (Cornebise et al., 2018). After the end of the project, in 2019, Amnesty published a whitepaper promoting the benefits and potential of the combinational use of satellite imagery and machine learning, as well as its challenges. Their main challenges were the need for data validation, sensitivity and transparency of data,

specifically images and graphics, and risks of malicious attacks and dissemination, and the risk of data getting in the wrong hands which could be used for future targeting (Marin et al. 2020). In order to prevent possible dangers and strengthen the responsibility, the report invites other NGOs to work together in addressing these challenges (Cornebise et al., 2018).

Engaging in such projects creates two-sided results. On the one side, it raises awareness of the use of AI for good, and it demonstrates the benefits that artificial intelligence brings to human rights monitoring, such as speed, accuracy, and public engagement. On the other hand, it shows the need for the cooperation of different international actors in order to use AI in the best way possible and avoid risks such as data sensitivity, transparency and protection, risks of malicious activities, etc. The model used in the Darfur case was also tested and showed success along the border between South Sudan and Uganda, which shows the applicability of this model and the ability to be adapted to different places and situations (Marin et al. 2020).

Engaging in initiatives inspired by partnerships between NGOs and technical experts shows the transformative potential of AI also in the detection of human rights violations beyond mere destruction. The approach in this example not only made the results more accurate, reduced time and resources but also encouraged collaboration among different international actors.

1.2 Media Monitoring Used for Tracking

Amnesty International has another example of the use of AI in practice, this time in cooperation with an artificial intelligence company Element AI, funded by the Canadian government (Amnesty International & Element AI, 2019). In 2018, they joined in the effort to develop a tool in order to track media information on death penalty cases (Amnesty International & Element AI, 2019). Before this tool, all the data was collected and input manually by volunteers, automating a significant part of this process and therefore significantly reducing the time, something that for a volunteer would take hours this tool did instantaneously (Amnesty International & Element AI, 2019). Although human intervention remains necessary in this process to authenticate, verify and correct errors, these results obtained with the help of an AI tool expedited the overall process.

1.3 Monitoring ethnic violence in Myanmar

Element AI also partnered with Human Rights Watch intending to track human rights violations in Myanmar (Dulka 2022). This partnership started in 2017 intending to create a machine-

learning tool that would help in the tracking of human rights violations of Rohingya populations (Dulka 2022). This tool had a task to identify, track and document human rights violations by using satellites and remote sensing thermal data (Dulka 2022). Data was collected by tracking smoke, fire and destruction, in which AI had an important role by combining the data collected by thermal monitors with satellite images and also with data from social media of individuals who were present at the scene (Salian, 2019).

In this case, the significance of remote sensing in human rights monitoring is shown as essential for a better collection of data in areas that perhaps can't be accessible or are dangerous to access. The traditional human rights investigation requires data collectors and researchers to directly enter this zone and collect information, which is often impossible or puts their lives in danger. In this way, civil society organisations, like Human Rights Watch in this case, have much-needed technical expertise which helps them to enhance monitoring capabilities, improve the accuracy of the reports, and collect data promptly and without putting the lives of their researchers in danger. Further, this way of using AI for reports also can mean that the organization has wider geographic reach since the physical presence of the researchers is not required – which is beneficial for human rights and means that even previously unreachable areas can now be included in the process of human rights reporting. In this way, areas such as conflict zones or closed countries become reachable for human rights monitoring (Dulka 2022).

Moreover, this way of the utilization of AI shows how AI can also contribute to higher transparency and accountability (Dulka 2022). Civil society organisations, like Human Rights Watch in this case, do the process of monitoring and reporting in order for them to hold states accountable for the violations of human rights (Dulka 2022). Data gathered and analyzed in this way means that state authorities cannot restrict its gathering or influence the outcome of the report by altering evidence because the entire process was conducted by civil society and independent organizations in the areas where that option was previously impossible (Dulka 2022). Also, the type of data analyzed makes it hard to alter since the satellite imaging, thermal data and social network inputs by citizens. This provides organizations with the possibility of conducting objective, unbiased monitoring and being sure that the collected information has not been altered by the state. Also, this means that denying accusations by states will be much

harder. Therefore, this type of AI implementation not only serves the safety of researchers but also provides higher accuracy and efficiency in monitoring and reporting.

1.4 Using Machine Learning to Track Abuse

After previous cooperation, Element AI and Amnesty International worked on another project together in 2018. This case was specific because it used machine learning to analyze the Twitter online abuse against women in the United States and the United Kingdom (Dulka 2022). As a result of this project, Element AI designed the "Troll Patrol" report which explained how AI contributed to the results (Amnesty International 2018).

The data-gathering part of the project included politicians, journalists and volunteers and technical experts, a data analysis was conducted by AI models, that also filtered findings on the basis of demographic markers (Dulka 2022). In the process of the final evaluation of the performance of the tool used, they concluded that the AI model was giving good results but not in comparison to human data experts: "The AI was able to correctly identify 2 in every 14 tweets as abusive or problematic in comparison to experts who identified 1 in every 14 tweets as abusive or problematic" (Amnesty International 2018). The result of this project was the world's largest crowdsourced dataset about online abuse against women. Based on these conclusions, Amnesty published recommendations aimed at the examined actors, separately for Twitter and states.

The recommendations for Twitter include (Amnesty International, 2018b):

(1) Publishing of the efforts and actions in which they show to the public in which way they are addressing and handling violence and abuse on the Twitter platform;
(2) Explaining and simplifying the reporting process in which they transparently show how decisions on the content restriction were made, and ensuring that those decisions are in line with human rights law;
(3) Clarifying the procedures taken for handling abuse are dealt with and deployment of the moderators;
(4) Enhancing security measures, and privacy, and addressing other safety risks or features.

These are the recommendations to states that include:

(1) Implementation of legislation measures in order to combat the pervasive issue of the abuse of women online;
(2) Allocate funding to programs that would offer better education and training of the state law enforcement on the issue;
(3) Create education campaigns for the public about abuse online and the promotion of gender equality more broadly; and
(4) invest in publicly available services or programs tailored to support women who have experienced abuse online.

Civil society organisations are conducting this kind of project in order to shed light on the violation of human rights but also to keep private actors and the state accountable for the respect of fundamental human rights. This collaboration between Element AI and Amnesty International resulted in the largest crowdsourced collection of its kind and laid the groundwork for other international actors to follow by using AI for

Amnesty specifically uses international human rights law and the obligations of states and private actors as a way to push for accountability and policy change based on the information that the AI and machine learning tools were able to identify and report on. More broadly, this case study provides a salient example of how AI might be used not only to evaluate the status of women and shed light on challenges women face but also to push other actors for accountability in their treatment of them. By utilizing AI in this capacity,

2. Using AI to Forecast International Displacement

AI has also proven to be useful in forecasting international displacement. In 2021 and 2022 the Danish Refugee Council (DRC) employed AI and machine learning to project displacement trends (Danish Refugee Council 2021). Based on the ability of AI to make sense of data and identify patterns, correlations and trends, it was possible to correctly forecast how many people would be displaced in the years ahead. The AI tool which was created thanks to the funding of the European Union, has collected and analyzed data on 148 indicators that lead to international displacement, including conflicts, governance, climate, violation of human rights, societal trends and many more, out of which, conflict was identified as one of the major reasons for displacement (Dulka 2022). Besides the involvement of the EU, other international actors contributed to the success of this project even indirectly. AI models that were used to predict

displacement were trained on open-source data from major international organizations and actors, such as the International Monetary Fund and World Bank (Dulka 2022).

It's proven in this case that AI's capability of making sense of information and creating patterns while analyzing big data is beneficial since these patterns likely wouldn't have been detected by humans (Dulka 2022). For example, in this case, the model used was also able to pair conflict, which was found to be the major reason for displacement, with other indicators and predict different forecasts thus creating novel correlations that were not immediately noticeable (Dulka 2022). Using AI for creating models and predicting future trends leads to more efficient and better strategic planning and response, but also more effective allocation of resources, that are typically limited (Dulka 2022). This forecasting tool can also be easily replicated and adjusted for it to be used in other needs beyond just displacement, therefore it can help in advocating for increased humanitarian assistance in regions where the models forecast high levels of displacement (Dulka 2022). This model also has its limitations, out of which the biggest one is the incapability of including unpredictable events and unexpected developments in the geopolitical sphere (Dulka 2022).

After the end of the project, the Danish Refugee Council made a report for the European Union, that financed the project, which contained recommendations for the assessment of displacement with the help of AI models and recommendations for future humanitarian response and forecasting (Danish Refugee Council 2021).

3. AI and conflict prediction

Conflicts are in their nature complex and unpredictable. The vast majority of countries have them, in one way or another, and they often arise from numerous interconnected factors, such as economic issues, political and religious tensions, territorial disputes, historical grievances, social injustice, etc. The consequences of violent conflicts result in the deaths of thousands of people every month across the globe and forcibly relocate even more, they can create or deepen poverty, undermine the development and weaken the functioning of political systems (Hegre et al. 2019).

In democratic and peaceful societies, conflict is often handled through local and national institutions, and at the international level, they are handled through diplomacy, international treaties, negotiation and the involvement of regional and global institutions (Panic 2024). These

institutions support societies in resolving conflict in a peaceful manner and often provide effective governance to them, but given the complexity of conflicts, sometimes these instruments and institutions are not sufficient to resolve them, leading to an escalation of the conflict such as increased violence, humanitarian crisis, intensification of hostilities, etc. Therefore, in order to prevent these escalations different actors tend to predict conflicts.

Based on the report of the United Nations and World Bank, *Pathways for Peace: Inclusive Processes to Preventing Violent Conflict* there are many drivers of violent conflict, so their identification is not an easy process, and therefore a simple or unique formula for the prevention of conflicts doesn't exist (United Nations and World Bank, 2018). Based on this report, prevention can be defined as the "avoidance of the outbreak, escalation, recurrence, or continuation of violent conflict" (United Nations and World Bank, 2018). This report also mentions structural factors that lead to violent conflicts such as inequality, low trust and insecurity, still, they alone are not enough to predict the circumstances and timing of escalation and violent conflict (Panic 2024). Right now, it's considered that a key challenge is how to predict a conflict in societies that have been peaceful, since in the opposite situation, in countries that had a previous history of conflicts, it's more or less expected for them to emerge again and therefore predict them. (Panic 2024). Policymakers and first responders would have significant benefits if there was a possibility to predict escalations in areas where it's not expected for conflict to emerge.

Based on the UN report, there is a pessimistic tone in the public discourse when it comes to the effectiveness of the international actors in effectively predicting violent conflict (United Nations and World Bank, 2018). The question is, can AI help?

Since new AI technologies are emerging daily, and current uses of AI continue to expand deeper and deeper into the field of human rights, conflict prediction could be another area that explores the use of AI. As shown previously in this chapter, international actors are delving deeper into the potential use of AI in the protection of human rights with a goal to help prevent a crisis, avoid human suffering, prevent damage to the economy, democracy and societies, or mitigate the potentially devastating impacts of conflicts. Social scientists all over the world have started tapping into the potential of technology to predict the place and time of possible violent conflict (Panic 2024). The reason why many social scientists haven't approached AI models before is

that big data alone cannot be the only transformative force of conflict prediction study and practice, the rigorous work and testing done by social scientists is still needed. A good number of them still rely on traditional well-known sources of data such as the Political Instability Task Force (PITF), the Armed Conflict Location and Event Data (ACLED), and the Uppsala Conflict Data Program (Panic 2024). But also, an increasing number of them have started using data sets that rely on AI, specifically in natural language processing and text classification techniques, for example, the AI-driven Global Database of Events, Language, and Tone (GDELT) (Panic 2024).

3.1 Violence & Impacts Early-Warning System (ViEWS)

A review from the International Journal of Forecasting from 2023 showed that half of the forecasting systems that were analyzed are already using machine learning algorithms to detect patterns and produce forecasts (Rød, Gåsste, and Hegre 2023). One good example of researchers integrating innovative techniques and overcoming past failings is the Violence & Impacts Early-Warning System (ViEWS), a collaborative effort between the University of Uppsala and the Peace Research Institute–Oslo, which is designed to forecast the likelihood of political violence in Africa and the Middle East (Panic 2024).

ViEWS's primary objectives are transparency and replicability, therefore the main principles lying in its process of development are "public availability, uniform coverage, transparency, and methodological innovation" (Hegre et al. 2019). ViEWS is ensuring its transparency by only using data that are available to the public, publishing all information on its website and encouraging stakeholders to use the replication material (Hegre et al. 2019). ViEWS sets an example for high-resolution prediction of conflicts available to the public, even though the AI models that ViEWS is using have been trained only on data based on Africa and the Middle East, this model can serve as an inspiration model for other parts of the world (Hegre et al. 2019). Public accessibility of these findings and transparency are beneficial both for domestic and international stakeholders, especially NGOs who could learn about the use of AI in all stages of forecasting. These forecasts are developed based on the findings of decades of quantitative analysis on peace and conflict, in areas of the economy, politics, geopolitics and history (Hegre et al. 2019).

Panic considers that the future of using AI in conflict prediction can go in 3 different directions (Panic 2024):

(1) use of AI in detecting the factors that contribute to a society becoming peaceful or to understand why some societies are peaceful and some not (Panic 2024).
(2) AI will advance enough that the research can include not only small data sets but also big-data ones (Panic 2024).
(3) To approach the prediction by putting more limits on the examination such as analyzing smaller geographical areas in shorter time periods instead of whole countries during a long period of time, which would lead to higher credibility (Panic 2024).

In conclusion, predicting violent conflicts still doesn't have a perfect solution, and it's a question if it ever will. But, efforts in employing AI are still greatly needed based on the positive contributions to the protection of human rights so far.

3.2 Predicting conflict and detection of danger

Inside the broader violent conflict prediction is also included the early warnings system that danger is close and that people need to escape the place or seek for protection (Panic 2024). An example that could better give insight into how this system works is based on the application used in Syria developed by Hala System, a US-based company (Loveluck 2018). The Hala System Sentry application has the possibility to provide early warnings of airstrikes by detecting aircraft activity and by providing users the possibility to evacuate and seek safety on time (Loveluck 2018). When there is a potential bombing, even minutes are enough to save people's lives, this app warns people of potential danger by sending textual messages and spreading across social media channels containing the chances of an air strike, potential target and place where the attack will probably occur, but also with sound alarm (Loveluck 2018).

In order to develop the system they needed "plane spotters" – plane observers, who would detect aircraft activity and insert the information into the application, usually the ones chosen for this task were teachers, engineers and often even regular citizens like farmers spread across Syria (Loveluck 2018). The insertion of data is done in a rather simple manner, by watching the skies in shifts and entering the information on the aircraft they see into the application, which would then be compared and refined by AI program with data from remote sensors hidden across the country which capture the sounds of aircraft, aiding in identifying the type and speed of the planes (Panic 2024).

This example shows the importance of not only predicting violent conflicts in the upcoming months or years, like in the example previously shown but also the importance of "early warnings" systems, such as the Hala system. These mechanisms rely greatly on local people without whom it probably wouldn't be possible to get this real-time information. On top of the contribution of local people in order to get the final warning it's necessary to also use natural language processes, remote sensors, artificial intelligence data processing and refining and similar, which contribute to the precision and speed of the warnings, which could save lives by giving people a chance to seek for protection and safety (Panic 2024).

The Hala System isn't the only example of using the early warnings system, Panic provides examples of initiatives that function in the similar way (Panic 2024):

- The Economic Community of West African States (ECOWAS) in order to protect stability of the region, also has networks of local people across West Africa that also contribute to the early warnings system based in Nigeria which is designed to analyze threats not only in conflict situations but also with potential political threats and natural disasters (Panic, 2024).
- The UN Refugee Agency's (UNHCR) Project Jetson experiment which is particularly especially used in situations of mass displacement, also involves artificial intelligence. The main task of the experiment is to improve the processing of asylum seekings and refugee applications. This experiment is also tackling issues like food insecurity and violence to forecast cases of displacement in Somalia (Panic, 2024).
- The Danish Refugee Council has also developed and implemented a machine learning prediction model intending to forecast displacement from one to three years in advance. When those models detect possibility of displacement, international actors such as governments or the United Nations are informed so they can react to the crisis in a timely manner and maybe even prevent it by addressing the root causes of the displacement (Panic, 2024).
- An example that could use the benefits from the contribution of local people is the Kivu Security Tracker. This system is used in eastern Congo and is developed by the Congo Research Group (CRG) and Human Rights Watch. The system is based on the

contribution of human monitors, local reports, interviews, and incident documentation, but which could be further improved by the higher use of AI (Panic, 2024).

As shown in these examples, AI cannot be used in the entire process of the early warning systems but it certainly does contribute to the speed and preciseness of getting the time-sensitive results. AI contributes to faster analysis of data, connecting different big data sets, by reducing the time and resources used, providing more precise prediction and reducing errors.

4. Climate change and AI

Climate change is one of the most impactful challenges to contemporary society, creating resource scarcity, natural disasters, geopolitical tensions, environmental refugees, governmental challenges, etc. (Coeckelbergh 2021). The effects of climate change can't be denied anymore, and international actors are urged to action in order to prevent what is though to be one of the most significant challenges for humanity—if not the most significant (Coeckelbergh 2021). These challenges imposed by climate change are not only causing natural disasters but also present a threat to international peace and security by being connected to the emergence of violence and armed conflict, since climate can worsen political, social, and demographic conditions (Panic 2024). The ones most affected by these challenges are the most vulnerable parts of global society - developing and conflicted states (Panic 2024). In order for scientific research to be able to follow the rapid changes in climate change and tackle it's challenges, the use of technology has become crucial, in terms of big data analytics, collection of large amounts of data and research (Panic 2024).

Today many researchers are exploring the connection between conflicts and climate change, there are many disagreements between them about the exact correlation of these two, on the first look, distinct topic, there is a mutual agreement that strong climate shocks need to be effectively managed on the local and international level in order to avoid the risk of the emergence of violent conflicts (Panic 2024). Besides violent conflict, climate change also poses a threat to economic development which is crucial for developing societies, especially the ones relying on agriculture (Panic 2024).

In order to tackle this challenge, technology is necessary in order to make easier the process of data collection in areas that are inaccessible or dangerous to approach because of natural

disasters. As previously examined, researchers and civil society organisations are already using AI and machine learning in order to predict conflicts. The models used for this purpose could also be adapted by having climate change factors incorporated into them, and be useful for the timely informing of decision-makers about possible conflicts, humanitarian crises and security escalations (Panic 2024). The use of AI could especially be useful in areas that are unapproachable not only because of climate change challenges but also in societies that already have security issues, such as war zones or conflict-affected areas Technology, and specifically AI, can be helpful in the automated collection of real-time data, tracking of online discussions on climate change and possible signs of the escalation of protest and violence by AI natural language processing, creation of geopolitical datasets that show the real-time vulnerability of affected societies, trends and risks (Panic 2024). Satellites can be used to track weather, temperature variations, movement of military forces, destruction of nature and infrastructure and many more (Panic 2024).

To many, climate change already represents the biggest crisis of our time and the most urgent challenge humanity needs to face based on the catastrophic consequences it can cause to our planet (Panic 2024). The question of why international community has still not fully grasped on the potential benefits in the usage of AI in tackling climate change could be the fact that, on the global perspective, some societes are not “affected” neither by AI nor by climate change (Coeckelbergh 2021). Those are mostly societies that already have enormous domestic problems, such as poverty, conflicts, lack of water, food insecurity and other challenges critical to basic survival (Coeckelbergh 2021). Climate change is a global problem and it will be more and more obvious as the time passes by and the consequences become visible everywhere. Some scholars consider that it’s important to “negotiate the distribution” of available resources and activism between countries that have the capacity to do that and between those who don’t have it, which are most probably societies that struggle with challenges already threatening to cause humanitarian catastrophies (Coeckelbergh 2021). Coeckelbergh highlights that “without taking such a wider global political perspective and without addressing these matters of global justice, the discussion about AI for climate may well be perceived as a neo-colonial hobby” (Coeckelbergh 2021).

These examples show the importance of the integration of AI which offers numerous benefits for tackling the challenges to human rights imposed by climate change. The capacity that AI has in the analysis of data gives organizations the necessary advantages in monitoring climate-related threats, and therefore the greater opportunity for the protection of human rights. Further, international actors can advocate more effectively by using AI-derived data to show potential risks and promote global sustainability initiatives.

5. The opportunities of AI for society

It's no longer a question whether AI has an impact on society but more on how we can control and use this impact in the favor of society. Floridi provides four main opportunities that are offered to society by AI (Floridi 2021):

(1) Who we can become – AI can give people the opportunity to thrive by implementing AI possibilities to improve their own skills, projects and others, and with that offering personal growth. The use of technology in everyday life, such as the use of the washing machine, dishwasher, smartphones, etc. may offer people more free time dedicated to things that are in their spheres of interest and would give them greater satisfaction or ease some difficult and complicated processes. Further, jobs that previously required long manual work can now be solved in a matter of minutes by using AI and machine learning models, which ease difficult and complicated processes (Floridi 2021).

(2) what we can do - AI is giving the opportunity for "smart agency" which utilized together with human expertise and intelligence can enhance human capabilities that otherwise likely wouldn't be reached. If society approaches the development of AI in a responsible manner, the possibilities for human agency can lead to greater societal benefits such as the improvement of efficiency, better decision making and other benefits which would contribute to a better society (Floridi 2021).

(3) what we can achieve - Artificial intelligence can offer society vast of opportunities for improving the achievements both of individuals and society. AI can be used in human rights protection, medical areas, logistics, solving complex problems and all other areas which can enhance what humans are already capable of achieving. Human intelligence and capability by being enhanced by AI opportunities can therefore contribute to societies' problem solutions (Floridi 2021).

(4) how we can interact with each other and the world - In tackling global issues like climate change, antimicrobial resistance, nuclear threats and similar, a joint approach in finding a solution to these issues needs to exist. This can only be achieved with the involvement of all international actors who can contribute to the regulation of AI and its use for good in the international arena. While AI can be used as an opportunity to manage complex coordination can enhance societal cohesion, there is also the risk of AI misuse which needs to be addressed by international actors. It's crucial to use the benefits and control the risks that AI may impose in order for societies to thrive (Floridi 2021).

Floridi considers that society can approach to each of these opportunities in 3 different ways: "AI can be used to foster human nature and its potentialities, thus creating opportunities; underused, thus creating opportunity costs; or overused and misused, thus creating risks".

On the other hand, the same as AI has an impact on society, humans also impact AI. Since AI can be both beneficial and harming to human rights, it's important for it to be globally regulated by the guidance of human rights. The idea behind this is that by emphasizing the importance of respecting human rights in the creation of AI, more rights-sensitive AI will be developed (Panic 2024). The global initiative "AI for Good" is in the public interest since it emphasizes the use of AI to create better outcomes for society in general, both for individuals and societies (Panic 2024). As shown in the second chapter, international actors such as the European Union have already taken important steps to demand more transparent and accountable development of AI. One of the EU initiatives is also guidance on "future-proofing human rights in the age of AI", which calls for conducting human rights impact assessments before the implementation of AI solutions (Panic 2024). Good examples of this are also the UN Guiding Principles on Business and Human Rights which highlight the importance of businesses and other entities having a responsibility to respect human rights when conducting business (Panic 2024). Private actors and industry leaders started calling for global regulation of AI and demanding more governmental regulation. Amnesty International has argued that the EU AI Act is critical to "ending the use of discriminatory and rights-violating artificial intelligence (AI) systems." ("Rome Call for AI Ethics: A Global University Summit," 2022).

Today, human defenders, whether as an organization or individuals, thanks to technology have an enormous availability of real-time data that's easily accessible and mostly shared through social

media platforms. But there are two sides to this story. While in cases that do not involve serious harm and graphic images, this way of sharing information is highly beneficial and undisturbed, when violent conflicts are in question the situation gets complicated. Many social media companies have algorithms that automatically detect and block violent content, for the sake of not promoting violence and protecting users from disturbing images, but at the same time, this function can be a step back in the protection of human rights. It prevents real evidence about violence from becoming available to the public, which often could be a piece of crucial evidence but also important for forming public opinion on some situation or conflict.

Moreover, there is a significant challenge in the situations of using AI in the peacebuilding process, since that work is usually implemented in critical moments during the ongoing crisis (Panic 2024). This peacebuilding efforts often have to happen during the unstable period which requires urgent reaction. In situations like that, a challenge might appear on how to carefully apply the ethical principles to AI projects needed for peacebuilding, given the tension between the long ethical evaluation and the urgent need for immediate action (Panic 2024) Of course, this doesn't implicate that ethical norms should be disregarded in urgent situations, but rather, that there needs to be a system for embedding ethics with urgency, which is crucial for the "promise of AI for peace" (Panic 2024).

AI governance has proven to be quite paradoxical. It represents a novelty in the world of technology but also a novelty in the area of protection of human rights, policy and ethics. It promises transformative benefits in the fields of medicine, transport, data processing and collecting, but also creates significant challenges and dilemmas in terms of privacy, bias, accountability and transparency. AI empowers human rights defenders with access to tools that provide more representative and accurate results, but it's becoming a force from which humans need to be protected. A regulatory lag that exists because of the fast and unpredictive development of AI creates additional pressure on the international actors to regulate it, but it also takes more time for human rights defenders to find a way to use those developments for good, and find a perfect balance between innovation and regulation has proven to be quite challenging. As innovations continue and AI influences our lives more and more, in order to find a balance between using the full potential of AI and addressing its risks the governance of this technology

will remain a critical area of focus for technological experts, policymakers, researchers, private actors and industry leaders.

In order to promote the use of AI for peace, it's important that globally accepted ethical principles get translated into practical methods. This is especially important with AI since it's not only in the hands of governments or big organisations, AI is available to everyone so its timely regulation aligned with AI advancements is crucial. Therefore, ethical principles need to be the first step in this demanding process, and as it was shown in the second chapter, by ongoing international efforts.

Conclusion

With the rapid and unpredictable development of AI technologies, it's easy to be pessimistic about technological evolution and focus on the possible threats AI can impose on human rights. This chapter aimed to shift the focus on the examples where AI can contribute to the protection of human rights and show how these technologies can be employed in a different way that could serve humanity by protecting and advancing human rights, such as in human rights monitoring, violent conflict prediction, early warning systems and climate change. Even though AI requires a lot of effort from international actors in regulating it ethically, it also provides numerous opportunities for its use for good. Ideally, actors that should focus on the potential of AI in the protection of human rights are civil society organisations, states, international and regional organisations and similar. By a collective approach to searching for the positive ways AI can be implemented to improve human rights protection, huge advancements can be made. By cooperation of human rights experts with technological experts in the field of AI, many initiatives could be created or current initiatives could be improved.

Predicting conflicts, especially in peaceful countries, is difficult, the monitoring and reporting processes are still not perfect, and early warning systems heavily depend on the reporting of locals. The AI models that are currently in use in these situations are still not perfected and the developers and users still encounter many challenges. Still, all these mechanisms should continue with their improvement in the use of AI since the results so far have shown many benefits, especially in giving faster and more precise analysis of big data sets. Even with the models that are currently used, there are a lot of issues regarding the quality and quantity of data since

precise prediction and analysis require data that are reliable and representable (Panic 2024). One of the ways researchers and organisations are approaching this field now is the use of smaller data sets instead of big ones (which could "overwhelm" the machine), which creates additional troubles because small data sources can lead to underrepresentation or biases (Panic 2024).

These are just some of the initiatives in which the combination of AI and human rights protection and monitoring was combined. Most of them presented pivotal tries in the collaboration of these two, at first look, distinct fields, but nevertheless, they show great potential in this space. In order to emphasize this potential, the mentioned case studies delve much deeper and also provide the challenges that might have occurred during those projects.

One of the biggest challenges is achieving a positive attitude toward AI which requires a lot of effort from international actors in encouraging the public's trust in AI and its possibilities for a positive impact. Over the last couple of years, AI has become part of our lives but also an important tool for achieving goals in protecting human rights and maintaining peace.

The positive examples of the interaction between human rights protection through international humanitarian law and artificial intelligence have become more and more often since many civil society organizations have started to cooperate with tech organisations and delve into the scope of possible opportunities AI can provide in reaching their goal. Therefore, this chapter aimed to provide a positive example of the implementation of AI in the protection of human rights by different international actors.

AI has already become a crucial part of the human rights protection mechanisms. Thermal imaging, use of satellites, use of machine learning systems are just some of the ways that AI is contributing to the work of human rights defenders. Since AI innovations happen almost on a daily level, their use will become even greater in the future. Since these technologies can produce both benefits and harm to human rights, it's important that legal and regulatory regulations stand in place as soon as possible.

Even though the benefits of using AI in human rights protection mechanisms has proven as highly beneficial, the use of AI in human rights protection still has its limits and challenges, one of which refers to the limited acceptance of AI-assisted evidence in judicial proceedings (Panic 2024). This example shows the need for further development in the legal framework and

standards, which should be done at the international level in order to ensure the reliable and ethical use of the evidence gathered thanks to AI. What makes this limitation even deeper is the recent emergence of deepfake content, which are discussed in the first chapter on the AI challenges, and represents one of the ways in which AI can be misused, especially since there is always a question of technological development and whether we will even be able to have a "counter-technology" which would have the ability to detect AI altered content.

The second limitation is the technical expertise required for the AI systems to be implemented, trained and used within the human rights organization, this is especially troubling for local human rights organisations (Panic 2024).

In order to fully harness the potential of AI for good, the accent is again put on the global governance of AI. International actors need to ensure that through education, training programmes, conferences, global actions, reporting and international collaborations and partnerships provide information on how AI can be used for good and what can be done in order to achieve that. It's important to bridge the gap between technology and its benefits for human rights, that can be done by establishing a global network of AI and human rights experts who would share good practices that are developed and deployed among human rights defenders on a strong ethical foundation. This will not only promote AI for peace initiatives but will also shed light on how to avoid and protect from AI misuse.

Regardless of all the challenges and skepticism around the use of AI for good, there is still potential for the creation of a positive impact on human rights with the help of technology. Even if we can't know with certainty whether will AI ever be helpful in the creation of the perfect human rights protection mechanisms, the fact that AI already contributes to that goal should be a reason enough for the international actors to keep putting efforts in this field. One is sure, that human oversight, expertise and involvement in these mechanisms are still greatly needed, and maybe only by focusing on how we can utilize AI to add value to human work, and less on how AI will "steal our jobs" we can contribute to a greater cause.

CONCLUSION

Regulation of Artificial Intelligence has proven to be an urgent yet demanding mission in the international arena. In recent years, the growing global interest from countries, individuals, organizations, private companies, and businesses in the possibilities and risks of AI has only confirmed how important this technological evolution is for humanity. Policy and ethical discussions have been flooded with possible outcomes with the involvement of AI in their respective fields, spawning numerous different theories. While AI innovations present novel benefits and opportunities, fears about the dangers and its potential misuse are also a part of everyday discussions, often taking precedence. These debates did not bypass international relations, where the impact of AI on global security, economy, and ethical standards are thoroughly discussed. Based on the scope of this new technology, international actors are dealing with complex challenges while trying to define a regulatory framework that simultaneously fosters AI innovation and promotes human rights protection.

To answer the main research question: *"How are international actors balancing AI regulation with its development while upholding fundamental human rights?",* firstly, the research focused on defining AI and examining how AI technology directly impacts human rights. This was done by exploring the effect of this technology on multiple principles such as transparency, oversight, accountability and responsibility, privacy and data protection. This chapter focused on the challenges and risks that can be imposed on these principles, emphasized recommendations on how to ensure those principles and especially highlighted the importance of global cooperation in resolving possible threats to them.

Further, the focus of the research shifted to exploring strategies and responses of international actors to the emerging need for regulation. The analysis of the regulatory frameworks by major international actors was conducted, focusing on the gaps and main issues that emerged from these debates and examples of regulating AI from a human rights perspective. Special attention was paid to international cooperation efforts and their obstacles.

The study then continued to analyze how AI can be used for good. This was done by demonstrating the areas in which AI can be applied practically, such as human rights monitoring

and reporting, forecasting international displacement, conflict prediction, climate change and AI, through contemporary examples of this practice.

1. Overview of main findings

Unpacking of different dimensions of the governance of AI in international relations in this study has provided a comprehensive answer to the main research question. The findings of this research highlight the importance of global governance of AI in ensuring proper regulation and protection of human rights and find that while the unified global approach to the regulation of AI still doesn't exist, there have been numerous important steps of the international community toward its development. Among many policies and regulations achieved by the EU and other international actors such as the United Nations, OECD and UNESCO, two examples that stand out are the European Union Artificial Intelligence Act, which is the first regulation on Artificial Intelligence and the Council's Framework Convention on artificial intelligence and human rights, democracy, and the rule of law, which represents the first legally binding international treaty on Artificial Intelligence.

The EU AI Act can certainly be taken as a great example of regional cooperation with a global impact. Many studies have highlighted the influence of this act on future regulations around the world, not only on the European ground, which draws inspiration from its provisions. Although often referred to as the "regulator rather than the inventor", the EU has done an enormous job in regulating AI which is beneficial for the protection of human rights and represents an example of regulation on a regional level combined with the protection of ethical and societal values and the emphasis on the principles analyzed in the first chapter.

The Council of Europe's Framework Convention on artificial intelligence and human rights, democracy, and the rule of law, is also representing a significant step towards global AI governance. The fact that the treaty can be signed by the Council of Europe member States, the European Union member States and non-member States that participated in its elaboration allows this treaty to have a direct global impact on the regulation of AI.

This study also acknowledges the important work done by major international human rights organisations, such as Amnesty International and Human Rights Watch, who showed their dedication to human rights by being courageous and curious enough to incorporate AI into their

work. They represent a good example and motivation for other human rights defenders to recognize the positive impact AI can have when implemented in the protection mechanisms of human rights. Their impact overall contributes to the efforts made by the other actors of the international community in showing the importance of the international approach towards AI regulation.

2. Possible areas of future research

Based on the generally pessimistic attitude towards the effect of AI on human rights, not a lot of attention is dedicated to the potential use of AI in the enhancement of human rights protection mechanisms. Ideally, a lot more research should be done in the area to explore how AI can be leveraged to strengthen human rights frameworks, and how that can be done by a collective approach of international actors. The examination of innovative applications of AI could further shed light on its positive impact and motivate the international community to consider and develop these opportunities further.

Further, the question of the misuse of AI remains unresolved, since technological advancements continue to evolve rapidly and are getting harder and harder to detect and regulate. This raises significant concern whether in the future we'll be able to control autonomous weapons systems and be completely sure of their ethical use. Additionally, there is a growing concern about whether we will have the tools necessary to differentiate deepfakes from real content, which is a threat that could lead to significant global tensions.

So far, the international community seems to have the motivation to govern AI on a global level. However, the question remains if the investors and technology companies will support this level of regulation or will rather opt to invest in locations with fewer restrictions, which could in the end show as undermining for human rights. On the other side, there is a possibility that investors and users will favor AI that guarantees their safety and adheres to strict ethical standards, therefore positioning the EU as a potential leader in this "AI technology race". Therefore, this dimension of the regulation of AI governance presents a valuable area of future research, particularly in determining whether the protection of human rights or the drive for innovation will ultimately prevail.

Bibliography

Ai, Hleg. 2019. “High-Level Expert Group on Artificial Intelligence.” *Ethics guidelines for trustworthy AI* 6.

Amnesty International. 2018a. *Troll Patrol Findings: Using Crowdsourcing, Data Science, & Machine Learning to Measure Violence and Abuse Against Women on Twitter*. Vol. 2018, Accessed June 5, 2024. https://perma.cc/76ND-GJLE.

Amnesty International 2018b. “Toxic Twitter—The Solution: Chapter 8.”, Accessed June 5, 2024 https://perma.cc/BZW2-9TWR.

“PricewaterhouseCoopers Consulting.” 2020. *Artificial Intelligence for Reporting*. https://perma.cc/99NP-P8VS.

“AI & Robotics Researchers.” 2015. *“Autonomous Weapons Open Letter.”*, Accessed June 5, 2024 https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/.

Barbosa, Lutiana Valadares Fernandes. 2023. “Exploring the 2023 U.S.” *Directive on Autonomy in Weapon Systems*.

Bommasani, Rishi, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. *‘The foundation model transparency index.’ arXiv Preprint ArXiv:2310.12941*.

Bowlsby, Drew, Erica Chenoweth, Cullen Hendrix, and Jonathan D. Moyer. 2020. “The Future Is a Moving Target: Predicting Political Instability.” *British Journal of Political Science* 50, no. 4: 1405–17. https://doi.org/10.1017/S0007123418000443.

Bussler, Frederik. 2020. “A History of Artificial Intelligence – From the Beginning.” *Towards Data Science*, Accessed May 30, 2024. https://towardsdatascience.com/a-history-of-artificial-intelligence-from-the-beginning-10be5b99c5f4.

Coeckelbergh, Mark. 2021. “AI for Climate: Freedom, Justice, and Other Ethical and Political Challenges.” *AI & Ethics* 1, no. 1: 67–72. https://doi.org/10.1007/s43681-020-00007-2.

Cornebise, Julien et al. 2018. “Witnessing Atrocities: Quantifying Villages Destruction in Darfur with Crowdsourcing and Transfer Learning.” In *Proc. AI for Soc. Good NeurIPS2018 Workshop*.

Council of Europe. Nov. 4 1950. European Convention on Human Rights: Article 8. Rome.

Council of Europe. Sept. 11, 2019. *Terms of Reference for the Ad Hoc Committee on Artificial Intelligence (CAHAI)*. Accessed May 10 2024.

Council of Europe, May 17, 2024. Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law.

Doshi-Velez, Finale, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott et al. 2017. *'Accountability of AI under the law: The role of explanation.' arXiv Preprint ArXiv:1711.01134*.

Dulka, A. 2022. "The Use of Artificial Intelligence in International Human Rights Law." *Stanford Technology Law Review* 26.

European Commission. 2018a. "Artificial Intelligence – A European Perspective." https://publications.jrc.ec.europa.eu/repository/bitstream/JRC113826/aiflagship-report-online.pdf, Accessed 6/22/2024.

European Commission 2018b. "Communication from the Commission 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." Brussels: European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN, Accessed 6/22/2024.

European Commission 2021a. "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts." *Com* 206 final, Accessed April 21, 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

European Commission 2021b. "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts." *Com* 206 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

European Commission, Mar. 26, 2024. “Commission Publishes Guidelines Under the DSA for the Mitigation of Systemic Risks Online for Elections.” Brussels, Accessed June 6, 2024. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1707.

European Parliament. 2017. *Resolution of 14 March 2017 on Fundamental Rights Implications of Big Data: Privacy, Data Protection, Nondiscrimination, Security and Law Enforcement (2016/2225(INI))*.

European Parliament Committee on Civil Liberties. Nov. 23, 2016. “Justice and Home Affairs”. "Opinion of the Committee on Civil Liberties, Justice and Home Affairs for the Committee on Legal Affairs with Recommendations to the Commission on Civil Law Rules on Robotics." 2015, no. 2103(INL).

European Union. 2016: 1–88. “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing” Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*: L119.

Europol. 2022. *Facing Reality? Law Enforcement and the Challenge of Deepfakes: An Observatory Report from the Europol Innovation Lab. Luxembourg*. Publications Office of the European Union.

Floridi, Luciano, ed. 2021. *Ethics, Governance, and Policies in Artificial Intelligence*. Cham, Switzerland: Springer.

Franke, Ulrike. 2020. “France and Germany: Where Do They Agree on AI?.” *European Council on Foreign Relations*, edited by Carla Hobbs, Essay collection. https://ecfr.eu/publication/europe_digital_sovereignty_rulemaker_superpower_age_us_china_rivalry/#Artificial_intelligence:_Towards_a_pan-European_strategy.

Franke, Ulrike, and Paola Sartori. 2019. “Machine Politics: Europe and the AI Revolution.” *European Council on Foreign Relations*. https://ecfr.eu/publication/machine_politics_europe_and_the_ai_revolution/.

G20. Jun. 2019. “Ministerial Statement on Trade and Digital Economy.”, Accessed May 9 2024 https://g20trade-digital.go.jp/dl/Ministerial_Statement_on_Trade_and_Digital_Economy.pdf.

Goddard, Michelle. 2017. "The EU General Data Protection Regulation (GDPR): European Regulation That Has a Global Impact." *International Journal of Market Research* 59, no. 6: 703–5. https://doi.org/10.2501/IJMR-2017-050.

Hacker, Philipp. 2018. "Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law." *Common Market Law Review* 55, no. 4: 1143–85. https://doi.org/10.54648/COLA2018095.

Hegre, Håvard, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, Naima Mouhleb, Sayyed Auwn Muhammad, Desirée Nilsson, Håvard Mokleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Gerald Schneider, Nina von Uexkull, and Jonas Vestby. 2019. "ViEWS: A Political Violence Early-Warning System." *Journal of Peace Research* 56, no. 2: 155–74. https://doi.org/10.1177/0022343319823860.

High-Level Expert Group on Artificial Intelligence (HLEG). 2019a. *A Definition of AI: Main Capabilities and Disciplines*. ec.europa.eu.

Expert Group on Artifical Intelligence, 2019b. *Ethics Guidelines for Trustworthy AI*. European Commission. ec.europa.eu.

Hilton, Jacob, Daniel Kokotajlo, Ramana Kumar, Neel Nanda, William Saunders, Carroll Wainwright, and Daniel Ziegler. "'A Right to Warn about Advanced Artificial Intelligence.' Last Modified June 4, 2024." *Assessed June 10, 2024 https://righttowarn.ai/*.

Horowitz et al. Jul. 10, 2018. "Artificial Intelligence and International Security". https://perma.cc/AVP2-9XSF.

Human Rights Watch. 2014. "Shaking the Foundations: The Human Rights Implications of Killer Robots". https://www.hrw.org/report/2014/05/12/shaking-foundations/human-rights-implications-killer-robots.

Human Rights Watch, and Harvard Law School International Human Rights Clinic. 2012. *Losing Humanity: The Case Against Killer Robots*.

ICRC. 2014. "Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects." In Expert Meeting, Geneva, Switzerland, March 26–28 2014. Vol. 3.

Larsson, Stefan, and Fredrik Heintz. 2020. “Transparency in Artificial Intelligence.” *Internet Policy Review* 9, no. 2. https://doi.org/10.14763/2020.2.1469.

Lee, Kai-Fu. 2018. *AI-Superpowers, China, Silicon Valley and the New World Order*. Boston: Houghton Mifflin Harcourt.

Littman et al. 2021. “Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence” *Study Panel* Report 9. Vol. AI100.

Loveluck, Louisa. Aug. 18, 2018. Assesed. “The Secret App That Gives Syrian Civilians Minutes to Escape Airstrikes.” *The Washington Post*, Accessed on June 14 2024. https://www.washingtonpost.com/world/the-secret-app-that-gives-syrian-civilians-minutes-to-escape-airstrikes/2018/08/17/e91e66be-9cbf-11e8-b55e-.

Maras, Marie-Helen, and Alex Alexandrou. 2019. “Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos.” *International Journal of Evidence & Proof* 23, no. 3: 255–62. https://doi.org/10.1177/1365712718807226.

Marin, Milena et al. Jul. 6, 2020. https://perma.cc/5YMV-MFJP. “Using Artificial Intelligence to Scale Up Human Rights Research: A Case Study on Darfur, Citizen EVIDENCE.” *La.B.*

Mazzini, G. 2019a. “A System of Governance for Artificial Intelligence Through the Lens of Emerging Intersections Between AI and EU Law.” In *Digital Revolutions – New Challenges for Law*. Vol. 1, edited by A. de Franceschi, and R. Schulze: 3–4. Munich: C.H. Beck.

Mazzini, G. 2019b. “A System of Governance.” In *Digital Revolutions – New Challenges for Law*. Vol. 4, edited by A. de Franceschi, and R. Schulze. Munich: C.H. Beck.

Meta. “Misinformation.” *Meta Transparency*, Accessed June 7, 2024. https://transparency.meta.com/en-gb/policies/community-standards/misinformation.

Mügge, Daniel. 2024. “Eu AI Sovereignty: For Whom, to What End, and to Whose Benefit?.” *Journal of European Public Policy*: 1–26. https://doi.org/10.1080/13501763.2024.2318475.

Novelli, Claudio, Mariarosaria Taddeo, and Luciano Floridi. 2023. “Accountability in Artificial Intelligence: What It Is and How It Works.” *AI & Society*: 1–12. https://doi.org/10.1007/s00146-023-01635-y.

OECD. 2019b. “OECD Principles on Artificial Intelligence.” OECD, Accessed June 22, 2024. https://www.oecd.org/going-digital/ai/principles/.

Organization for Economic Co-operation and Development (OECD). 2019. “Recommendation of the Council on Artificial Intelligence.” OECD/LEGAL/0449, May 22. OECD: 2019a. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

“Troll Patrol Findings: Using Crowdsourcing, Data Science, & Machine Learning to Measure Violence and Abuse Against Women on Twitter, Amnesty Int’ L”. https://perma.cc/76ND-GJLE.

Politico. 2021a. “AI: Decoded: UNESCO’s AI Ethics Framework—(Some) EU Countries Want AI Bans for Tech Companies—UK Benefits Algorithm Under Fire.”, Accessed November 24, 2021 Accessed: June 5, 2024 https://www.politico.eu/newsletter/ai-decoded/unescos-ai-ethics-framework-some-eu-countries-want-ai-bans-for-tech-companies-uk-benefits-algorithm-under-fire-2/.

Politico. 2021b. “China Backs UN Pledge to Ban (Its Own) Social Scoring.”, Accessed November 23, 2021 Accessed: June 5, 2024 https://www.politico.eu/article/china-artificial-intelligence-ai-ban-social-scoring-united-nations-unesco-ethical-ai/.

Politico. Dec. 12, 2023. *“What’s Actually Happening with Europe’s Big AI Law,”* Digital Future Daily, Accessed June 5, 2024. https://www.politico.com/newsletters/digital-future-daily/2023/12/12/whats-actually-happening-with-europes-big-ai-law-00131353.

“Politico”. "EU Political Parties Promise to Steer Clear of Deepfakes Ahead of Election." 2024, Accessed June 7, 2024. https://www.politico.eu/article/eu-political-parties-promise-to-steer-clear-of-deepfakes-ahead-of-election/. *Politico*.

Politico. 2024a. “EU Countries Strike Deal on AI Law,”, Accessed February 2, 2024. Accessed: June 5, 2024 https://www.politico.eu/article/eu-countries-strike-deal-ai-law-act-technology/.

Politico. 2024b. “International AI Rights Treaty Hangs by a Thread.”, Accessed March 11, 2024 Accessed: June 5, 2024 https://www.politico.eu/article/council-europe-make-mockery-international-ai-rights-treaty/.

Reeves, Shane, Ronald Alcala, and Amy McCarthy. Dec. 1, 2020. “Challenges in Regulating Lethal Autonomous Weapons Under International Law.” *Southwestern Journal of International Law* 28: 101–18.

Reuters. Dec. 7, 2023. “EU Still Hammering Out Landmark AI Rules After Marathon Overnight Talks.” *Reuters*, Accessed June 5, 2024. https://www.reuters.com/technology/eu-still-hammering-out-landmark-ai-rules-marathon-overnight-talks-2023-12-07/.

Reuters. Apr. 5, 2024. “Meta Overhauls Rules for Deepfakes, Other Altered Media.”, Accessed June 6, 2024 https://www.reuters.com/technology/cybersecurity/meta-overhauls-rules-deepfakes-other-altered-media-2024-04-05/.

Rød, Espen Geelmuyden, Tim Gåsste, and Håvard Hegre. 2024. “A Review and Comparison of Conflict Early Warning Systems.” *International Journal of Forecasting* 40, no. 1: 96–112. https://doi.org/10.1016/j.ijforecast.2023.01.001.

Rodrigues, Rowena. 2020. “Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities.” *Journal of Responsible Technology* 4: 100005. https://doi.org/10.1016/j.jrt.2020.100005.

Russell, Stuart J., and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. Pearson.

Salian, Isha. Apr. 4, 2019. “AI in the Sky Aids Feet on the Ground Spotting Human Rights, Nvidia”. https://perma.cc/J8TE-BH27.

Sgueo, Gianluca. Feb. 26, 2024. “The AI Act vs Deepfakes: A Step Forward, but Is It Enough?.” *Euractiv*, Accessed June 6, 2024. https://www.euractiv.com/section/artificial-intelligence/opinion/the-ai-act-vs-deepfakes-a-step-forward-but-is-it-enough/.

Smuha, Nathalie A. 2019. “The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence.” *Computer Law Review International* 20, no. 4: 97–106. https://doi.org/10.9785/cri-2019-200402.

Smuha, Nathalie, A. 2021. “From a ‘Race to AI’ to a ‘Race to AI Regulation’: Regulatory Competition for Artificial Intelligence.” *Law, Innovation & Technology* 13, no. 1: 57–84. https://doi.org/10.1080/17579961.2021.1898300.

European Parliament Research Service. 2021. *"Tackling Deepfakes in European Policy"*, Accessed June 5 2024. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.

"Rome Call for AI Ethics: A Global University Summit." 2022. *Tech. Ethics Lab*, Accessed June 1, 2024. https://techethicslab.nd.edu/news-and-events/rome-call-for-ai-ethics-a-global-university-summit/.

Ufert, Fabienne. 2020. "AI Regulation Through the Lens of Fundamental Rights: How Well Does the GDPR Address the Challenges Posed by AI?." *European Papers- A Journal on Law & Integration* 2: 1087–97.

UN. Mar. 15, 2024. *News*. "UN Forum Spotlights Need to Harness Potential of Artificial Intelligence for Good." United Nations, Accessed June 2, 2024. https://news.un.org/en/story/2024/03/1147831.

UN news. 2024 " Forum Spotlights Need to Harness Potential of Artificial Intelligence for Good." UN News, March 12, 2024." *News*, Accessed June 11, 2024. https://news.un.org/en/story/2024/03/1147831.

UNESCO. 2021. "UNESCO Recommendation on the Ethics of Artificial Intelligence." United Nations Educational, Scientific and Cultural Organization, Accessed June 22, 2024. https://unesdoc.unesco.org/ark:/48223/pf0000376789.

United Nations. Dec. 2023a. "Advisory Body on Artificial Intelligence." Interim Report. *Governing AI for Humanity*.

United Nations. May 2023b. *Common Agenda Policy Brief 5-A Global Digital Compact—An Open, Free and Secure Digital Future for All*.

United Nations and World Bank. 2018. *Pathways for Peace: Inclusive Processes to Preventing Violent Conflict*. UN–World Bank Group.

United Nations General Assembly. December 2023. *Resolution* 78/241.

US National Institute for Standards and Technology. Jan. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*.

Van den Hoven van Genderen, Robert. 2017. "Privacy and Data Protection in the Age of Pervasive Technologies in AI and Robotics." *European Data Protection Law Review* 3, no. 3: 338–52. https://doi.org/10.21552/edpl/2017/3/8.

Veale, Michael, and Frederik Zuiderveen Borgesius. 2021. "Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach." *Computer Law Review International* 22, no. 4: 97–112. https://doi.org/10.9785/cri-2021-220402.

Von der Leyen, U. 2020. "A Union That Strives for More: My Agenda for Europe. Political Guidelines for the Next European Commission 2019–2024." In *European Commission. URL*.

Wrigley, Sam. 2018a. "Taming Artificial Intelligence: 'Bots,' the GDPR, and Regulatory Approaches." In *Robotics, AI and the Future of Law*. Vol. 187, edited by M. Corrales, M. Fenwick, and N. Forgó: 183–208. Berlin: Springer. https://doi.org/10.1007/978-981-13-2874-9_8