UNIVERSIDADE DE BRASÍLIA

INSTITUTO DE RELAÇÕES INTERNACIONAIS

CAMILA ARAUJO TANÚS GALVÃO

OECD'S AND UNESCO'S RECOMMENDATIONS ON ARTIFICIAL INTELLIGENCE IN THE 2020s:

Exploring Approaches in the Context of Multilateralism and the World Order

BRASÍLIA
2023

UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE RELAÇÕES INTERNACIONAIS

OECD'S AND UNESCO'S RECOMMENDATIONS ON ARTIFICIAL INTELLIGENCE IN THE 2020s:

Exploring Approaches in the Context of Multilateralism and World Order

CAMILA ARAUJO TANÚS GALVÃO

Final Term Paper
International Relations Institute of the University of Brasilia
Supervisor: Prof. Dr. Rodrigo Pires de Campos

BRASÍLIA
2023

ACKNOWLEDGEMENTS

Gostaria de expressar minha sincera gratidão ao meu orientador, Professor Rodrigo Pires de Campos, cuja dedicação e acolhimento foram essenciais neste processo. Seu conhecimento, incentivo e apoio foram pilares fundamentais para o desenvolvimento deste trabalho. Agradeço imensamente pelas contribuições, que ajudaram a dissipar as dúvidas que surgiram ao longo do caminho.

Além disso, é essencial registrar um agradecimento especial à minha mãe, Marisa, à minha irmã, Stephanie, e à minha madrinha e tia, Maria Eugênia. Mãe, como sempre enfatizo, todas as minhas conquistas foram e serão possíveis graças ao seu incansável esforço, à sua determinação e à sua garra ao longo da vida. Agradeço por todos os preciosos ensinamentos, especialmente sobre perseverança, e por sempre acreditar em mim. Stephanie, minha gratidão pelo carinho e por sempre celebrar minhas vitórias. Saiba que, apesar da nossa diferença de idade, você influenciou muitos gostos que cresceram junto comigo, tornando-se partes importantes da minha essência. Tia Eugênia, agradeço por ser constante luz e alegria. Sua presença, como uma segunda mãe, é não só uma fonte de inspiração de força, mas também de encarar a vida com leveza.

Aos meus amigos e entes queridos, tanto aqueles que acompanho desde outras fases da vida quanto os que compartilharam a jornada universitária comigo, expresso minha profunda gratidão por me proporcionarem felicidade, conexão, sinceridade e suporte. Independentemente do tempo em que estiveram presentes, saibam que cada um marcou a minha vida. Um agradecimento especial àqueles que contribuiram com este trabalho, seja por meio de sugestões, apoio emocional ou simplesmente por ouvir meus desabafos: Gabriel Gangana, Letícia Costa, Gabriela Liotto, Laura Bartolozzi e Enzo Ramos.

Ao Instituto de Relações Internacionais e à Universidade de Brasília, onde passei os últimos cinco anos, obrigada por me acolherem e por serem fundamentais no meu desenvolvimento profissional e pessoal. Chegar aqui foi um sonho por muitos anos, e é com alegria que afirmo tê-lo realizado. Agora, ao concluir esta trajetória, reconheço que não foi isenta de desafios, mas permito-me afirmar que os superei, aproveitando ao máximo tudo que a universidade tinha a oferecer, desde projetos e disciplinas até a experiência enriquecedora de um intercâmbio no Japão.

Por último, mas não menos importante, reservo um espaço para agradecer a mim mesma, não apenas pelo empenho dedicado a este trabalho, mas a tudo a que me comprometo. Obrigada, Camila, por nunca desistir, mesmo quando tudo parece difícil.

ABSTRACT

This study analyzes the proposals of global approaches to Artificial Intelligence (AI) advanced by the Organisation for Economic Co-operation and Development (OECD) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) in the beginning of the 2020s. The research is framed within the context of the neoliberal capitalist world order and new multilateralism, aiming to provide an analysis of the similarities and distinctions between the AI regime proposals advanced by each Organization within the current context of new multilateralism and transition in the world order. The study seeks to observe how these two international actors address AI in a geopolitical context where multilateralism faces a lack of legitimacy and the world order faces a potential hegemonic transition. The central hypothesis of this paper is that each Organization addresses AI in line with its backgrounds, objectives, and fundamental purposes, adjusting to the larger context: the OECD addresses it through a market-oriented stance, while UNESCO focuses on a socially oriented perspective. This study utilizes a historical approach and incorporates the concept of multilateralism as articulated by Robert Keohane, John Gerard Ruggie, and Robert Cox; analyzes data from literature reviews and official documents; and employs inductive coding for examination and interpretation of each document. Based on these elements, the research concludes that, while the OECD and UNESCO share similar overall objectives, their perspectives and approaches diverge significantly. The OECD, with a focus on the market economy, places considerable emphasis on fostering collaboration between the private sector and governments to achieve sustainable economic growth. Meanwhile, UNESCO, more aligned with social concerns, highlights the ethical use of AI for broader benefits, reflecting its commitment to peace and security through intergovernmental cooperation. These differences not only shape the content of their documents but also influence their structural and discursive approaches, outlining how their institutional origins and objectives guide their strategies in addressing the global challenges of AI in a context of emerging multilateralism and hegemonic transition in the world order.

Keywords: Artificial Intelligence; OECD; UNESCO; Recommendations; Multilateralism; World Order.

RESUMO

Este estudo analisa as propostas de abordagens globais à Inteligência Artificial (IA) preconizadas pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE) e pela Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO) no início da segunda década do século XXI. Esta investigação enquadra-se no contexto da ordem mundial capitalista neoliberal e do novo multilateralismo e pretende fornecer uma análise das semelhanças e distinções entre as propostas de regimes de IA apresentadas pelas respectivas Organizações no contexto atual de novo multilateralismo e de transição da ordem mundial. Desse modo, o estudo visa observar como esses dois atores internacionais estão lidando com as implicações geopolíticas da IA num contexto em que o multilateralismo enfrenta uma falta de legitimidade e a ordem mundial passa por uma potencial transição. Assume-se como hipótese de trabalho que cada Organização aborda a IA alinhada com seus antecedentes, objetivos e propósitos fundamentais: a OCDE com uma orientação para o mercado e a UNESCO com um enfoque social. Utiliza-se uma abordagem histórica e o conceito de multilateralismo por Robert Keohane, John Gerard Ruggie e Robert Cox; analisam-se dados de revisão de literatura e de documentos oficiais; e aplica-se uma codificação indutiva. A partir desses elementos, a pesquisa conclui que, embora a OCDE e a UNESCO partilhem objetivos gerais semelhantes, as suas perspectivas e abordagens divergem consideravelmente. A OCDE, focada na economia de mercado, destaca a colaboração entre o setor privado e os governos para o crescimento econômico sustentável. Enquanto a UNESCO, mais alinhada com as preocupações sociais, enfatiza o uso ético da IA para amplos benefícios, o que reflete seu compromisso com a paz e a segurança por meio da cooperação intergovernamental. Essas diferenças não apenas moldam o conteúdo de seus documentos, mas também influenciam suas abordagens estruturais e discursivas, evidenciando como suas origens e objetivos institucionais direcionam suas estratégias na abordagem dos desafios globais da IA em um contexto de novo multilateralismo emergente e de transição hegemônica da ordem mundial.

Palavras-chave: Inteligência Artificial; OCDE; UNESCO; Recomendações; Multilateralismo; Ordem Mundial.

Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks.

Stephen Hawking

SUMMARY

LIST OF ABBREVIATIONS AND ACRONYMS

AGI	Artificial General Intelligence
AI	Artificial Intelligence
ALPAC	Automatic Language Processing Advisory Committee
ANI	Artificial Narrow Intelligence
ANN	Artificial Neural Networks
ASI	Artificial Super Intelligence
BAT	Baidu, Alibaba, and Tencent
DAC	Development Assistance Committee
DAG	Development Assistance Group
DL	Deep Learning
EU	European Union
GAFAM	Google, Amazon, Facebook, Apple, and Microsoft
GATT	General Agreement of Tariffs and Trade
IBM	International Business Machines
ICIC	International Committee of Intellectual Cooperation
IO(s)	International Organization(s)
IR	International Relations
ICSU	International Council for Science
ML	Machine Learning
NLP	Natural Language Processing
NSCAI	National Security Commission on Artificial Intelligence
OECD	Organisation for Economic Co-operation and Development
OEEC	Organization for European Economic Cooperation
R&D	Research and Development
UK	United Kingdom
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization
US	United States
USSR	Union of Soviet Socialist Republics
WHO	World Health Organization
WTO	World Trade Organization

1. INTRODUCTION

The use of Artificial Intelligence (AI) has advanced in less than a century in unimaginable ways, becoming essential due to its analytical, manipulable, evolutionary, and predictive capabilities. This technology has become an integral part of individuals' daily lives owing to its pervasive use across diverse sectors of society. Consequently, it is now present in cultural artifacts, media platforms, academic and research institutes, healthcare systems, civil-society organizations, private companies, and governmental administration. This widespread of AI was aligned to the recognition of the improvements and advances it would bring to humanity, resulting in substantial public-private investments. Nonetheless, it soon became clear that such advantages would be counterbalanced by pitfalls and problems stemming from the geopolitical implications of AI. These challenges underscore the complexity of managing this innovation, which raises questions about how the international community is attempting to handle it amidst a crisis in multilateralism. In order to understand this crisis, it is crucial to present the origins of multilateral cooperation's institutionalization and the changes brought by the new liberal world order, known as neoliberalism.

After the Second World War, the United States (US) consolidated itself as a *hegemon* and "international organizations emerged as potential mediators and as pillars of the new order" (Lima; Albuquerque, 2021, p. 9). In theory, these organizations would bring a greater democracy to decision-making due to its multilateral character and wide agenda scope, from human rights to development, commerce and health issues (Lima; Albuquerque, 2021). However, this potential did not materialize in the years following the Cold War. The bipolar order presented great impasses for international organizations (IOs), and its dissolution was expected to enable the effective implementation of multilateral principles. Nevertheless, this expectation was not realized due to the gradual revival of neoliberal fundamentalism, which began and solidified afterward. This new world order led to a series of economic crises, the rise of new actors – state, private and cross-border –, and the adoption of new political-ideological strategies – strengthening ultra-right governments (Almeida; Campos, 2020, p. 20). Consequently, the dynamics of multilateralism were directly impacted by principles of individualism, reduced state involvement, and the unilateral strategies pursued by the *hegemon*.

The historical context mentioned above illustrates the origins of the lost of legitimacy of multilateral institutions. These roots established during and specially after the Cold War are the basis for unfolding events of the 21st century that serve as diagnoses for the multilateralism crisis. According to a recent article, the struggles of multilateral IOs in dealing with the competitive dynamics between the US and China – reflecting the Cold War tensions between the US and the Union of Soviet Socialist Republics (USSR) – and the World Health Organization (WHO) inability to effectively manage the global Covid-19 pandemic further demonstrates the challenges faced by these institutions (Lima; Albuquerque, 2021). This study highlights the growing academic interest in the topic, prompting readers to contemplate additional factors that could further underscore the crisis in multilateralism, such as the emergence and implications of AI.

The still limited information on what ethical principles should guide AI's design, development and deployment is also raising academic attention. A recent article on the topic reveals that "no previous study has systematically assessed a global consensus on ethics for AI in education" (Nguyen *et al.*, 2022, p. 4222). In this research, by examining and matching six ethical guidelines and reports of diverse entities and conducting a thematic analysis, the authors sought to prescribe a set of unified ethical principles, given that this could meet the demands of a widespread digitalization of education (Nguyen *et al.*, 2022). This study serves as an inspiration, delving into the role of ethics in guaranteeing the quality of delivery. However, it does not encompass an analysis of the differences between the organizations that it references, missing out on the nuances that could demonstrate which approaches could fit under the current neoliberal global order.

Therefore, considering the recent problems outlined by both topics, this study recognizes the imperative need to explore the integration of AI within the International Relations (IR) debate and its geopolitical implications as another facet highlighting the crisis in multilateralism. The premise behind this perspective stems from the perception that AI can only be handled collectively with all stakeholders involved, and not only relying on intergovernmental efforts.

Due to the rapid advancement of AI and its application in diverse sectors, the international community was caught unprepared for the urgent necessity of formulating standard definitions, values and policies to guide and shape the use of this technology. In this context, two IOs have already issued recommendations concerning the use of this innovation. The Organisation for Economic Co-operation and Development (OECD) published the *Recommendation of the Council on Artificial Intelligence* in 2019, followed by the United Nations Educational, Scientific and Cultural Organization (UNESCO)'s *Recommendation on the Ethics of Artificial Intelligence* in 2021.

This study endeavors to answer the following research question: *In light of the recommendations proposed by each International Organization, how is Artificial Intelligence being addressed by the OECD and UNESCO in the 2020s?* This inquiry is framed within the context of the neoliberal capitalist world order and new emerging multilateralism, and intends to provide, within that context, an examination of the similarities and distinctions between the two aforementioned documents.

Therefore, this study aims to observe how these two international actors are attempting to handle AI's geopolitical implications in a context where multilateralism is facing a lack of legitimacy and the world order is facing a potential transition. Some secondary goals that will assist in accomplishing the primary aim of this paper include understanding AI and the greater attention it is raising in the international arena, comprehending its geopolitical implications, and evaluating whether the nature, functioning, and historical background of OECD and UNESCO shape their respective approach to AI.

Amidst a declining legitimacy of multilateral IOs observed in the latter part of the 20th century, the underlying causes of this situation trace back to shifts in the capitalist world order during the post-Second World War and post-Cold War periods (Lima; Albuquerque, 2021; Almeida; Campos, 2020). These shifts have resulted in complexities concerning how IOs navigate the challenges posed by emerging powers and actors, along with the need to encompass a wider range of values that extend beyond the foundational liberal and Western principles. In this context, this study posits a hypothesis that each organization approaches AI in accordance with their background and foundational aims. Therefore, the OECD addresses it through a market-oriented stance, while UNESCO focuses on a socially oriented manner.

This paper's theoretical framework will be based on five key articles. Three foundational pieces – Robert O. Keohane's *Multilateralism: An Agenda for Research* (1990), John Gerard Ruggie's *Multilateralism: the Anatomy of an Institution* (1992), and Robert W. Cox's *Multilateralism and World Order* (1992) – lay the groundwork for comprehending crucial theoretical insights of multilateralism. Adding historical context and contemporary relevance, Maria Regina Soares de Lima and Marianna Albuquerque's *Instituições Multilaterais E Governança Global* (2021) examines the failure to implement multilateral principles within the United Nations (UN) System. In addition, it highlights how the Covid-19 pandemic underscored the crisis in multilateralism post-Second World War and post-Cold War. Moreover, Celia Almeida and Rodrigo Pires de Campos' work, *Multilateralismo, Ordem Mundial e Covid-19: Questões Atuais e Desafios Futuros para a OMS* (2020), narrows its focus to the World Health Organization's (WHO) management of the Covid-19 crisis. This article offers a detailed analysis of the challenges arising from transformations within the post-Cold War neoliberal global order.

The methodology for carrying out this research consists of documentary and bibliographical analysis. The instruments of methods for collecting information will be primary sources available on official OECD and UNESCO platforms and secondary sources, like articles, books, and researches delineated below.

Foremost, for introductory insights into AI, this research will be in accordance with the book *Artificial Intelligence: A Non-Technical Introduction* (2017), authored by Professor and PhD Tad Gonsalves, academic at the Department of Information and Scientific Communications at Sophia University, Tokyo, Japan. This work dissects the two terms in 'Artificial Intelligence' separately with the final aim of comprehending the whole concept.

Moreover, the geopolitical implications of AI will be based on the considerations of three articles on the topic. *Artificial Intelligence Diplomacy: Artificial Intelligence Governance as a New European Union External Policy Tool* (2021), produced by Ulrike Franke as a request of the European Parliament's special committee in AI in a Digital Age; *The Geopolitics of Artificial Intelligence: The return of Empires?* (2018), by Nicolas Miailhe; and *The Geopolitics of Artificial Intelligence* (2020), by Anastasia Kapetas.

Lastly, the historical formation and characteristics of both OECD and UNESCO will be examined with reference on the *Convention on the OECD* (1960); the article *OCDE: uma visão brasileira* (2000) by Denis Pinto; the *UNESCO Constitution* (1945) and the fourth chapter of the book *Ciência, política e relações internacionais: ensaios sobre Paulo Carneiro* (2004), by Aant Elzinga. This historical exploration is essential as it will provide context for understanding the origins and goals of these organizations.

These pieces of information will be analyzed in an explanatory manner, supported by an inductive analysis assisted by MAXQDA software. This method involves identifying patterns and frequencies of words and themes that will be later categorized by an open coding approach. Subsequently, these codes will be classified into wider meaning clusters. While this method could be categorized as either qualitative or quantitative, this paper leans toward a qualitative approach. Rather than quantifying the mentioning frequency of words and themes, the focus will lie on identifying similarities, if any, between OECD's and UNESCO's recommendations. This choice was deliberate, considering that the extensive structure of UNESCO's recommendations compared to OECD's could provide an unproportional and unfair comparison.

This dissertation emerges from an outsider perspective, which does not exempt it from potential bias and preconceptions. However, reflexivity will be an ongoing practice, both on a personal and epistemological level to mitigate any semblance of dogmatism and ensure the research’s reliability and validity.

In order to achieve the defined objective, the work is divided into five sections, in addition to this introduction and the final considerations. The first section will elucidate AI’s key foundational concepts and the historical progression, offering insights of its current developmental stage. The second section aims to combine AI and IR by initially examining the integration of technology within the field, which will be attempted from drawing insights from diverse data sources. Subsequently, it will address the ongoing geopolitical implications of AI. Furthermore, the third section will explore concepts and historical context potentially useful for attempting to interconnect multilateralism, the new world order and AI. Afterward, the fourth section will delve into the nature, operational mode and historical backgrounds of both the OECD and UNESCO, besides addressing their recommendations. Finally, the last section will present an effort to respond the main question under investigation and assess the initially posited hypotheses. Hopefully, the final considerations will be able to synthesize the findings, observations, and conclusion drawn from this research, while also acknowledging its limitations and proposing potential future steps for this study.

2. ARTIFICIAL INTELLIGENCE: CONCEPT AND HISTORICAL EVOLUTION

This section, divided into two subsections, will play a pivotal role in establishing a foundational understanding of AI in a non-technical approach. The first subsection will elucidate AI's key concepts, ensuring clarity on the study's core theme, accessible to all readers regardless of their background in the field. Subsequently, it will delve into the historical progression of AI, offering context to its evolution and insights into its current state of development. This comprehensive overview seeks to clarify AI's essence and trajectory, providing the groundwork for further exploration of its implications within the context of this study.

2.1. What is Artificial Intelligence?

The advancement of AI in recent years brought forth the need to discuss its far-reaching implications within IR. Nonetheless, it is necessary to comprehend what AI is before connecting these two topics. Therefore, the goal of this subsection is to introduce this concept using the book *Artificial Intelligence: A Non-Technical Introduction* (2017), by PhD and professor Tad Gonsalves. Due to this paper's focus on academic disciplines within the humanities, the selection of Tad Gonsalves' work was based on its non-technical and non-mathematical approach for students "with or without computer science background" (Gonsalves, 2017, p. xi). This book dissects the two elements that constitute the term Artificial Intelligence separately with the final goal of comprehending them together.

First, something 'artificial' implies a resemblance and/or, more importantly, has an analogous function to the corresponding real or natural object. Some examples presented by Tad Gonsalves demonstrate the possibility of artificial artifacts with only one of these characteristics or both. While artificial legs are physically and functionally akin to natural legs, artificial eyes are merely similar in appearance to its natural counterpart, given that, until now, they cannot perform the seeing function. Despite these differences, both objects are 'artificial'. To introduce a higher level of complexity to this matter, there are also functions that exist in reality but can be artificially executed. Ships and planes, for example, are, respectively, "far-fetched imitations of fish and birds" (Gonsalves, 2017, p. 2), but their function closely resembles those of these animals. Therefore, a ship artificially swims and a plane artificially flies, which is somewhat similar to AI as it will be clarified further on.

On the other hand, the term ‘intelligence’ carries several definition possibilities, especially when specifying an area of study such as biology or sociology. Given the clear objective of understanding existing AI models, Professor Gonsalves uses a simple definition of the Webster’s Dictionary: “Intelligence is the ability to learn and solve problems”. In this context, in the same way that living beings unable to adapt and do no more than “execute the hard-wired program in their brains or DNA” (Gonsalves, 2017, p. 6) are not qualified as “intelligent”, the same applies to computers that only capable of executing what is within their software program.

As a result, assembling the terms above described formulates the definition of AI as the science of making machines that think and behave like human beings and that can do things which, at least at present, can be done only by human beings (Gonsalves, 2017). As outlined before, the performance of a function artificially, as in the case of ships and planes, is in some way relatable to AI. This function similarity to something natural does not mean that humans will imagine ships sailing like a fish swimming or a plane flying flapping its wings. The same logic applies to the idea of machines being able to think intelligently, since it does not have a form or appearance. Machines do not have a brain in which electrochemical communication forms a pattern of activities, which makes it hard to imagine how the fast analysis of basic units of information programed in their core works. Despite this difficulty, the capacity of coming up with an action, behavior or answer in human language makes people identify this as “manipulating knowledge in an intelligent way to solve a given problem” (Gonsalves, 2017, p. 4). Therefore, while AI may not replicate natural intelligence in form or function, the resemblance of its achieved outcomes to uniquely human capabilities makes this association reasonable.

2.2. Historical Advancement of AI and its Current Stage

The study of AI started in 1956 with John McCarthy and an alpha-beta search (Russel; Norvig, 2020). Although there were earlier researches – such as in 1943 with the model of artificial neurons by Warren McCulloch and Walter Pitt, and the invention of the Turing Test[1] by Alan Turing[2] –, it was McCarthy who first coined the term. The pioneering

[1] One of the most important experiments for understanding when machines reached intelligence (Marr, 2021). “A computer passes the test if a human interrogator, after posing some written questions, cannot tell whether the written responses come from a person or from a computer” (Russel; Norvig, 2020, p. 32).

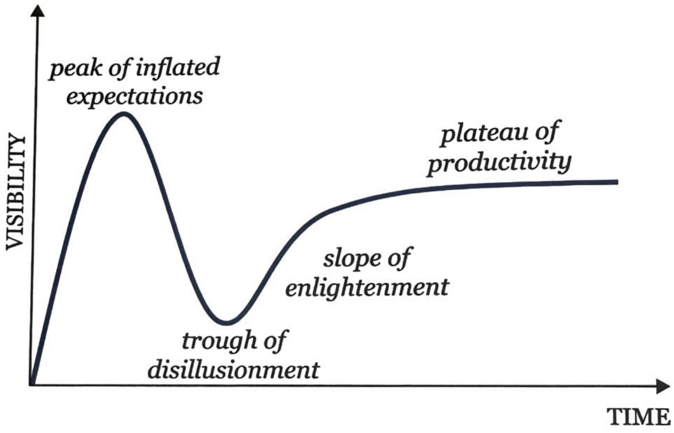
[2] British mathematician and logician who made major contributions to mathematics, cryptanalysis, logic, philosophy, and mathematical biology and also to the new areas later named computer science, cognitive science, artificial intelligence, and artificial life (Copeland, 2023).

use of this designation was in his proposal for a two-month workshop, during which ten man would investigate the potential for machines to simulate the learning process and exhibit intelligence. This investigation considered that these two features could be precisely described. Even without breakthroughs in this workshop, the term AI caught on for good.

Important milestones of AI after the term was first used started in the late 1950s until before the 1970s, with "the early programs for proving theorems and playing checkers" (Gonsalves, 2017, p. 9), the creation of Expert Systems (1965), the implementation of Natural Language Processing (NLP) in the chatbot ELIZA (1966), and the definition of the foundations of the field by Marvin Minsky (1969) and John McCarthy (1971). All of these achievements raised expectations and investments that were soon disappointed by poor results on machine translation experiments – reported in the Automatic Language Processing Advisory Committee (ALPAC) in the US in 1966 – and in unsuccessful automatic aircraft land system experiments in England in 1973 – stated in the Lighthill Report. These two reports highlighted deep problems within AI and defined it as both time-consuming and expensive. All of these events contributed to the transition from the early hyped period of AI, known as 'peak of inflated expectations', into a 'trough of disillusionment'. Consequently, funds were suspended, AI received negative portrayal in the public media, and AI research almost came to a grinding halt.

In an attempt to refrain from labeling studies as AI, some researchers continued their work, eventually guiding the innovation towards a 'slope of enlightenment' only in the 1990s. Some examples are the invention of the internet by Tim Berners-Lee (1991), and the defeat of world chess champion Garry Kasparov by International Business Machines Corporation's (IBM) *Deep Blue* supercomputer (1997). As a result, "by the end of the last century, AI had reached the 'plateau of productivity'" (Gonsalves, 2017, p. 10), with breakthroughs in Machine Learning (ML) and Game Playing. In 2011, IBM's supercomputer *Watson* won the Jeopardy US television quiz show against reigning champions Brad Rutter and Ken Jennings. Watson counted with combinations of NLP, semantic analysis, information retrieval, automated reasoning and ML to generate answers. Also, Google Deep Mind's supercomputer *AlphaGo* defeated Go's – popular game in China, Korea and Japan – world champion Lee Sedol in 2016 using, among other strategies, the Deep Learning (DL) method. An interesting fact about *AlphaGo* is that, during its Artificial Neural Networks (ANN) training, it was never once exposed to any games played by Lee Sedol, but only to other Go professionals and against itself.

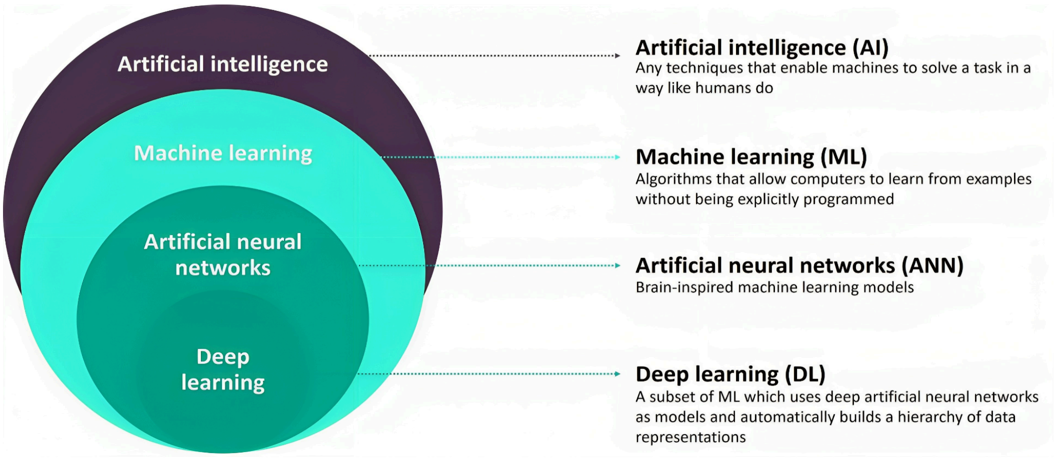
Figure 1: The hype cycle of AI



Reference: Gonsalves, 2017, p. 10.

Prior to proceeding, it is important to stress that ML, ANN and DL are parts of AI. Also, not all ML is DL (see fig. 2). An easy way to understand ML is thinking of it as pattern recognition. Given the infinite number of potential situations, creating rules to cover each one is nearly impossible. Therefore, when a machine learns general patterns to address problems, it falls under statistical ML. While there exists a second type of ML known as model-driven, which comprises approaches that build an abstract model of representation, this model has not yet achieved widespread success (OECD, 2023). However, for the purposes of this paper, the technical distinction between these types will not be relevant. Subsequently, DL is a subset method within ML that consists of employing layers comprising millions of ANN, which draw inspiration from the human brain learning process to solve complex problems. An ANN is "generally trained on a training dataset and then tested on an entirely different test dataset" (Gonsalves, 2017, p. 158) to learn patterns and make predictions from data.

Figure 2: Components of Artificial Intelligence


Artificial intelligence
Machine learning
Artificial neural networks
Deep learning
Artificial intelligence (AI)
Any techniques that enable machines to solve a task in a way like humans do
Machine learning (ML)
Algorithms that allow computers to learn from examples without being explicitly programmed
Artificial neural networks (ANN)
Brain-inspired machine learning models
Deep learning (DL)
A subset of ML which uses deep artificial neural networks as models and automatically builds a hierarchy of data representations

Reference: ForumIAS, 2022.

In summary, AI is a piece of software – commands that can always be changed, edited or erased, unlike hardware after it is designed – and an application program built to solve problems in an intelligent and autonomous way. In order to achieve the final creation of a program, it is necessary to use algorithms, which are a set os clear steps to solve a problem, written in natural languages. Subsequently, the establishment of algorithms allows for the creation of a pseudocode, that is, a combination of natural language and computing programming language (codes). Finally, with these in hand, a code system that makes up the program is designed. There are three different kinds of AI: weak/narrow, strong/general and extended/super, which will be explained below in order to present in what level humanity is currently in and what is trying to be achieved.

Artificial Narrow Intelligence (ANI), also known as weak AI or Expert System, refers to programs that perform extremely well in given and limited domains, but fail in others. In other words, "it cannot be extended to tasks for which it was not designed" (Gonsalves, 2017, p. 195). The three steps to create one are (i) acquire knowledge from a domain-specific experts through interviews, questionnaires, books, websites, etc; (ii) formalize it into knowledge representation through semantic networks[3] and frames[4] to remove ambiguities within the text and diagrams formats that the knowledge acquired is usually first structured; (iii) and make inferences[5], which is "drawing conclusions by matching the facts provided by the user to the antecedents of the rules in the knowledge base" (Gonsalves, 2017, p. 50).

Weak AI is already playing roles in many sectors of society, like financial services robo-advisors, telecommunication technologies, medical diagnoses systems, traffic control, legal and civil cases evaluation systems and crop damage forecast programs (Lutkevich, 2023). The famous ChatGPT is also categorized as narrow AI, given that it is only a NLP trained using DL, but was not yet designed to understand more complex tasks. Although these still weak systems provide several advantages and improvements for individuals, there are a series of ethical and responsibility matters that need to be established in case it fails or presents any risky mistake.

Proceeding to Artificial General Intelligence (AGI), or strong AI, which was not yet achieved, refers to the goal of creating machines with intellectual and spiritual capabilities of Homo Sapiens Sapiens, like possessing emotions, intentions, intuition, creativity and other

[3] "A graph of labeled nodes and labeled directed arcs to encode knowledge" (Gonsalves, 2017, p. 46).

[4] A frame contains slots that hold the attributes and fillers that hold the corresponding values of the attributes (Gonsalves, 2017).

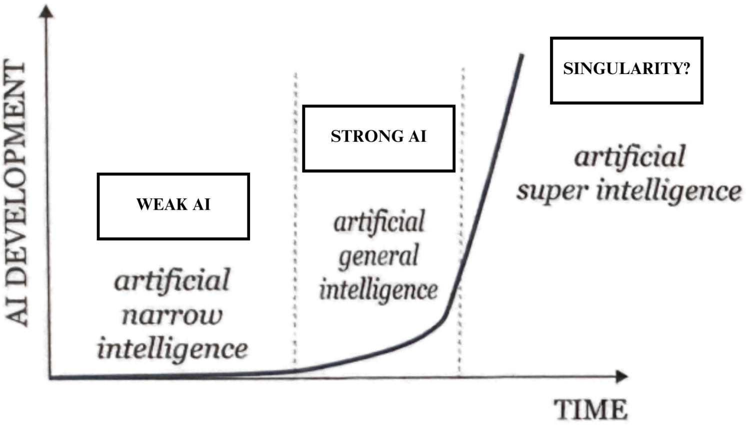
[5] Two possible engines for this last step are the Forward Chaining – based on facts (data) that connect with IF rules, tries to draw final conclusions on an IF-THEN rule basis – and Backward Chaining – in light of a hypothesis considered to be true, works backwards to understand the facts that led to that outcome.

uniquely human abilities. “They will know that they know [...] [and] will have consciousness [...]” (Gonsalves, 2017, p. 8).

Finally, a next step called Artificial Super Intelligence (ASI), or extended AI, is already idealized by some developers. This final advanced level refers to “humanoid robots that look like and behave like human beings” (Gonsalves, 2017, p. 9), including the robotics that people associate so rapidly to the term AI. Overcoming human intelligence, this type of innovation would be ubiquitous and omniscient in an autonomous and self-developing way. The point where AGI will self-improve into ASI is addressed as Singularity. This term is a reference to the name of the center of the black hole, where classical laws of physics cease to apply, representing a phenomenon that defies natural explanation.

Within the AI community, there is not a consensus regarding the endpoint of AI development. For conservatives, AGI will be the representation of the final goal of AI founding fathers, which is “building machines that think and behave like human beings” (Gonsalves, 2017, p. 197). In contrast, radical opinions stress that AGI will self-develop into ASI with or without human intervention. Regardless of these different visions, the AI community seeks to achieve AGI and believes it will evolve from ANI on par with human intelligence and control.

Figure 3: AI Growth Stages


AI DEVELOPMENT
WEAK AI
artificial narrow intelligence
STRONG AI
artificial general intelligence
SINGULARITY?
artificial super intelligence
TIME

Reference: Adapted from Gonsalves, 2017, p. 197.

However, experts are again divided when it comes to the takeoff of AGI to ASI, with two possible scenarios predicted by AI experts. While some believe that the growth from AGI to ASI will have a gradual and *soft takeoff* – taking decades or even centuries – with human beings in control of the situation, others fear that, after the subtle development of ANI to AGI, it will become too complex to be controlled by human beings. This fear revolves around the potential for AGI to suddenly transition to ASI through self-learning loops, characterizing a *hard takeoff*. Professor Tad Gonsalves added his own conjecture to these predictions, suggesting the potential for an *aborted takeoff*. This proposal addresses limitations not imposed by human constraints, but rather by nature's inherent limits on levels of intelligence. The author draws a parallel with the General Theory of Relativity, where lightspeed is the limit that no material body or information can overcome, and with Quantum World, in which matter and energy fundamental properties are limited by nature. In other terms, AI will keep growing, but could encounter a critical barrier and not develop into ASI.

The brazilian neuroscientist and physician Miguel Nicolelis, in an interview for the news portal *Opera Mundi* (2023), also highlights the existence of achievements that would be impossible for AI – referencing Alan Turing's thesis. According to Nicolelis, the unknown aspects of what these systems can actually accomplish have given rise to a significant myth that has become larger than reality. In his view, the great attributes of the human mind, like love, empathy, solidarity, intuition, imagination, creativity, wisdom, and ethics, cannot be computable and reducible to digital algorithms, despite ongoing attempts by AI developers to achieve this.

Regardless of how far AI will develop, there are still numerous limitations and challenges that arise global apprehensions. In the report *Artificial Intelligence in Science* (2023), published by the OECD[6], the main obstacles for scientists addressed on this matter were:

- Scalability: ML requires large amounts of data, which are often unavailable in theoretical or very descriptive areas of science;
- Annotation and labels: it takes time and resources to label large databases by hand, and there are variation in data access areas of science, which may not allow generalizations;
- Representation of data: capture data and matrices using symbolic representations, like words, images and sounds, to help computers grasp the meaning and connection between data represents a big challenge. This is due to the inherent complexity of human language, which does not readily conform to the structured format essential for computer processing;

[6] It is crucial to note that, while this document presents an interesting exploration of the areas where AI still requires enhancement, its origin from the OECD might align with certain aspects of this organization's recommendations on AI that will be further analyzed.

- The need of a model-driven approach: ML and DL, that are the current popular approaches, struggle when it comes to making logical deductions and abductions. Consequently, for scientific discovery, there is a necessity for abstract modeling to facilitate these kinds of logical inferences;
- The black-box dilemma: the black-box technique present in most ANN approaches is a method that successfully reveals a mass of data correlations, but provide little to zero insight or explanation regarding how these correlations are achieved. In scientific contexts, having a classification without a causal explanation holds limited value, as understanding the underlying causes is crucial;
- Bias: as a legacy of human involvement in science, if individuals labeling data for ML possess varying levels of competence, commit errors, or introduce personal insights and prejudices to do it, the system will learn biased information;
- Classification: even if a ML or DL system accurately classifies an image, altering just a single pixel can lead to numerous misclassifications of similar images. Despite the existence of ANN designed to mitigate this vulnerability, they do not produce a model completely immune to this issue;
- Big difference from human intelligence: humans do not need to think of all possible images derived from flipping pixels to classify something. Instead, humans rely on abstract models of the world, which allows mental simulations of possible modified versions of an object or situation without multiple driving tests like AI;
- Arithmetic operations: statistical ML cannot ‘understand’ arithmetic operations because one cannot feed it with every possible sum between any two numbers, which makes it difficult to replicate human reasoning;
- Overfitting: when AI undergoes extensive training, there is a risk of it memorizing examples without truly ‘understanding’. This lack of comprehension is detrimental, particularly in tasks like arithmetic, where mere memorization falls short in problem-solving;
- Symbolic systems: statistical ML struggles with arithmetics representations and small variations in data, which are tasks that involve symbols. Thus, identifying different types of things or living beings becomes challenging when they exhibit slight differences in appearance, such as dogs’ breeds for example;
- Beyond data size: for scientific applications, AI should not focus on accumulating vast amounts of big data. Instead, an efficient methodological frameworks should be developed for each specific scientific domain;

- Symbolic regression: certain model-driven approaches prioritize symbol manipulation through matching and classification techniques. Yet, the ideal AI system would combine both statistical ML data classification and symbolic computation rule-based inference capabilities;

In conclusion, all this great potential loaded with an extensive list of challenges compose the current stage and progress of AI. The explanations outlined above have been condensed to offer a non-technical perspective, contributing to the comprehension of this new landscape whose implications are emerging in IR. Additionally, the information has been made more accessible, recognizing that not all readers of this study may have a background in the technical aspects of AI. If AI will ever overcome the human mind, innovate, and achieve a utopian stage as depicted in movies and games, is a concern for the scientific community involved in its development. AI, as it exists currently, undoubtedly presents both benefits and risks for individuals, societies, nations, and thereby the international community. However, the only solution to mitigate these risks is to understand AI and regulate it (Nicolelis, 2023).

3. ARTIFICIAL INTELLIGENCE AND INTERNATIONAL RELATIONS IN THE 2020s

This second section endeavors to integrate both topics of this study: AI and IR. The first subsection will focus on the increasing attention that AI is garnering in the international arena, drawing insights from various data sources. This examination aims to illustrate how technology, particularly AI, intertwines with and influences the landscape of IR. Subsequently, it will explore the geopolitical implications stemming from the advancement and utilization of AI. By delving into these aspects, the section will provide an understanding of how AI interfaces with and impacts the dynamics of international relations.

3.1. AI's Growing Presence on the International Arena

AI's expanding capabilities and applications turned it into an essential tool for data processing, analysis, manipulation, and predictions. It starts from basic tools and advances into complex programs designed to perceive and discern patterns through the analysis of extensive databases, which assists in tackling governmental administrative challenges. In this context, AI has garnered attention in the international arena, with escalating public and private investments. Hence, AI's growing presence has ignited debates about the advantages and pitfalls it brings forth.

Nonetheless, an innovation with far-reaching impacts requires comprehensive regulation and preparedness. As a result, an increasing number of countries are developing official national AI strategies and enacting more legislation on the subject. This subchapter aims to explore the heightened attention governments are devoting to AI and how this focus leads to the growth of a burgeoning sector of competition in the international arena.

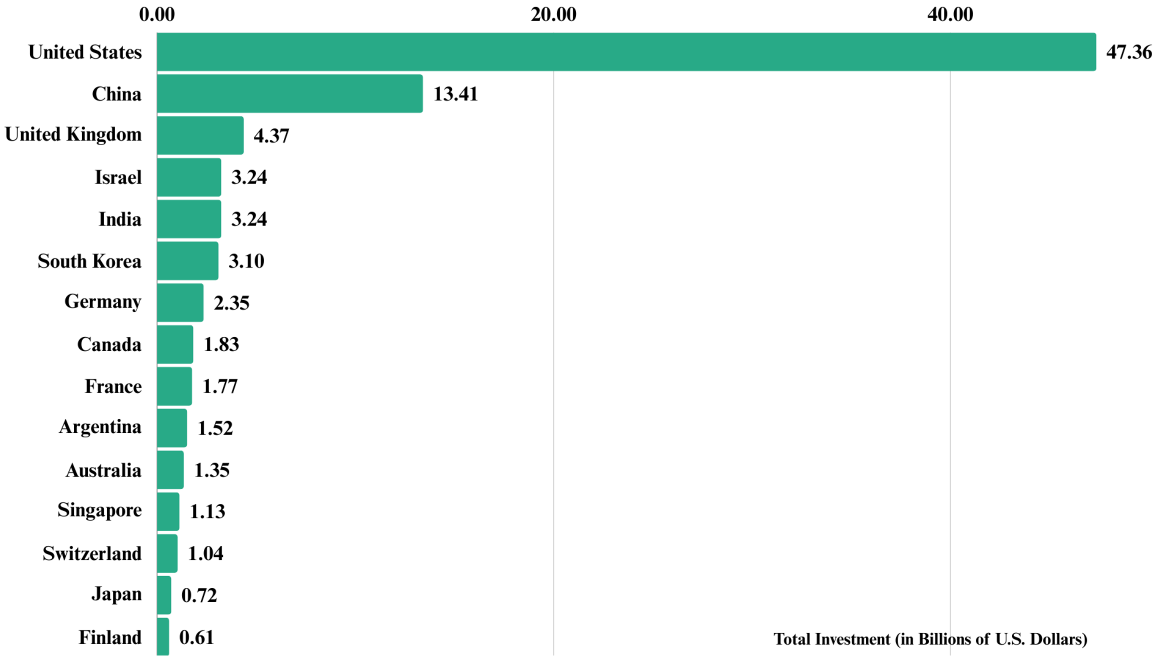
The use of AI by states is present in diverse sectors, from military and defense to trade and diplomacy, becoming essential for governments worldwide. The improvements and facilities that AI can bring to these governance areas have prompted nations to substantially invest in this new source of power to attain a competitive advantage. Although the concept of AI originated in the private sector, "its growth depended largely on public investments, from fundamental, long-term research into cognition to shorter-term efforts to develop operational systems" (Maslej *et al.*, 2023, p. 14). In this scenario, governments' Research and Development (R&D) efforts persistently lead to rapid advances in AI capabilities.

Unfortunately, "no effective benchmarks exist for total nor for government-funded R&D on AI, especially for enabling international comparisons as is commonly aimed for"

(Yamashita *et al.*, 2021). Nevertheless, in the 2022 fiscal year, only nondefense US government agencies allocated $1.7 billion dollars to AI R&D spendings (Maslej *et al.*, 2023). Despite the impossibility of presenting this type of data from a global comparative perspective due to measurement challenges, the US investment in this specific area of AI exemplifies the immensity that this technology is gaining in the international arena.

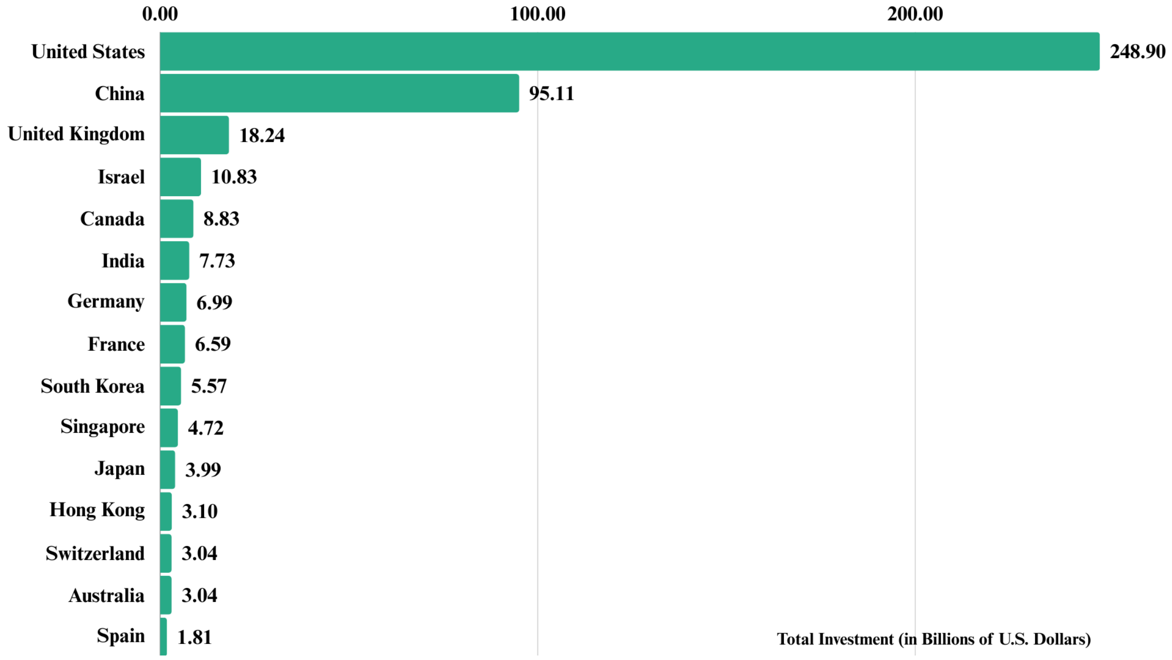
On the other hand, the possibility to track corporate investment contributes significantly to constructing a more comprehensive perspective on how AI is increasingly integrating into the global economy. A curious observation is that the Covid-19 pandemic acted as a catalyst for the widespread adoption of AI, evidenced by a 40% surge in private sector investment between 2019 and 2020 driven by the urgency to embrace digital transformation (Thillien *et al.*, 2022). According to *The AI Index 2023 Annual Report* produced by the Stanford University, in 2022 the US once again led the world in terms of total AI private investment, with $47.4 billion, followed by China with $13.41 billion. When examining this dataset starting from 2013, there are some minor shifts in ranking, but with the US and China consistently maintaining their dominant positions.

Figure 4: AI's Private Investment by Geographic Area in Billions USD, 2022



Reference: Adapted from NetBase Quid, 2022 *apud* Maslej *et al.*, 2023, p. 189.

Figure 5: AI's Private Investment by Geographic Area in Billions USD, 2013-2022 (sum)



Reference: Adapted from NetBase Quid, 2022 *apud* Maslej *et al*., 2023, p. 190.

The US and China stand out as the primary countries making substantial private sector investments in AI. While the gap between China's investment (second place) and that of the third-ranked country is considerable, the difference compared to the US (first place) is even more striking. This allows for three key analyses: (i) in terms of AI private investment, the world remains marked by North American hegemony due to its considerable lead in investment; (ii) although the difference between the first and second place is significant, the gap between the first and third is even wider, showcasing China's growing presence in the sector as an emerging and formidable competitor; (iii) notably, while a developing nation like Argentina stood out in 2022, a long-term analysis of cumulative investments from 2013-2022 reveals that no country in similar conditions consistently maintains enough prominence to appear in the top ranking.

Despite the US' predominance over China in the ranking, it is essential to outline that these data solely consider private investments. This limitation prevents this paper from conclusively determining whether this indicates a reestablishment of the US' complete prevalence over China, thereby not discrediting academic discussions on China's emergence as a significant competitor to North American power. Moreover, the data underscores AI as another sector contributing to the growing inequality between developed and developing nations.

In addition to the number of investments, the rising number of governments worldwide launching official national AI strategies highlights the increased perception around the proportion that AI is taking. These strategies are policy plans to steer the development and deployment of AI. Canada led the way by launching its strategy in 2017, followed by China and Finland in the same year. Since then, another 59 countries released theirs as well. This demonstrates an increasing emphasis on the management and regulation of AI technologies.

Table 1: Yearly Release of AI National Strategies by Country

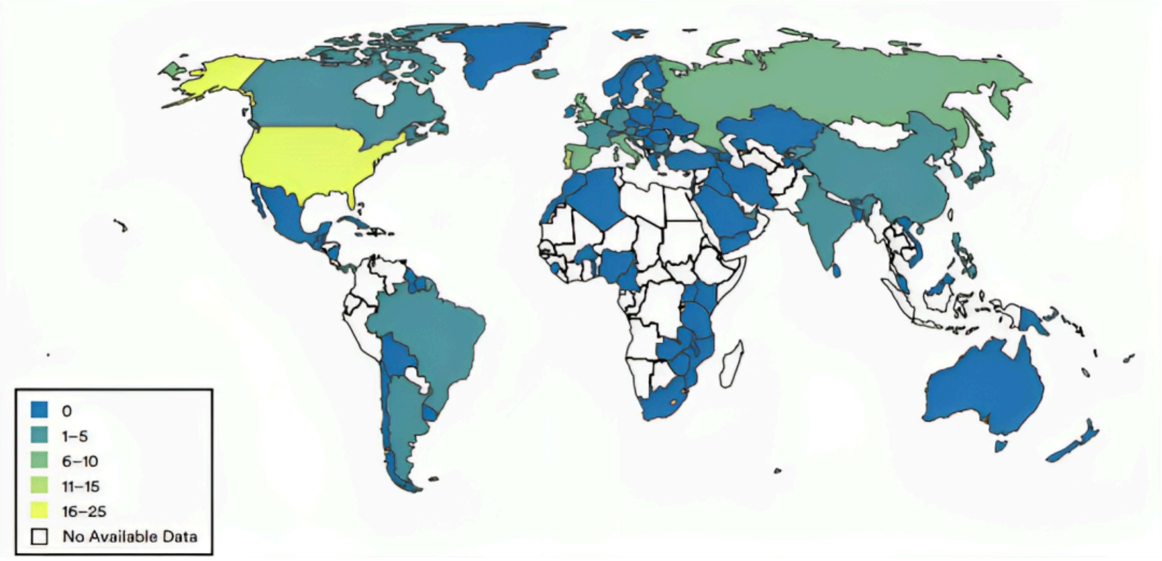
Year	Country
2017	Canada, China, Finland
2018	Australia, France, Germany, India, Mauritius, Mexico, Sweden
2019	Argentina, Austria, Bangladesh, Botswana, Chile, Colombia, Cyprus, Czech Republic, Denmark, Egypt, Estonia, Japan, Kenya, Lithuania, Luxembourg, Malta, Netherlands, Portugal, Qatar, Romania, Russia, Sierra Leone, Singapore, United Arab Emirates, United States of America, Uruguay
2020	Algeria, Bulgaria, Croatia, Greece, Hungary, Indonesia, Latvia, Norway, Poland, Saudi Arabia, Serbia, South Korea, Spain, Switzerland
2021	Brazil, Ireland, Peru, Philippines, Slovenia, Tunisia, Turkey, Ukraine, United Kingdom, Vietnam
2022	Italy, Thailand

Reference: Adapted from Maslej *et al.*, 2023, p. 285.

In addition to the development of national strategies, more countries are enacting AI-related legislation. An analysis conducted by *The AI Index 2023 Annual Report* covering 127 countries, from 2016 to 2022, revealed that 31 of them have implemented at least one law on AI, collectively accounting for 123 in total. Notably, the US alone has passed between 16 and 25 AI-related bills into law, with 9 only in 2022. This acceleration in the US' AI regulation is a result of the release of OpenAI's ChatGPT (Lazard, 2023).

Despite China being recognized in the *Ethics Guidelines Global Inventory* for having fewer domestic guidelines than the US and the European Union (EU), a study from the University of Turku, featured in a section of *The AI Index 2023 Annual Report*, showed that the Chinese research communities do not have a significant overlap with Western ones. In an analysis of 328 papers related to AI ethics in China published from 2011 to 2020 included in the China National Knowledge Infrastructure platform, privacy issues emerged as the most discussed topic among these papers. In addition, other topics were included, such as equality – bias and discrimination –, moral agency, AI arms race, ethics of predatory marketing, and media polarization. Therefore, the range of concerns for Chinese researchers were very similar to the ones from the West (Maslej *et al.*, 2023).

Figure 6: Number of AI-Related Bills Passed Into Law by Country, 2016–2022



Reference: Maslej *et al.*, 2023, p. 267.

The reason behind an innovation receiving billion dollar investments and heightened attention for more regulatory framework can be related to its capacity to expand the power margin of countries. According to Dahl, “A has power over B to the extent that he can get B to do something that B would not otherwise do” (Dahl, 1957, p. 202-203). Nonetheless, there are multiple dimensions of power that are deeply interdependent, which means international power is a complex system (Granados; Peña, 2021). Hard power, rooted in economic, financial, and military might (Mearsheimer, 2001; Waltz, 2010), contrasts with the concept of soft power, which emphasizes cultural and diplomatic influence in this world interconnected by information technologies (Slaughter, 2017). However, within these concepts lie dimensions considered as powers in themselves or means to attain power, such as military power, financial power, and diplomatic ties. These elements are interconnected, as success in one dimension can catalyze achievements in others or facilitate their attainment (Granados; Peña, 2021). Nevertheless, “the old formula of developing economic power and then transforming it into military capabilities is already useless, because complex modern science and technology are necessary conditions to achieve economic and military advance" (Brooks; Wohlforth, 2016 *apud* Granados; Peña, 2021).

The concept of AI aligns with Charles Weiss's definition of technology as "any application of organized technical knowledge about the natural world for a practical purpose, or the capacity to develop and use such knowledge" (Weiss, 2015, p. 412). Furthermore, technology does not act by itself on international relations. The reason why it is a condition to achieve power is because it is combined with economic, political, legal, and cultural forces (Weiss, 2015). Therefore, AI is becoming another tool for achieving foreign policy goals and competing on the global stage due to its capabilities as a technological force to obtain power. Consequently, and given that "advances in [...] technology frequently put new issues on the agenda of the international community", AI has turned into another source of international competitiveness, which can be observed in the data below.

According to *The Global AI Index* (2023) by Tortoise published in June, the current top 10 countries ranking in AI capacity are the US, China, Singapore, the United Kingdom (UK), Canada, South Korea, Israel, Germany, Switzerland and Finland. This research considers 111 indicators collected from 28 different public and private data sources and 62 governments (Cesareo; White, 2023). On a three pillars basis – investment, innovation and implementation –, the US scored 100 out of 100 taking first place on each pillar and overall. China came in second scoring 62 out of 100 in the total, maintaining a significant gap from the North Americans. Subsequently, with Western Europe, Eastern Asia and North American countries following next, South America and Africa lag behind, not making into the top 10 list. It is important to highlight that this rank combines 'scale' – a nation's absolute AI capacity – and 'intensity' – AI capacity relative to the size of a country's population or economy – to obtain a holistic perspective.

When analyzing all the previous data, it is possible to conclude that the US is largely ahead in the competition for AI leadership, from investments and regulation to capacity. This leadership possibly aligns with the fact that a majority of AI business are headquartered in the US. As of September 2023, the US had around 15,000 companies engaging in the field (Thormundsson, 2023), including some of the the largest ones in the world, such as Google, Amazon, Facebook, Apple, and Microsoft (GAFAM). These enterprises have already integrated ML into the core of their technology (OECD, 2023).

Despite being behind the US, China is also a big competitor, featuring companies like Baidu, Alibaba, and Tencent (BAT), and is actively striving to catch up with the US. Notably, "although the United States and China continue to dominate AI R&D, research efforts are becoming increasingly geographically dispersed" (Maslej *et al.*, 2023, p. 22) with other developed countries trying no to fall behind. It can be observed that there is a global ecosystem with thriving AI research sectors, such as the UK, France, Russia, Israel, Japan and South Korea (Villasenor, 2018). On the other hand, it is important to acknowledge that in developing countries this technology can exhibit an even more profound dual-edged nature. While it has the potential to tackle economic and social challenges, it can also exacerbate them by contributing to job outsourcing, deteriorating human rights and dignity, and widening the gap with developed nations.

The data presented underscores the stark reality of an unequal distribution of AI advancements, creating a profound gap. Therefore, "this new 'global digital' landscape significantly impacts the ability of states to remain competitive, fostering an unbridled competition where governments and companies perpetuate inequalities, both at local and global levels" (Tworek, 2022 *apud* Pedroso; Capeller; Santos, 2023, p. 244). The disparity in AI investments between the US and China, as well as between the US and other nations, notably underscores this divide. Furthermore, this gap notably widens when analyzing developing countries. Consequently, AI emerges as a significant force in shaping the global landscape, exacerbating the widening technological chasm between nations.

In light of the facts presented, it is possible to observe how AI is playing an ascending and significant role in the international stage. This multifaceted AI landscape already demonstrates the presence of possible geopolitical implications deeply embedded within the sphere of IR that might be significantly impacted by the progress of this technology, which will be explored next.

3.2. Geopolitical Implications of AI

The advancement of AI is giving rise to inevitable and consequential implications in geopolitics, particularly when considering the alterations it triggers in the interplay of new relations and dynamics between territories, space-time dimensions, and intangible elements (Miailhe, 2018). In a interview podcast titled *How AI Could Upend Geopolitics* (2023), conducted by the Foreign Affairs magazine, Ian Bremmer, the founder of the Eurasia Group, and Mustafa Suleyman, founder of the AI companies DeepMind and Inflection AI, highlight what they term the 'AI Power Paradox'. This concept encapsulates how this rapidly evolving technology is driving a transformation in power and giving rise to a new reality where geopolitical forces extend into a cyber and border-free dimension. Given that geopolitics covers matters of industry, trade, economy, cooperation, security and diplomacy, these changes, to which policymakers are struggling to keep pace, have a profound impact on the balance of power and its actors. This major shift in the international system exacerbates already existing problems and rivalries to the same extent that creates new challenges.

The capacity of AI to handle large amounts of information to be utilized in ways not possible before makes it a political power amplifier (McBride, 2023), including for hard and soft power applications. In the first case, AI being used to enable military and defense improvements is likely to affect the geopolitical balance of power in terms of warfare (Franke, 2021). The implications that derive from this fact are still a subject of debate with different opinions ranging from extreme to more moderate positions. While some believe it will alter the nature of war and psychological essence of strategic affairs, others only focus on limited changes in weaponry (Franke, 2021). The current Ukraine and Russia conflict is turning into a living lab for AI warfare. Ukraine, for example, is using neural networks to combine ground-level photos, footage from drones and satellite images to obtain fast analysis to produce strategic and tactical advances (Bendett, 2023). This situation may serve as a case-study to identify the potential changes AI is actually bringing to warfare with governments investing to "give their military forces the decisional edge on the battlefield" (Kapetas, 2020).

Some of the areas and functions in security and defense that AI may support are (i) intelligence, surveillance, and reconnaissance due to its ability to deal with big data; (ii) logistics – maintenance and transfer of personnel and material; (iii) cyberoperations – from espionage activities to shutdown of a country's infrastructure; (iv) command and control of military operations; (v) swarming – multiple remotely controlled units of systems, like drones; (vi) nuclear-related purposes; and (vii) desinformation. This last case is harmful both in AI-enabled cases or not and intentionally done or not. For military purpose, deep fakes – images and videos altered to spread false statements – could be spread to incite conflicts, persuade people to take a stance to advocate for, and justify and mask a frowned-upon behavior. A recent and evident example is the conflict between Israel and Hamas militants in Gaza, in which deep fakes have been deployed by both sides and spread online (Murphy, 2023).

However, focusing on the development of AI as a tool exclusively for hard power would be a mistake (Miailhe, 2018). This innovation is impacting soft power aspects of economy, commerce, politics and culture. The first two areas are mainly affected by a country's development of digital infrastructures underpinning the technology, such as the investment in microchips, data centers and AI projects. This type of activity enhances a country's strategic economic and commercial development capabilities by boosting productivity and efficiency. Therefore, this has the potential to spur economic and labor growth, enhance customer experiences, optimize e-commerce, and attract international negotiations. Politically, culturally, and also commercially, AI exerts indirect influence globally (Miailhe, 2018). Nonetheless, the role of AI-driven recommendation algorithms in heightening political divisions is a significant concern (World Economic Forum, 2023). These algorithms manipulate and polarize societal viewpoints, keep users within content bubbles aligned with their interests and beliefs, and restrict the access to diverse perspectives. As a result, the functionality of AI facilitates the creation of deep fakes that mislead the public.

It is still uncertain if AI will definitely be the key tool for military, economic and ideological dominance and if its advantage will lead any one nation to acquire pre-eminent power (Kapetas, 2020). Still, this belief, spurred by AI's integration within the discourse of geopolitical competition, is driving intense national races for AI-powered monopolies in almost every sector, like energy, infrastructure, health, online gaming, telecommunications, news, social media, and entertainment (Kapetas, 2020). In light of the facts, this type of hard and soft power amplifier – positively or not – stimulates the proliferation of AI mission statements by states. This competition of related benefits and know-how for AI dominance has the potential to deepen, divide and aggravate international rivalries (World Economic Forum, 2023), especially the Sino-American one.

According to Franke (2021, p. 13), "no geopolitical development is likely to shape global stability as much as Sino-American competition [...]. And AI plays an important role". The technology was even highlighted by the Director of the CIA, Bill Burns, as the central component of the competition with China (Franke, 2021). As mentioned in the first section, the *AlphaGo* american supercomputer beating the Chinese in its own game Go in 2016 was a representation of how AI is boosting this contest by being used to humiliate the competitor also with simplistic elements. This event was called China's AI 'Sputnik moment'[7] and motivated the Chinese to pursue even further research and development (Franke, 2021).

This rivalry was further intensified by China's 2017 *Next-Generation Artificial Intelligence Plan*, released in July, which aimed to catch up with the US and achieve strategic objectives by 2020, 2025, and 2030. The State Council issued a notice outlining that:

> In the face of new situations and new demands, [they] must proactively seek and adapt to changes, firmly grasp the major historical opportunities for the development of artificial intelligence, closely follow development, analyze the

[7] Reference to when the USSR brought the first satellite into orbit in 1957 during the Cold War against the US.

> general trend, proactively plan, grasp the direction, seize opportunities**, lead the new trend of artificial intelligence development in the world**, and serve economic and social development, supporting national security. This will drive the overall improvement and leap-forward development of **national competitiveness** (State Council, 2017, p. 8).

This statement underscores China's long-term strategic ambitions to surpass the US and assert global leadership in AI development. However, the US was not idle in its efforts to maintain a leading position in this competition. In December 2017, the US unveiled the *National Security Strategy of the United States of America*, emphasizing concerns about increased risks to national security stemming from adversaries merging personal and commercial data with ML and AI-driven intelligence gathering and data analysis capabilities (White House, 2017). Additionally, the document affirmed that:

> To maintain [their] competitive advantage, **the United States will prioritize emerging technologies critical to economic growth and security**, such as data science, encryption, autonomous technologies, gene editing, new materials, nanotechnology, advanced computing technologies, and **artificial intelligence** (White House, 2017, p. 20).

These strategic approaches depict the global competition for AI dominance, heightening the rivalry between China and North America. While *The Global AI Index* for 2020-2021 highlighted China's advantage over the US in specific sub-pillars, the data showcased for 2023 positions the US as the primary frontrunner. Apart from rankings, the US leads in private investments, regulation, and capacity in AI, as indicated in the previous subsection.

The North American and Chinese governments are fully funding efforts in AI applicabilities in diverse sectors, including in the military, with a focus on each other as mutual mirror images. These digital empires are, therefore, benefitting from the "acceleration of their concentration of power in the economic, military and political fields thanks to AI" (Miailhe, 2018, p. 107). This situation creates a scene where countries unable to tax ultra-profitable AI companies to subsidize their workers will be forced to negotiate with either the US or China and become dependent on supplies (Franke, 2021). Therefore, while the North Americans and the Chinese are "increasingly locked in a competition with each other" (Franke, 2021, p. 16), accumulating and concentrating this power amplifier, certain nations find the need to divert their attention to matters of regulation and comparative advantages to safeguard against dependency. Meanwhile, other countries are serving as testing grounds for AI applications, potentially leading to a scenario where they become unable to detach from this reliance.

In this context of Sino-American leadership, in 2017, countries worldwide began launching official AI national strategies – as stressed in the previous subsection –, aiming to identify and develop comparative advantages they could aspire to (Miailhe, 2018). Canada, Mexico, France, Italy, the UK, the European Commission, Denmark, Finland, Sweden, the Baltic region, the United Arab Emirates, India, Singapore, South Korea, Japan, Taiwan, and even China – despite also aspiring to global leadership – established strategies in various sectors tailored to each individual country's needs, like education, R&D, digital infrastructures, public services, and ethics (Miailhe, 2018).

According to the 2021 *Coordinated Plan on Artificial Intelligence* of the EU, the bloc is also characterized by its ambition for global leadership in trustworthy and ethical AI, propelled by its strong research community (European Commission, 2021; Franke, 2021). Nevertheless, it lacks a comfortable or secured lead in any of the AI capabilities, besides being very dependent on foreign Sino-American supply (Franke, 2021). The EU's potential necessity to eventually choose sides is hindered by its current stance in this competition. As expressed, "while the US is the EU's most important and closest ally, and China a systemic rival, China also is a cooperation partner on some topics, and an important partner in trade" (Franke, 2021, p. 16). In this context, any side chosen by the EU would be contradictory to some of its interests. This is leading the Europeans to adopt strategies of non-alignment and prioritizing groundbreaking regulatory initiatives. This approach aims to position the EU as a role model for others and a mediator between the divergent technology policies of the two dominant powers in this domain.

In regard to developing and emerging countries, the situation varies geographically. In Latin America[8], the national AI strategies emphasize priorities like cultivating local talent, strengthening the technological infrastructure and ensuring responsible AI (Thillien *et al.*, 2022). Nonetheless, this technology's policies in this region have a high degree of discontinuity as administrations change, resulting in several policy stagnations. In addition, given that home-grown AI-focused workforce is critical to the region, policymakers end up relying on global expertise and foreign inflows (Thillien *et al.*, 2022).

Subsequently, in Southeast Asia, only Singapore is ranking as an AI leader. Some other cases encompass countries that embrace AI, but lag behind in implementing concrete regulations, like Thailand, Malaysia, Indonesia, Brunei, Vietnam and the Philippines.

[8] This is based in a study from The Economist Group in 2022. Despite also mentioning Paraguay, Uruguay, Bolivia in some observations, the document used to analyze this region only considers the potential evolution of AI in five Latin American countries: Argentina, Brazil, Chile, Colombia and Mexico.

Moreover, Myanmar, Cambodia, and Laos do not prioritize AI, as their attention is directed towards issues related to poverty and internal conflicts (Pan, 2023). In the Asian South, India is the most prominent country with AI capabilities, “specializing in applications specific to developing countries” (Miailhe, 2018).

When it comes to African countries, this region is yet-to-develop in terms of digital infrastructures oriented towards AI. Of the 54 countries in the continent, only six have national AI strategies. Consequently, developed nations with advanced AI capabilities see this enormous potential for exploring the technology’s applications and inventing new business and service models (Miailhe, 2018). The Chinese investment in African countries has intensified and created a very unequal techno-industrial partnership grounded in exports of solutions, technologies, standards and company models. To avoid falling behind, the GAFAMI companies are multiplying startup incubators and support programs to develop African talent as well.

Considering the context provided, it is possible to observe that AI’s influence in the Sino-American rivalry is expanding the effects this competition has in other countries worldwide. This contest also extended the competition for geopolitical power to one between AI-enabled authoritarianism and liberal democracies (Franke, 2021), especially in the eyes of liberal governments. According to a study on the promotion of internet content control norms in regional and international institutions, countries like China and Russia are accused of censorship and promotion of illiberal content control norms with justification on information security (Flonk, 2021). These internet penetration initiatives are more accessible to authoritarian regimes as they enhance the government’s ability to control the population (McBride, 2023), turning it into a more useful control than liberation tool. Consequently, China’s exports of surveillance systems, telecommunication equipment, and facial recognition software to countries like Zimbabwe, Malaysia, and Ethiopia as part of the Belt and Road Initiative has raised concerns among liberal and democratic competitors. They are skeptical about the ethical problems of these investments in developing countries, even though the topics of concern for Chinese researchers in AI policymaking from 2011 to 2020 are similar to those of their Western counterparts (Maslej *et al.*, 2023), as highlighted in the previous subsection.

Moreover, AI’s implications on governance are not restricted to relations between democratic and authoritarian regimes, but also within governments. The learning capability of algorithms have supercharged the possibilities for ‘sharp power’, that is, “the manipulation of public sentiment through computational propaganda, disinformation and conspiracism by foreign actors and their domestic proxies" (Kapetas, 2020). Deep fakes and bubbles of matching-interest content have been pushing internet users towards more extreme content (World Economic Forum, 2023), encouraging behavior change. Some concrete examples of AI-generated disinformation affecting internal processes and, thereby, provoking consequences in international relations include partisan operatives attempting to discredit Joe Biden's son, Hunter Biden, by falsely connecting him to the Communist Party of China (Collins; Zadrozny, 2020). Another instance occurred when China interfered in the Taiwanese presidential election in 2020 using AI-generated disinformation (Kapetas, 2020).

The difficulty of imposing globally recognized ethical and normative patterns is also enhanced by another change that AI is causing in the balance of power. AI creators are emerging as geopolitical actors, joining nation-states, and increasing big tech's influence to shape politics. As a result, in this new context where companies are the agencies possessing resources, creating algorithms, and developing AI, there is a certain loss of state control regarding accountability in the technology's use. Furthermore, "it means that the state has less sway to influence the direction of research in a direction that is beneficial to it" (Franke, 2021, p. 20). This situation is drawing attention to the significance of the public-private relationship for determining whether states are really benefited by the private sector achievements. In regard to the military and security sector, for example, companies do not see it as a driving seat of AI's innovation as much as governments would want, and usually prefer not to work with it due to ethical and economic reasons. Although many countries have a division line between public and private entities, there are nations, like China, where this separation is less clear (Franke, 2021).

> When the US government, following a terrorist attack in California, wanted to break into one of the terrorist's iPhone, [...] Apple, refused to [...] provide a back door into the phone's operating system. In the end, the FBI had to hire a private firm from Israel, which used a technology unknown to the FBI to break the phone's encryption. (Franke, 2021, p. 20).

In this context of public-private relations, it becomes evident that AI's competition also includes the access and control of data, talent, and computing infrastructure. Considering these three aspects, despite China's control of the private sector, the US holds a slight advantage. While China has a larger population and, therefore, access to great amounts of data, the North American companies are widely used globally, amplifying both diversity and quantity. When it comes to talent, despite the Chinese efforts to attract individuals from abroad and bring back Chinese talents residing overseas, the US remains highly appealing for foreign researchers and experts. Finally, computer hardware encompasses a large number of products, but the US still appears to be leading, specifically in semiconductor (chip) capabilities (Franke, 2021; Fitch; Ip, 2023).

Finally, this competition between major countries grounded in a new public-private digital era may lead to a new type of nationalism. In order to foster national AI research and companies, states might adopt protectionist policies to safeguard their independence, giving rise to AI Nationalism. In addition to policies that support and fund local AI researchers and firms, this measure could be operated through blocking acquisitions of domestic AI companies by foreign ones. While this stage has not yet been reached and is not guaranteed, there are already some developments in this direction that are taking place (Franke, 2021). For example, the US has imposed export bans on chips and has convinced European firms, under certain conditions, not to export chips to China (Sterling, 2023).

Analyzing this scenario through the lens of traditional IR theories offers insights into the implications of AI. Realism, in particular, views the emergence of AI as a significant concern in areas like security and power (Ndzendze; Marwala, 2023). This perspective aligns with realism's emphasis on the "importance of power in the international system and an appreciation of its centrality to states' capacity for action" (Ndzendze; Marwala, 2023, p. 56-57). AI is recognized as a potent tool for gaining geostrategic advantages, prompting nations to intensify their efforts in defining national strategies. This is an evident indicator of their perception of "how much they deem themselves as being behind the others – this is the AI security dilemma in practice" (Ndzendze; Marwala, 2023, p. 61). This concern ties into states' apprehensions regarding their sovereignty.

Nevertheless, realism's focus on the state as the primary actor in international affairs faces a challenge with AI, primarily reliant on the private sector. In this context, "the impotence of the government outside the cooperation and assistance of non-state actors such as industry, academia, and civil society" (Ndzendze; Marwala, 2023, p. 63) is an affront to realism, especially its structural approach. The acknowledgment of this reality is evident in the stance of the US National Security Commission on Artificial Intelligence (NSCAI), as Chairman Eric Schmidt emphasized that the US government cannot navigate the AI era without collaboration from various sectors (Schmidt *et al.*, 2020, p. 7 *apud* Ndzendze; Marwala, 2023), especially the private one. Similarly, China's approach recognizes the limitations of the state and underscores the importance of the private sector (Ndzendze; Marwala, 2023). This highlights how AI poses a challenge to traditional IR theories when examined through their frameworks, even if it aligns with some of their premises.

Conversely, the liberal perspective allows for explanations of contexts involving non-state actors and can, thereby, ofer valuable analysis regarding AI's economic and transnational economic facets (Ndzendze; Marwala, 2023). This viewpoint aligns with liberalism's emphasis on fostering cooperation and trade to mitigate confrontations between countries by enhancing interdependence among nations (Ndzendze; Marwala, 2023). Yet, analyzing AI through a liberal lens brings some controversies that challenge the notion of AI as "an expansion of the democratic peace thesis[9] and the concept of economic intercedence as a prerequisite for peace" (Ndzendze; Marwala, 2023, p. 73-74).

The idealistic notion of a free market has faced significant flaws due to the intertwined relationships between governments and major corporations, particularly evident in the funding ties to military purposes in AI R&D. The military aim previously addressed in this section and the possibility of an AI Nationalism aligns with this criticism. Moreover, the belief in markets' self-regulation to find ideal prices is challenged by AI's disruption of traditional dynamics, undermining the notion of a 'single market'. This transformation is evidenced by research showing "that AI can funnel prices to individual consumers on online platforms" (Marwala; Hurwitz, 2017 *apud* Ndzendze; Marwala, 2023, p. 76). Consequently, AI has segmented and personalized markets, with algorithms predicting individual consumer behavior based on personal data (Ndzendze; Marwala, 2023). Additionally, "democracy and artificial intelligence appear to be having a negative correlation with one another. The more AI has become diffused, the fewer countries have qualified as free societies" (Ndzendze; Marwala, 2023, p. 76).

In summary, the geopolitical debate surrounding AI is instigating a major transformation in the international system environment and creating a cycle where new competition sectors exacerbate existing rivalries, and vice versa. When trying to analyze it through core IR theories, the emergence of AI and its implications also challenge traditional premises. Regardless of theories' perspectives, it is the "deterministic and potentially transformative influence on military power, strategic competition, and world politics more broadly" (Johnson, 2019, p. 147) that explains the need to establish standard definitions, values and policies to guide and shape the ethical and trustworthy use of AI by states. If all this context is rising with only ANI in the picture, the achievement of AGI, or even ASI, could deepen even more the problems in IR.

[9] "Originating from the work of Immanuel Kant, Democratic Peace Theory proposes that democracies rarely, if ever, fight war against other democracies" (Adiputera, 2014, p. 21)

The rise of AI has ignited a new arena of competition among nations due to its potential to consolidate geopolitical power. This shift has transformed the global stage into a battleground for technological supremacy. The ongoing hegemonic rivalry in the AI sector between the US and China is not just a standalone geopolitical implication but also influences other spheres. In AI investments, the US leads in private ones, followed by China in second place. The substantial gap between the two countries underscores North American dominance, while the difference between the first and third places is even more pronounced, emphasizing China's burgeoning influence. In addition, this disparity contributes to the widening economic gap between developed and developing nations, highlighting AI as a contributing factor to global inequality.

However, this competition for AI hegemony extends far beyond economic implications, reaching into national security strategies and the pursuit of global leadership in a field crucial for the future. The rivalry is evident in the approaches to AI adopted by the US and China. The US, citing national security concerns, prioritizes emerging technologies, including AI, as vital for both economic growth and security. Meanwhile, China's Next-Generation Artificial Intelligence Plan outlines strategic ambitions to surpass the US and lead global AI development, aligning with long-term goals aimed at achieving strategic milestones by 2030.

Hence, the battle for AI dominance not only molds economic landscapes but also defines strategies for national security and the pursuit of global leadership. This competition is also evident in the respective state policy approaches between the US and China, each seeking to leverage AI's potential to fortify their positions on the world stage.

4. MULTILATERALISM AND THE WORLD ORDER IN THE 2020s

In light of this study's focus on two multilateral IOs embedded in the current world order, the theoretical framework was defined around the need to understand this context. This section aims to comprehend the correlation between the transformations of multilateralism and the world capitalist order throughout the post-Second World War and post-Cold War periods, the new multilateralism emerging in the 21st century, and the financial neoliberal capitalism that attempts to shape the world order. The conceptualization of multilateralism will rely on Robert Keohane's (1990), John Ruggie's (1992) and Robert Cox's (1992) contributions due to their different perspectives about the term.

With a more minimalist concept, Keohane – one of the main contributors to neoliberal institutionalism in IR – defines multilateralism as the "the practice of coordinating national policies in groups of three or more states, through ad hoc arrangements or by means of institutions" (Keohane, 1990, p. 731). According to his perspective, multilateralism is limited to arrangements involving states, not because transnational relations are meaningless, but because this theme's scope is already so broad that Keokane prefers to restrict the term (Keohane, 1990). Robert Keohane focuses specifically on multilateral institutions, which are "specific and connected sets of rules, formal and informal, that prescribe behavioural roles, constrain activity, and shape expectations" (Keokane, 1990, p. 732).

However, Ruggie criticizes this conceptualization due to its connection solely with the quantitative etymology of the term. Therefore, the scholar expands the concept to a qualitative dimension establishing that "multilateralism refers to coordinating relations among three or more states in accordance with certain principles" (Ruggie, 1992, p. 568). Such principles are indivisibility among the members of a collectivity, which means that one's action affects all of them, and the expectations of diffuse reciprocity (Ruggie, 1992), that is, "the arrangement is expected by its members to yield a rough equivalence of benefits in the aggregate and over time" (Ruggie, 1992, p. 571).

According to Ruggie, "the definition by the number conceals the fact that arrangements formed by multiple units can, in practice, be controlled by one or a few members, distorting the purpose of collective decision-making" (Lima; Albuquerque, 2021, p. 8). This perspective justifies his criticism towards Keohane's minimalist definition. Nonetheless, although both of these concepts are in accordance with the structure of multilateralism within the main IOs, such as the UN, the OECD, the World Bank and the International Monetary Fund, for example, they do not encompass other actors of the international arena. Considering AI's reliance on public-private collaborations, as addressed in the previous section, it becomes essential to introduce a definition that further emphasizes the importance of multilateralism.

Robert Cox (1992) posits that the concept of multilateralism encompasses at least two dimensions: one related to interactions between states and the other involving engagements among various public, private, and civil society actors, mediated by states and international organizations (Almeida; Campos, 2020). In addition, Cox affirms that "multilateralism can only be understood within the context in which it exists, and that context is the historical structure of world order. But multilateralism is not just a passive, dependent activity. It can appear in another aspect as an active force shaping world order" (Cox, 1992, p. 161).

In order to understand the considerations that may have influenced such definitions of multilateralism, it is important to understand the context of its emergence and the trajectory of its loss of legitimacy. Therefore, three periods will be addressed below, the post-Second World War, the post-Cold War, and 21st century events.

The world order is not based on fixed structure, given that it changes according to the historical context. Nevertheless, regardless of each period's *hegemon*, that is, a power or national state that is able to "ensure control over political and economic territories maintained in the form of colonies, dominions, or (inter)dependent peripheries" (Almeida; Campos, 2020, p. 17), the world system is ruled by "two contradictory economic political forces" (Fiori, 2005, p. 68 *apud* Almeida; Campos, 2020, p. 16). These consist of (i) the pursuit of constructing an empire or global state, often marked by uncooperative tendencies and the imposition on other nation-states, and (ii) the resistance of other nation-states aiming to protect their sovereignty (Almeida; Campos, 2020).

The first *hegemon* after the Peace of Westphalia was Great Britain (1845-1875), whose industrial capabilities echoed across Europe and the world with liberal values transforming the economy. Afterward, the insertion of new world actors marked a counter-hegemonic period until the end of the Second World War, when free trade was replaced by protectionism and the European empires failed to maintain themselves as a power due to the destruction caused by the war. This decline is also associated with the "active support to the anti-imperialist cause, especially in SouthEast Asia and in Africa" (Black, 2008, p. 144), given by the communist powers. "Furthermore, although operating in a different fashion, the US also helped undermine the European empires" (Black, 2008, p. 144), since a communist threat in a weakened environment would limit the propagation of their business interests. With a devastated Europe and an economic boom of the North-American economy, the US emerged as a hegemonic state.

Along with the advent of the *pax-americana*, the institutionalization of coalitions gained ground, initiating a period in which IOs emerged reconfigured, aiming to reformulate multilateralism into a "governance model that recognized both the constant possibility of conflict and the responsibility that great powers have in stabilizing the system" (Lima; Albuquerque, 2021, p. 9). This effort aimed to prevent a recurrence of the failures witnessed in the League of Nations, the first IO established in contemporary terms (Lima; Albuquerque, 2021). However, the expectations on the mediation function of IOs were soon disappointed with the beginning of the Cold War. The dispute between the US and the USSR resulted in stagnations and impasses due the ideological differences of both nations. An example of this situation is the paralyzation, within the Security Council of the UN, of debates that were related to the conflicts between the US and the USSR due to the veto power that both had as permanent members (Lima; Albuquerque, 2021).

Since the bipolar order was what hindered the full implementation of the multilateral premises, it was expected that its end would enable IOs to be more active in diverse thematics of global interest. Nonetheless, the gradual revival of neoliberal fundamentalism, which is based on individualism as an ethical-moral value, posed a challenge for this potential. Even during the Cold War, certain indicators hinted at the reemergence of neoliberalism, such as the 1973 and 1978 Oil Crises, and the disastrous results of the macroeconomic adjustments from the 1980s onwards, conditioned by World Bank loans (Almeida; Campos, 2020). However, it was after the Cold War, along with technological, economic, social, political, cultural and ethical changes, that the restructuring of the "new global neoliberal economic order" (Kamat, 2004 *apud* Almeida; Campos, 2020, p. 24) became more evident.

Towards the end of the 1990s, reforms extended beyond the economic domain to encompass state policies. As transnational actors gained prominence in the global arena, the traditional perception of the State as the solitary force behind development was questioned. This can be observed in the World Bank's *World Report 1997*, which highlighted that "the central role of the state would no longer be to drive economic and social development or directly provide services, but rather to act as a catalyst and facilitator of that development" (Almeida, 2017, p. 3 *apud* Almeida; Campos, 2020, p. 24).

The aftermath of the Cold War marked a significant shift in the global landscape, particularly in the adoption of neoliberalism. As the US emerged as the 'victor', its strategy shifted from an informal imperialism to a unilateral one (Martins 2020, p. 27 *apud* Almeida;

Campos, 2020, p. 22). The spread of North American Western values, including democracy, free trade, and human rights, occurred within a unipolar[10] order. Nonetheless, this new liberal world order, characterized by its unipolarity, led to "'an increase in inequality among countries, classes, and individuals' and was associated with 'a succession of localized economic crises [...]'" (Fiori, 2020 *apud* Almeida; Campos, 2020, p. 23).

Despite attempts at a multipolar[11] approach, such as the formation of the G7-G8[12] and the discourse by established multilateral organizations hinting at a shift toward a multipolar world, the reality did not align. This is related to the fact that "the only superpower which had remained in the new unipolar order was reluctant to embrace the multilateral wave unconditionally and wholeheartedly" (Lazarou *et al.*, 2010, p. 12). This hesitation became apparent in the midst of a series of local financial crises that were inherent to the neoliberal paradigm.

These crises eventually culminated in the 2008 Global Financial Crisis (Fiori, 2020 *apud* Almeida; Campos, 2020) – the most severe worldwide economic downturn since the Great Depression in 1929 –, prompting a search for viable alternatives. The rise of forums like the G20[13] and the formation of alliances such as BRICS[14] signaled an effort to investigate alternative global governance structures. This pivotal juncture, marked by the aftermath of the 2008 crisis, triggered a notable shift, indicating a transition from the established unipolar order towards a more genuinely multipolar global dynamic.

This historical context provides a view of how the "shift in the management of the global order and power dynamics among different actors also affected the dynamics of multilateralism and the interstate system, triggering increasingly frequent resistances and retaliations from other states" (Almeida; Campos, 2020, p. 23). The results of such effects can be seen through some events of the 21st century that outlined a lack of legitimacy in

[10] "Unipolarity is a structure in which one state's capabilities are too great to be counterbalanced" (Wohlforth, 1999, p. 9)

[11] "A structure comprising three or more especially powerful states" (Wohlforth, 1999, p. 9)

[12] Formed in 1975, the group includes the US, Japan, Germany, the UK, France, Italy, and Canada, representing major world economies. Briefly expanded to the G8 with Russia joining in 1997, it returned to the G7 following Russia's annexation of Crimea in 2014 (Senado Federal).

[13] The G20 was created in response to the global financial crisis that followed the collapse of Lehman Brothers bank in 2008. The participating countries are South Africa, Germany, Saudi Arabia, Argentina, Australia, Brazil, Canada, China, South Korea, United States, France, India, Indonesia, Italy, Japan, Mexico, United Kingdom, Russia and Turkey, in addition to of the African Union and the European Union. Its Members meet annually to discuss economic, political and social initiatives (G20 Brasil 2024, 2023).

[14] The BRICS are a group made up of Brazil, Russia, India, China and South Africa, initially proposed by economist Jim O'Neil in 2001. With strong economic weight, they represent a bloc that stands out globally due to rapid growth, boosting dialogue and cooperation in various areas, without a permanent structure or specific funding. They work on economic-financial topics, security, agriculture, energy, and seek convergence and collaboration on strategic issues, increasing their interaction and dialogue (IPEA, 2014).

multilateral organizations, leading to what is described as a crisis in Western multilateralism and the rise of a new multilateralism. According to the article *Global Reorganization and the Crisis of Multilateralism* (2020), potential diagnoses for this new multilateralism include unilateral breaches of rules, the emergence of China, and the impact of the Covid-19 pandemic (Lima; Albuquerque, 2020).

The first factor is mostly linked with the US – interpreted as – 'unipolar moment' after the end of the Cold War, which was accompanied by an increasing unilateralism. This positioning was aggravated after George W. Bush invasion of Iraq without the UN's Security Council authorization in 2003, which promoted even more the vision of the US as a 'lonely superpower' (Lima; Albuquerque, 2020). The problem associated with rule breaches is that "if one power breaks the rule unilaterally (as in the case of the invasion of Iraq), the others tend to follow suit (such as the annexation of Crimea and a new security law in Hong Kong)" (Lima; Albuquerque, 2020, p. 8). After all, "since multilateral institutions lack a formal sovereign authority to decide exceptions, the legitimacy of such decisions stems from adhering to the premise of equality among states" (Lima; Albuquerque, 2021, p. 18). Consequently, middle powers are inducted "to behave in the same way in violating the UN Charter, (as in the interventions of Saudi Arabia in Yemen, Turkey in Syria and Israel's positions in the occupied Palestinian territories)" (Lima; Albuquerque, 2020, p. 8).

In 2017, Donald Trump's government agenda *America First* reinforced this unilateral geopolitical conduct. This new national security strategy abandoned the 'moral messianism' and exchanged its liberal and humanitarian convictions for the pure and simple defense of its own 'national interest' (Fiori, 2020 *apud* Almeida; Campos, 2020, p. 23). The primacy position strengthened after the Cold War found in both the Republican governments mentioned identify no limits to power (Lima; Albuquerque, 2020). According to Maria Regina Soares de Lima and Marianna Albuquerque,

> Its logic is realistic in the best contemporary translation of this theory. Its objective is to preserve and increase absolute and relative power and, at the same time, prevent the increase of power of its 'peer competitors' (Mearsheimer, 2001). The result is the belief that the international norms do not operate in favor of the USA. Donald Trump embodies the current version of the primacy policy, further accentuated since China has ceased to be a potential competitor and has become an actual competitor. (Lima; Albuquerque, 2020, p. 8)

This correlation between the US reinforcement of its unilateral and primacy position and China's rising potential as a *peer competitor* is linked with the second factor that underscored a new multilateralism: the rise of new actors. This evolving multilateral landscape is characterized by a departure from the traditional Western-centric multilateral order, indicating a new era where Eastern institutions and influential state actors, particularly China, are exerting a growing influence on the global stage, redefining the contours of international cooperation and power dynamics.

The Chinese emergence as an economic power in a competition against the US is spreading to global multilateral organizations. Some events that exemplifie this situation are "the decision of paralyzing collective security instances, such as the UN Security Council, with the absence of an efficient multilateral management of the pandemic, and with the suspension of US funds and the country's withdrawal from WHO" (Lima; Albuquerque, 2020, p. 9). Furthermore, the World Trade Organization (WTO) is encountering a legitimacy crisis as the Dispute Settlement System stagnates, hampered by challenges in dialogue arising from trade disputes and divergences between the US and China (Fundação Alexandre Gusmão, 2022).

These events highlight IOs' lack of representativity in light of the "refusal to incorporate values different than those of liberal Western normativity" (Lima; Albuquerque, 2020, p. 10). In order for China – and other emerging actors – to be treated and considered fairly as an equal partner, more diverse values need to be incorporated.

Lastly, the Covid-19 pandemic presented a transnational threat that highlighted the crisis in multilateralism, showcasing inadequate management in addressing the need for international regulation. The WHO faced a series of criticism related to its lack of transparency, inoperability, and inefficiency (Almeida; Campos, 2020). Nonetheless, all these criticisms are assigned to the WHO since its creation and are mainly related to the fact that this institution only holds normative power (Lima; Albuquerque, 2021). The paradox, therefore, "lies in the fact that, at a time when collective multilateral regulation is most needed to address public ills, institutions such as the UN and WHO are currently weakened" (Lima; Albuquerque, 2021, p. 17).

In essence, international organizations rooted in multilateralism grapple with a mounting legitimacy crisis, a product of the profound shifts in the global landscape since the conclusion of the Cold War. The post-US-USSR era signaled an early onset of difficulties for multilateralism amid sweeping technological, economic, social, political, cultural, and ethical transformations. As the world transitioned from a bipolar to a unipolar order, significant shifts unfolded, eventually leading to subsequent phases toward a multipolar configuration later in the 21st century. In both these transitional phases, multilateralism encountered formidable challenges.

As unipolarity is related to the concentration of power in one state, “multilateral institutions are inherently vulnerable to hegemonic/unilateralist power, demonstrated vividly during the UN Security Council’s failure to constrain the US misadventure in Iraq” (Newman; Thakur; Tirman, 2006, p. 3). On the other hand, in a context of multipolarity, “multilateral institutions will be less and less able to meet their objectives because states within the international system will disagree on the process for pursuing the common good and the attendant sharing of responsibility” (Laïdi, 2014, p. 351). This does not mean multilateralism is doomed, especially considering its importance to deal with sectors that are difficult for states to monitor, like AI.

Initially entrenched within a unipolar framework and later transitioning into a multipolar scenario, the landscape shaped by the consolidation of the neoliberal global order offers insights into unfolding events that have profoundly impacted multilateralism. At first glance, it seems plausible to envision multilateral institutions playing a pivotal role in guiding AI usage through the formulation of principles, given the technology’s border-transcending nature. However, aligning the perspectives of aforementioned theorists and the multilateralism crisis with the geopolitical implications of AI reveals limitations in this prospect.

Firstly, the development of AI as a power amplifier, both in hard and soft power applications, might incentivize nations to prioritize their individual AI agendas, potentially straining multilateral approaches grounded in norms and rules – considering Ruggie’s definition. Consequently, this divergence could make it challenging for countries to collectively address the multifaceted implications, regulations, and ethical considerations spanning military, economic, commercial, and societal spheres

Secondly, the intensification of Sino-North American competition extends into AI, becoming another arena for rivalry. Multilateral organizations, like the UN and the WTO, already grapple with reconciling divergent interests between the US and China. Therefore, attempting to establish common rules and ethical frameworks for AI could exacerbate this challenge. In addition, the race for AI dominance is creating asymmetrical interdependencies among nations. Countries that are unable to develop their own cutting-edge AI technologies might find themselves reliant on either the US or China for supplies and advancements. This asymmetric interdependence can influence their roles in multilateral settings, leading to challenges in maintaining a balanced and inclusive decision-making within IOs.

Moreover, the growing influence of the private sector, particularly in AI development, demonstrates the need to expand beyond state-centric notions of multilateralism proposed by

Keohane and Ruggie. The changing power dynamics, with tech companies holding considerable influence in AI research and deployment, necessitate a recalibration of power and accountability within the multilateral framework. Cox's definition of multilateralism, incorporating non-state actors, gains relevance, yet it requires adapting to accommodate the pivotal role of companies in AI.

Furthermore, the emergence of AI Nationalism, marked by protectionist policies to foster national AI research and independence, directly challenges the principles of multilateralism and international cooperation. This divergence from shared norms toward prioritizing national interests over collective cooperation creates obstacles for IOs. In line with Ruggie's principle of indivisibility, the adoption of AI Nationalism by a country affects others, disrupting collaborative efforts.

In conclusion, as AI increasingly integrates into the international arena, robust regulations are imperative to balance its advantages and drawbacks. However, this task becomes more difficult in a landscape where the implications of AI may contribute to diminishing the legitimacy of IOs. Therefore, AI emerges as yet another innovation entangled with multifaceted variables, which may pose further challenges for multilateral IOs in this scenario of legitimacy crisis.

5. OECD, UNESCO, AND RECOMMENDATIONS ON ARTIFICIAL INTELLIGENCE IN THE EARLY 2020s

This section aims to delve into the nature, functioning, and historical backgrounds of both the OECD and UNESCO while also examining their respective documents outlining recommendations for AI. The goal is to comprehensively understand these organizations, paving the way for an analysis of how they address AI. By consolidating insights into their characteristics and examining their recommendations, this section will set the stage for revealing the correlation between organizational attributes and their approaches to AI in the last section of this paper.

5.1. OECD: Nature, Functioning and Background

The OECD was founded in 1961 to promote liberal policies aiming at higher sustainable economic growth, employment, rising standard of living, and economic development. Its mission revolves around expanding global trade on a multilateral and non-discriminatory basis while adhering to international obligations. The organization currently counts 38 countries, 20 of which being founding Member States – main European countries, the US and Canada – and 18 having joined over time. Nowadays, out of the total Membership, 26 countries on the list are European. The OECD basically provides a forum where governments can express their experiences and challenges, besides looking for solutions to pressing problems (EEAS, 2021). Through reports and research serving as guidelines and best practices recommendations, the organization plays an important role as a standard-setter that facilitates the negotiation of international agreements.

The Council of the OECD is the body from which all acts of the organization derive and it is composed of representatives from all Member States. Its Chairman, responsible for presiding over ministerial sessions, is designated on an annual basis. This body is assisted by more than 300 Committees, which are experts and working groups that cover areas of policy making, and a Secretariat that works with policy makers and shapers in each country. Furthermore, the activities carried out by the organization in its forums result in four types of instruments. Among these, the decisions and recommendations are adopted by the Council, while the international agreements or substantive outcome documents are adopted directly by the adherents. Although the first two regulamentations are legally binding, the last two mentioned are not.

The document to be analyzed further on – *Recommendation of the Council on Artificial Intelligence* (2019) – is classified as a recommendation, which basically represents "a political commitment to the principles they contain and entail an expectation that Adherents will do their best to implement them" (OECD, 2023). The decisions and recommendations are made by mutual agreements of all Members, unless agreed otherwise unanimously. Each Member holds the right to one vote and, in case of abstentions, this does not invalidate the decision or recommendation, as it remains applicable to all other Members.

The history of the OECD can be traced back to 1960 when the Convention that officially established it was signed. This marked the reconstitution of the Organization for European Economic Cooperation (OEEC), which was initially created in 1948 to oversee the administration of aid provided under the Marshall Plan for the post-Second World War reconstruction of Europe. However, this distribution of aid in order to resume European economic and production growth was only the short term plan for the OEEC. The extended term strategy encompassed also a regulatory facet, "involving the training of personnel in the scientific and technological field, as well as the establishment of mechanisms aimed at promoting trade liberalization and multilateralizing payments" (Pinto, 2000, p. 14). This fact highlights basic elements of the current functioning of the OECD that were inherited from the OEEC, that is, economic coordination between countries and the provision of data and statistics on the functioning of different economies (Pinto, 2000).

The need to reform the OEEC and reconstitute it under the current OECD was driven by three pivotal events and one logic reason. The first circumstance was the beginning of studies focused on integration projects for establishing a customs union within Europe. This initiative was primarily led by France and West Germany, receiving substantial support and incentive from the OEEC. Notwithstanding, differing opinions regarding a larger or smaller scale of integration led to two different tendencies, culminating in a polarization of the discussion and hindering the realization of a European Common Market under the auspices of this organization.

Subsequently, the second event relates to the wave of independence following the final phase of the Second World War. In this period known as 'Era of Decolonization', which was marked by the waning of colonial empires and a proliferation of nationalist movements striving for self-determination, numerous African and Asian countries achieved their independence, starting with India in 1947. This accumulation of independence led to a rise in developing countries joining the UN system. This enabled them to express their needs and opinions, contributing to some extent to the organization's decision-making processes during the 1960s. In this context, the imperative for a more cohesive organization of developed countries, committed to the enhancement of a global market economy system, became increasingly evident.

The last fact that culminated in the remodeling of the OEEC was the period in which all of these independence movements were happening, the Cold War. The new so called 'Third World' countries were seen by the USSR as a great new ground to serve as a role model to exercise influence. Hence, the threat this posed to the US hegemony led the largest development-assisting country to "demand that European countries, in full economic recovery, increase their participation in development aid programs and undertake greater coordination between donor countries" (Pinto, 2000, p. 16). Although a Development Assistance Group (DAG) – later restructured into the Development Assistance Committee (DAC) within the OECD – was created in the scope of the OEEC, it soon became evident that the organization lacked the mechanisms to effectively respond to this new international balance challenge of growing economic interdependence. Considering that its initial focus was on addressing European needs, the OEEC would not be able to handle the significant amount of new actors requiring investment to counter the communist expansion. Therefore, it became clear that concerted action was needed to safeguard and maintain their national economic stability. For this purpose, they needed an organization that would enable them to act collectively in shaping the international economic order towards the consolidation of the liberal market economy model.

Finally, the logical reason behind this reform lies in the fact that the OEEC had already fulfilled its main short term objective, which was the successful reconstruction of Europe. Moreover, it created the groundwork for its extended term goal by fostering discussions on the convertibility of the European currency and integration projects within Europe, rooted in liberal and multilateral principles. On that account, in light of the context highlighted throughout the series of events outlined above, it became evident that a reform was the more strategic direction to take. Therefore, in 1959, it was decided that the OEEC would be restructured regarding its working methods, set of normative decisions, and general objectives. In addition, it was determined that the OECD would act within the principles of the General Agreement of Tariffs and Trade (GATT) to encourage the cooperation between the Member States aiming at international economic stability and growth (Pinto, 2000).

The fact that the same countries that comprised the OEEC would form the OECD allowed for numerous continuities among them, like the "consolidation, among European countries, of the belief that economic development presupposed cooperation and interdependence" (Pinto, 2000, p. 17). Throughout the next 13 years, new members from Europe and from other continents joined the organization, like Australia, Finland, Japan and New Zealand. Despite this expansion, the OECD was still limited to developed capitalist economies.

With the further creation and advancement of the EU as a complex and complete bloc, its significance was reflected as well in the OECD. The EU has a full participant status in the organization, besides being a full Member of some of its bodies, such as the Development Assistance Committee, which deals with cooperation issues with developing countries, and the Development Centre, which comprises countries from Asia, Africa and Latin America. There is also an EU Delegation to the OECD that actively engages in several dialogues with agencies and technical committees – including in areas such as digitalisation, innovation, and development cooperation. The delegation also voluntarily provides financial support to budgeting bodies and actively participates in them to assist in organizing the OECD's resource planning. The European Commission "contributes to the work of the OECD on an equal footing with full Members, except for voting rights" (EEAS, 2021).

In the 1990s, with the end of the Cold War and the advancement of the globalization process, the OECD recognized the need to adapt to the new international context and expand its activities and interests beyond the limits of its restricted circle of members (Pinto, 2000). Emerging economies, like Mexico and South Korea, and transition countries in Eastern Europe, like Czechia, Hungary, and Poland, began to gain access to the organization since then. To join the institution, candidate countries must meet a set of requirements and criteria aligned with the principles that guide the Member States.

Nevertheless, as of 2023, the OECD comprises only four upper-middle-income economies – Colombia, Costa Rica, Mexico, and Turkey –, in accordance with the World Bank classification. Its membership still reflects disparities in geographical representation. In Latin America, for example, only Chile, Colombia, Costa Rica and Mexico are Members, while Brazil, Argentina and Peru are in the adhesion process from 2022. Developing countries in Africa remain absent from the members list. In this context, although the OECD Convention mentions a commitment to contributing to the development of non-Member countries, the organization continues to be "characterized by being restricted and closed [...] to its official participants and guest countries, and by its technical work being related to themes of economic growth and development" (Campos; Lima; Lopes, 2011, p. 33).

All things considered, the OECD is an organization without financial and supranational authority, but with political, legal and economic capabilities on an international level, providing a platform for consultation and coordination among its members. Moreover, it has the "ability to temper academic theory with factual analysis and generate policy recommendations that correspond to the needs of member countries" (Pinto, 2000, p. 19). Such needs, however, encompass a wide range of areas, delineating the OECD as a forum to address matters related to the fields of economy, statistics, agriculture, commerce, energy, environment, education, employment, social issues, science, technology, and financial, fiscal, and industrial policies. Therefore, being part of the organization means having access to a continuous exchange of data and information among its members across these various geopolitical domains. In the case of AI, the urge to establish standardized definitions for a new shared challenge in the global economy found a platform for discussion within the OECD.

The multidisciplinary nature of the OECD often results in a coincidence between their action fields and competencies of other specialized organizations, like the WTO, the Monetary International Fund, the World Bank and UN affiliated agencies, including UNESCO, the Food and Agriculture Organization, the International Atomic Energy Agency and the UN Environment Programme. Even so, the OECD does not seek to compete with these institutions. Instead, its goal is "to engage these other organizations in their discussions and negotiations, with a view to exchanging information and deepening the debates, as well as serving as channels for disseminating the standards developed within the Organization [OECD]". (Pinto, 2000, p. 67).

Finally, considering the origin of the OECD and its founding Members, it is undeniable that these nations continue to exert significant influence within the organization and the outcomes it attains. Additionally, the selectivity for big industrialized countries fosters a sense of homogeneity within it, shaping the decisions made from the perspective of its narrow circle of represented Members. While this might lead to a lack of diverse perspectives, it does foster efficiency in the decision-making process. In addition, it is important to note that, despite aiming to contribute to the economic development of both Member and non-Member countries, as well as to the global economy in general, the OECD was not created as a universal entity for inclusive participation for all countries. Even in its preamble, it is expressed as a belief that economically more advanced nations should cooperate in assisting countries in the process of economic development.

Therefore, the criteria for the type of influence and ideals sought within the OECD have always been explicit. There is a notable inclination toward greater inclusivity of other countries as members or observers, both developing and emerging ones. Nonetheless, the selection of countries for admission remains bound by a specific set of prerequisites aligned with the foundational principles outlined in the Convention by the organization's founding Members. This aspect of the OECD's membership criteria might contribute significantly to the dearth of critical academic analysis highlighting potential crises within the organization. In other words, the selection of Members that share common interests might possibly veil certain deficiencies or challenges inherent in the multilateral model embodied by the OECD.

The OECD operates within a framework significantly influenced by internal dynamics, profoundly shaping the organization's regimes and recommendations. These factors play a pivotal role in setting the agenda, defining objectives, and guiding the OECD's approach to global economic, social, and environmental challenges. Despite its recognition as a key global think tank and provider of comparative data, the OECD has historically been underestimated in its influence and governance compared to institutions like the IMF, World Bank, and WTO (Eccleston, 2011). The institution "has generally assumed a facilitating role, undertaking research and brokering agreements for member (and increasingly non-member) states through a network of what were, until very recently, relatively secretive committees" (Eccleston, 2011, p. 243).

For many years, the OECD remained a discreet force until certain events thrust it into the global governance spotlight. Notably, its prominence surged during discussions on financial regulation in the aftermath of the 2008 Global Financial Crisis. This period coincided with heightened scholarly interest in the OECD's ability to shape social norms and influence state actions (Eccleston, 2011). The OECD played a strategic role in supporting the G20, a group that emerged in response to the crisis, particularly in proposing fiscal transparency. This highlighted its growing importance due to expertise in economic governance. However, uncertainties loom over the OECD's future role, especially given the rise of new economic powers. Challenges related to expanding membership and diversification of visions and values also pose significant hurdles (Eccleston, 2011).

This context provided gives a glimpse of the OECD's approach after the consolidation of a multipolar order. Its response post-2008 crisis illustrates a tendency "to move with the times by following the economic orthodoxy of the day, from Keynesianism in the 1960s, to monetarism in the 1970s and neo-liberalism in more recent years" (Mahon; McBride, 2008, p. 15 *apud* Eccleston, 2011, p. 251). While the OECD's adaptability to evolving economic paradigms demonstrates flexibility, it also raises skepticism about a predisposition to endorse policies within existing frameworks, potentially avoiding challenges to fundamental systemic flaws. These considerations lead to doubts about the organization's willingness to support comprehensive reforms that might challenge the status quo, particularly concerning global capitalism within neoliberalism. As the global power dynamic expands from the Atlantic to the Pacific, questions about the OECD's relevance in an increasingly multipolar world continue to emerge.

5.2. UNESCO: Nature, Functioning and Background

UNESCO is an organization linked to the UN system and whose focus is primarily on the promotion of peace and security through the cooperation between nations in education, science and culture. Nowadays, it has 193 Member States and 7 Associate Members. The proposals within it are distinguished between recommendations and international conventions. In the former case, which is the nature of the document to be analyzed further on, a majority vote shall suffice (UNESCO, 1945). Given that these norms are not subject to ratification but are encouraged for adoption by Member States, their purpose is to influence the development of national laws and practices. Hence, even though Member States are not under legal obligation to adhere to these recommendations, they carry significant political weight, contributing to their moral legitimacy. Consequently, this instrument plays a crucial role in advancing shared international standards and practices among countries.

UNESCO's foundation dates back to 1945 and was preceded by a series of negotiations and deliberations that involved divergent ideas regarding its future. In a first moment, science was not included in the debate along with culture and education. However, the incorporation of science into UNESCO's name and statute can be related to two key factors. Firstly, the strong efforts and compromise of Joseph Needham – a marxist bernalist[15] – and Julian Huxley – affiliated with the Social Responsibility of Science movement – efforts to stimulate international cooperation in the scientific field. Secondly, the heightened awareness raised towards the role of scientists and the importance of collaborating and sharing scientific knowledge after the war, particularly after the recent atomic bombings of Hiroshima and Nagasaki (Elzinga, 2004). In this context, the british Minister of Education, Ellen Wilkinson, expressed in the negotiations for the foundation that "[...] it is important that they [the scientists] are closely linked to the human science and feel that they have responsibility towards humanity [...]" (Sewell, 1975, p. 78-79 *apud* Elzinga, 2004, p. 97).

There were two divergent intellectual movements conflicting to drive organization's international exchange character: one of French origin and the other Anglo-North American.

[15] Follower of John Bernal's (1901-1971) ideas about the social function of science and the organization of scientific research.

The first was derived from the International Committee of Intellectual Cooperation (ICIC), situated in Paris and created under the auspices of the League of Nations as a place for transnational cooperation. The ICIC had a strong elitist character that believed that the "pleiad of the world's brightest minds would be able to rise above the conflict that normally divided nations into political, ideological and other blocs" (Elzinga, 2004, p. 91). The French aimed to carry on the ICIC's legacy and advocated for strong non-governmental representation to keep an independence from the political power. For them, an organization with extensive government representation "could be impeded by the emergence of blocs that would further hinder mutual cooperation and understanding implied in the ideals to be pursued" (Elzinga, 2004, p. 97).

The second approach, inspired by the Conference of Allied Ministers of Education with meetings in London, favored the formation of a global and intergovernmental organization, that is, controlled by Member States. Contrary to the French, supporters of the Anglo-North American idea were convinced that a non-intergovernmental character would be powerless and perpetually stagnated in a stage of philosophy and idealization.

UNESCO ended up initially carrying a hybrid aspect with an intergovernmental structure aligned with universal principles and some non-governmental activities related to the French interests. The tensions within the organization, nevertheless, were related to the external context of the Cold War. In 1954, with the USSR's adhesion to the organization, there was an increased attempt to diminish ideological conflicts by establishing a more instrumentalist vision of science and culture, resulting in studies with a more technical and less critical tone (Elzinga, 2004). The inclusion of countries from the oriental block strengthened the intergovernmental nature and limited the opportunities for collaborations with non-governmental organizations. This shift was an attempt to mitigate ideological differences between the West and the East.

The first director of UNESCO, Julian Huxley, acted as a mediator between pragmatic liberal and left-wing forces. Therefore, "he was denounced both by the Cold War bernalists, who were consolidating forces on the right, and by the communists" (Elzinga, 2004, p. 96). Subsequently, Science Director Joseph Needham deepened the partnership with the International Council for Science (ICSU), contributing to a shift the focus of scientific internationalism towards policy-driven research. However, this evolution perpetuated a bias in favor of industrialized countries, a trend that began to draw criticism due to an enhanced awareness of the need to benefit 'Third World' countries (Elzinga, 2004). In the 1970s, the financial help for these more economically disadvantaged nations – along with the extension of scientific patterns and services – advanced. As a result, external social relevance criterias began to outweigh internal scientific quality considerations.

The ideological polarization within UNESCO got worse after Huxley and Needham left their positions (Elzinga, 2004). According to John Bernal, the organization became "the ideological front of the North American-led majority in the United Nations" (Kolasa, 1962, p. 132-133 *apud* Elzinga, 2004, p. 96). Consequently, although the greater attention and support for 'Third World' countries enabled them to finally have a voice within the organization, science and technology were turning into instruments for strengthening cultural imperialism (Elzinga, 2004).

Nevertheless, due to the influx of Oriental European countries in the 1950s and subsequent waves of decolonization, the number of Members almost doubled, increasing from 70 to 130 participants between 1954 and 1974. This new majority represented a challenge to the hegemony of Western developed countries in culture, technology and information, which caused a change in the ideology within Unesco itself. This shift became more explicit and pronounced with the appointment of Senegalese Amadou Mahtar M'Bow as the General Director in 1974, further directing UNESCO's focus towards the 'Third World'. Consequently, the US and the UK withdrew from the organization in the 1980s with the objective of paralyzing UNESCO, resulting in a one-third reduction in its budget. Due to the impactful withdrawal of the Anglo-North American nations and amid numerous criticisms directed at M'Bow for politicizing the organization, the former Director was replaced by the Spanish Frederico Mayor.

Further on, other stances and attitudes also provoked complaints from the US and other Western nations. The most significant one for this analysis is the controversy regarding the report produced by the Sean MacBride Commission, which exposed the struggle of countries non-aligned to the global power relations. Although it was shelved, the document was cited to support various critical analyses because it showed "that new media penetrates the 'receiving' culture more deeply than any other manifestations of Western technology, producing serious contradictions in developing countries" (Elzinga, 2004, p. 117).

Most critics highlight that the Statute created for the institution in 1945 with the "idea of science and internationalism as instruments for order and justice reflected a particular vision of the Western liberalism" (Elzinga, 2004, p. 117). The prevalence of Anglo-North American ideas for problem-focused orientation over the broad French cultural approaches marked internal discord within UNESCO more generally. However, an organization originally envisioned with a scientific component aimed at fostering cooperation and diversity deviated from its intended path. Instead, it evolved into an entity where content presented under the banner of science was deemed neutral only if it acknowledged the supremacy and universality of Western scientism as the benchmark for evaluating other forms of intellectual life and knowledge (Elzinga, 2004). “This instrumentalist vision was reinforced with the USSR’s entry into UNESCO” (Elzinga, 2004, p. 129).

Despite the dilemmas and controversial topics, all these diverse moments show that the intergovernmental character of UNESCO gave it more authority in the eyes of the public. Nevertheless, it is contentious to affirm if this was positive or negative considering the influences that now shape and guide certain recommendations in favor of specific interests using the organization’s voice. Undoubtedly, there were instances when this autonomy brought forth perspectives of an unheard segment of the international community, albeit it was the more influential voices that tended to receive greater consideration most of the time.

> It is evident that a transnational agency like UNESCO [...] serves as a platform for compromises between the interests of individual nations and those of geopolitical blocs. As an intergovernmental and, therefore, transnational forum, UNESCO has its own life and logic. This formal autonomy creates a space where internationalist ideals can be exposed and, in turn, influence public opinion, even when they are in constant contradiction with the more pragmatic behavior dictated by the interests of the realpolitik of Member States and their coalitions. In the early days of UNESCO, prominent figures from around the world could use it as a platform to embrace the ideal of scientific internationalism, while government representatives emphasized the need to leave the lofty realm of utopian dreams behind and confront the harsh reality of the challenges of what was possible (realpolitik). (Elzinga, 2004, p. 122).

The idea that UNESCO’s science component, along with education and culture, should be international and universal is promoted under the premise that establishing standards in scientific practices is essential, regardless of temporal or spatial considerations. However, it is important to acknowledge that this can also be manipulated for political purposes that serve national interests. All this evolution with scientific credibility “serves as a means of exchange in the political arena” (Elzinga, 2004, p. 125). Therefore, “the stronger the claims of purity and universality of knowledge, the higher the exchange rate for the currency of science” (Elzinga, 2004, p. 125). An organization that should supposedly promote a neutral science was multiple times influenced towards the reaffirmation of Western scientificism as superior and the ultimate reference pattern.

Therefore, the criteria to what type of influence UNESCO wanted to exercise was not always explicit due to the dictomic intellectual movements within it. Nowadays, both the French and Anglo-North American visions are still embedded in activities realized by the organization. Regime and recommendations stemming from UNESCO reflect the internal dynamics of the institution due to their profound impact on its structure and operations. The Anglo-North American influence has shaped UNESCO towards embracing an intergovernmental framework, empowering states with significant control over its decisions and policies. Despite a hybrid approach blending Anglo-North American ideals with the French perspective, these regimes and recommendations often echo the interests and priorities of Member States, which are frequently driven by their distinct political, cultural, and educational agendas. This becomes evident when noting that, despite being rooted in the UN System, which advocates for inclusivity and universality, the discussion in this subsection reveals a discernible influence from a specific group of countries on the notions and contents propagated by UNESCO.

As a UN-affiliated agency, UNESCO faces the ripple effects of the organization's crisis of legitimacy. Since its creation, the UN "was [...] based on Western and liberal values, advocated by the great powers of the time" (Lima; Albuquerque, 2021, p. 8). This fact is also reflected in UNESCO's foundation, which was marked by a divergence between Anglo-North American and French ideas – both Western. Currently, this results in structural resistance to embracing diverse values, leading to the referenced crisis of legitimacy.

This challenge is exemplified by the UN's internal reform impasse, which advocates for the "the expansion of the organization's 'representativeness' [...] in the Security Council, as well as the appreciation and broadening of the role of the General Assembly, which is still confined to recommendations" (Lima; Albuquerque, 2021, p. 16). Furthermore, the violation of the UN's reciprocity principle, initially by the US and later by other states, as previously mentioned, set off a chain reaction. This "breakdown of multilateralism's pillars deepened the crisis of legitimacy and accentuated the counter-hegemonic axis, leading to demands for greater democratization and participation" (Lima; Albuquerque, 2020 *apud* Lima; Albuquerque, 2021, p. 16). Moreover, with the emergence of China as an economic power, the absence of non-Western cultural values has become more evident as a factor exacerbating this crisis within the UN and its affiliated agencies. Paradoxically, UNESCO, which emphasizes culture as one of its focal areas, finds itself entangled in this issue.

5.3. OECD's and UNESCO's Recommendations on AI in the early 2020s

Although the OECD and UNESCO have intersected in certain international policies since their foundations, they inherently differ in their focus and objectives. As previously outlined, the OECD was established to promote policies aiming at higher sustainable economic growth and development through the expansion of world trade. On the other hand, UNESCO is linked to the UN's System and its focus is primarily on the promotion of peace and security through the cooperation between nations in education, science, and culture. In this subsection, an analysis will be conducted to discern whether this divergence in nature is evident in the ongoing discourse surrounding AI within the recommendations of both organizations. As such, it will commence with the presentation and considerations of the OECD's recommendations, followed by those of UNESCO. Finally, the study will present the codes identified through an inductive thematic analysis assisted by MAXQDA software, which will support the final insights drawn from the research.

The content of Annex – A is the OECD's *Recommendation of the Council on Artificial Intelligence*, adopted in 2019 and amended in 2023. A total of 38 OECD Members have signed the document, joined by 8 non-Members, including Argentina, Brazil, Egypt, Malta, Peru, Romania, Singapore, and Ukraine. It starts with a brief background information that highlights the aim "to foster innovation and trust in AI by promoting the responsible stewardship of trustworthy AI while ensuring respect for human rights and democratic values" (OECD, 2019, p. 3). Besides that, values-based principles, five recommendations for policy-makers, an explanation of the development process, and a follow-up and implementation monitoring are presented. After a preamble of considerations, recognitions, and agreements, the document proceeds to its primary focus: the recommendations concerning AI. These recommendations are divided into two substantive sections: (i) principles for responsible stewardship of trustworthy AI, and (ii) national policies and international co-operation for trustworthy AI.

From a pre-coding analysis perspective, the OECD's document holds a dual focus. One aspect is directed towards AI actors, emphasizing the development, application, and utilization of AI. The other aspect is geared towards the actions and governance responsibilities of national governments. This duality is divided between the two aforementioned sections, precisely defining the roles and responsibilities necessary to achieve the ultimate goal of trustworthy AI. It is important to highlight that the presentation of the recommendations begins by advocating that both Members and non-Members of the OECD, who adhere to the Recommendation, promote and implement the following principles for responsible stewardship of trustworthy AI. The Recommendation not only encourages the Secretary-General and Adherents to disseminate it but also urges non-Adherents to consider and adhere to it. Furthermore, it provides directives for ongoing efforts in the field and emphasizes continued oversight through supervision.

On the other hand, Annex – B contains UNESCO's *Recommendation on the Ethics of Artificial Intelligence*, produced and approved by its Member States in 2021. The document is nearly double the length of the OECD's. Before addressing the recommendations, it has parts dedicated to the preamble, scope of action, aims and objectives, values, and principles. At first glance, it is apparent that UNESCO, while encompassing similar introductory elements with subtle variations, opts for a more extensive and detailed approach compared to the OECD, which adopts a more concise strategy.

The divisions within the list of recommendations are notably more expansive in UNESCO, totaling 11 in all: (i) ethical impact assessment, (ii) ethical governance and stewardship, (iii) data policy, (iv) development and international cooperation, (v) environment and ecosystems, (vi) gender, (vii) culture, (viii) education and research, (ix) communication and information, (x) economy and labour, and (xi) health and social well-being. These delineated policy action areas reveal a profound level of detail within the recommendations. The document concludes by underlining the significance of monitoring and evaluating AI policies, programs, and mechanisms while reiterating the commitment to respect, promote, and protect the Recommendation. Furthermore, it emphasizes the necessity of viewing the propositions as an integrated whole, ensuring alignment with international law obligations and rights, and refraining from endorsing actions conflicting with them.

Prior to applying the coding method, it is evident that, while the document acknowledges multiple stakeholders, it predominantly focuses on policymakers. Member States are emphasized as the key agents responsible for translating the core values and principles proposed into action. While AI actors are referenced, the document primarily portrays them as entities to be encouraged, included, ensured, or required by states to take specific actions, rather than directly engaging with them regarding their responsibilities. Another crucial distinction to note is the document's core emphasis on ethics, evident from its title – a notable difference from the OECD's approach.

In summary, the OECD adopts a succinct and objective discourse more balanced between AI actors and governments, which can be interpreted as the necessity to combine public and private entities to accomplish the Recommendation's aim. This type of goal, discourse, and focus aligns with the OECD's nature and background towards fostering economic prosperity and equitable competitiveness. On the other hand, while not excluding AI actors and other stakeholders, UNESCO gives more emphasis to the governments' responsibilities with vast and detail-oriented suggestions. Due to its aim "to provide a basis to make AI systems work for the good of humanity, individuals, societies and the environment and ecosystems" (UNESCO, 2021, p. 5), the document uses a more human-centric linguistic focus on well-being and respect for human rights and fundamental freedoms. In addition, it explicitly highlights the aim "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" (UNESCO, 2021, p. 5).

As explained in the introduction to this study, this research also holds an inductive analysis assisted by MAXQDA software. The exercise produced during this method was to categorize patterns and frequency of words and themes in an open coding approach. In this type of analysis, "the maintenance of rigor [...] becomes overshadowed by the imprint of the researcher's subjective understandings, thus establishing correspondences between linguistic and psychological structures denotes a weakness" (Rocha; Deusdará, 2006 *apud* Menezes; Filho, 2022, p. 9). In spite of that, this factor was not overlooked during the process. As the codes will serve more specifically to identify if both documents ultimately address similar topics, it is believed that this inevitable subjectivity does not disqualify the goals intended.

The thematic analysis began with an initial reading of both documents, which were then revisited to note patterns in similar use of words and themes of discussion. Subsequently, an open coding method was employed and the 56 codes found were later classified into 8 clusters of wider meaning. It is essential to highlight that the coding process was restricted to Sections 1 and 2 – from page 7 to 9 – of the OECD's Recommendation and the 11 areas of policy action – from page 10 to 20 – within the UNESCO document. These limitations in observation were dictated by the software's constraints. Nevertheless, analyzing these specific parts will aid in assessing whether, despite the outlined disparities between the IOs, their recommendations fundamentally converge in their final objectives or not.

Table 2 reveals that both documents essentially address the same themes, except for six codes – auditability, proportionality, data governance, market and consumer protection, cultural preservation, and human oversight – uniquely mentioned by UNESCO. There are a couple factors that could explain these few differences. As aforementioned, UNESCO's document is more extensive and detail-oriented, featuring 3 to 17 paragraphs per policy action, translating each section's focus into actionable guidelines. As a result, the OECD's more direct approach might exclude some terms. Also, despite amended in 2023, the OECD's Recommendation was produced two years before UNESCO's one. Given AI's rapid evolution, this time gap allows for the emergence of new challenges and problems.

Table 2: Thematic Analysis: OECD’s and UNESCO’s Recommendations on AI

Clusters	Codes	OECD	UNESCO
		Recommendation of the Council on Artificial Intelligence	*Recommendation on the Ethics of Artificial Intelligence*
Governance Pillars	AI Actors	✓	✓
	Democratic Principles	✓	✓
	Digital Ecosystem (Infrastructure, Research, and Skills Development)	✓	✓
	Governance and Stewardship	✓	✓
	International Cooperation	✓	✓
	Monitoring and Evaluation	✓	✓
	Multistakeholder	✓	✓
	Redress of Harm	✓	✓
	Trustworthy and Integrity	✓	✓
Transparency & Accountability	Access to Data and Information	✓	✓
	Accountability	✓	✓
	Auditability		✓
	Awareness Promotion	✓	✓
	Data, Best Practices, and Benefits Sharing	✓	✓
	Explicability and Explainability	✓	✓
	Open and Understandable Communication	✓	✓
	Responsibility	✓	✓
	Traceability	✓	✓
	Transparency	✓	✓
Sustainability & Social Well-being	Business, Economy, and Labour	✓	✓
	Environment and Ecosystem	✓	✓
	Mitigation and Protection Actions	✓	✓
	Proportionality		✓
	Socio-economic Implications	✓	✓
	Sociocultural Implications	✓	✓
	Sustainability	✓	✓

	Data Governance		✓
	Data Privacy and Protection	✓	✓
	Impact/Risk Investigation, Assessment, and Prevention	✓	✓
Security & Safety	Markets and Consumer Protection		✓
	Robustness	✓	✓
	Safety	✓	✓
	Security	✓	✓
	Trialability	✓	✓
	Accessibility	✓	✓
	Algorithms Biases	✓	✓
	Diversity	✓	✓
	Fairness	✓	✓
Inclusivity	Gender Equality	✓	✓
	Inclusiveness	✓	✓
	Multi and Interdisciplinarity	✓	✓
	Non-discrimination and Equality	✓	✓
	Plurality	✓	✓
	Cultural Preservation		✓
	Freedoms	✓	✓
	Human Autonomy	✓	✓
	Human Dignity	✓	✓
Human-Centric	Human Health and Well-being	✓	✓
	Human Learning/Training	✓	✓
	Human Rights	✓	✓
	Human-Centric/Centered	✓	✓
	Human Oversight		✓
	Guidelines, Frameworks, and Policies	✓	✓
Policy Framework & Standards	Interoperability and Recognition	✓	✓
	International Alignment	✓	✓
	Rule of Law	✓	✓

Reference: Compiled by the author.

Nonetheless, to varying degrees, the coding underscores a convergence toward similar discussions. The categorization of codes into clusters offers insight into the underlying motives driving these recommendations. For instance, the Governance Pillars and Sustainability & Social Well-being clusters closely correlate with the current geopolitical implications of AI. Addressing these is crucial to prevent more profound issues that could potentially escalate into conflicts. These implications encompass managing AI's potential as a power amplifier, the engagement of new geopolitical actors, and the risk posed by deep fakes, capable of polarizing national and international arenas. Moreover, the Transparency & Accountability and Security & Safety clusters focus on necessities aimed at averting AI misuse, which could exacerbate associated problems. Furthermore, the Inclusivity & Human-Centric clusters are particularly centered on recommendations impacting humanity at large, directly affected by any AI-related issues. Lastly, the Policy Framework & Standards cluster directly aligns with the Recommendations' core intentions: establishing guidelines for global consideration.

Beyond its only original purpose of observing the alignment between OECD and UNESCO's recommendations, the categorization of these codes reinforces the pressing demand for regulating AI's usage. This imperative stems from AI's pervasive involvement across diverse domains, signaling the wide-ranging implications it carries. However, it becomes evident that this analytical method alone cannot adequately showcase the distinct approaches of each organization. Hence, delving into the nuances – how each assigns responsibilities, establishes priorities, and engages in discourse – alongside an exploration of each organization's characteristics will be instrumental in addressing the research question.

In concluding this section, it is essential to emphasize the nuanced disparities within the documents. These variations serve as a reflection of the internal dynamics and prevailing values within each organization, as explored in earlier subsections. Moreover, within the wider context of influences and ideals pursued by both entities, it becomes essential to scrutinize how their frameworks are shaped by internal dynamics and broader external contexts. Both organizations are now entrenched in a new form of multilateralism, tasked with navigating a landscape where the former hegemon still operates predominantly in a unilateral approach, and emerging actors, notably China, envision the potential to espouse values differing from those foundational to the OECD and UNESCO. Thus, analyzing the Recommendations on AI from these entities demands consideration not solely of the recommendations' content, but also of the intrinsic characteristics, values, and dynamics unique to each organization. Besides, equally significant is examining their broader geopolitical and historical contexts. These interconnections will be expounded upon in the following section, designated as a summary of findings.

6. SUMMARY OF FINDINGS

The primary aim of this study was to examine how the increasing emergence of AI and its implications are being managed by two specific international actors, the OECD and UNESCO. These organizations operate within the current neoliberal world order, where multilateralism is an active force facing a legitimacy crisis. The selection of these international organizations was driven by their pioneering efforts in producing multilateral documents for regulating AI.

To fulfill the paper's objective, the study encompassed explanations and contextualizations about AI, its escalating attention on the international stage, an analysis of its geopolitical implications, and an exploration of the characteristics of both the OECD and UNESCO. The results that can be drawn from each of these researches are presented below.

Firstly, AI is currently in a phase described as weak and narrow in comparison to the potential it theoretically holds. However, the risks and implications stemming from this initial stage have prompted discussions within the international community regarding the need for guidelines and regulations to manage its impact. The future trajectory of AI's development remains uncertain, even for experts engaged in its advancement. Regardless of how far AI will evolve, its rapid pace of improvement underscores the paramount importance of proactive policymaking and collaboration among various stakeholders, including governments, researchers, and private entities, to prevent humanity from encountering greater consequences due to falling behind in addressing AI's pitfalls.

Secondly, the substantial investments and heightened regulatory discussions surrounding AI demonstrate the rising attention it is gathering in the international arena. Furthermore, it also shows the integration of the technology within the debates of IR due to AI's potential to enhance a country's power dynamics. As a result, this technology has emerged as a pivotal source of international competitiveness, profoundly influencing various facets of global affairs, including the realms of economy, politics, law, and culture.

Moreover, AI's geopolitical implications are far-reaching, spanning various domains. It serves as a political power amplifier, intensifies the competition between the US and China, propels global endeavors to secure comparative advantages, witnesses the growing influence of the private sector as a geopolitical actor, poses threats through deep fakes, which heighten polarizations nationally and internationally, and fuels the emergence of AI nationalism, collectively reshaping the global scenario. Hence, this complex situation underscores the necessity for IOs to establish widely acknowledged guidelines, addressing AI's multifaceted impact to the greatest extent possible.

When examining AI's implications through the prisms of realism and liberalism within the realm of IR, there emerges a dynamic interplay of perspectives. These frameworks offer insights that both enlighten and fall short in capturing the multifaceted implications of AI. Realism accentuates AI's centrality in power dynamics and security concerns, aligning with its strategic utility while grappling with challenges stemming from AI's reliance on the private sector, revealing a gap in its state-centric approach. Conversely, liberalism sheds light on AI's economic potential and prospects for international cooperation but faces contention over AI's role in democratic peace and the complex interplay between government-corporate relationships and traditional market dynamics. The conclusion drawn from this analysis underscores the importance of evaluating AI's impact through established IR theories, while also acknowledging their inadequacies in fully comprehending the intricate nuances of AI's influence. This recognition prompts the need for establishing a theoretical framework that can better encompass a new domain like AI.

Consequently, this paper employs multilateralism and the world order as foundational frameworks to present an analysis centered around the OECD's and UNESCO's Recommendations on AI in the 2020s. Through this study, distinct phases of the world order have been delineated, illuminating how multilateralism has been intricately shaped within these periods. By the 2020s, the world order is entrenched within a capitalist neoliberal landscape, a paradigm facilitated by the proliferation of North American ideals within a unipolar structure following the Cold War. However, the continuous unilateral approach of the US, the advent of new actors, both public and private, and the Covid-19 pandemic have precipitated a crisis within Western-centric multilateralism, characterized by its alignment with liberal and Western values. Consequently, a new form of multilateralism has emerged to accommodate and address these evolving dynamics now present in a multipolar order.

In this context, the integration of AI into the evolving landscape of this new multilateralism within the neoliberal global order becomes increasingly evident. This alignment is particularly noteworthy as AI's emergence coincides with a period when Western international institutions are grappling with a legitimacy crisis. The outcomes of this integration not only mirror a transformative process but also signify a fundamental redefinition of the functions and dynamics within contemporary multilateral organizations. The emergence of a new multilateralism not only reflects these shifts but also denotes a significant reconfiguration of roles and interactions within multilateral structures.

Consequently, these IOs face notable difficulties, especially in light of escalating geopolitical tensions, the ascent of new global actors, and technological advancements, prominently illustrated by the rise of AI.

The transition from the former multilateralism, primarily anchored in Western values, to a new multilateral approach shows the need to the integrate and reconcile diverse and often conflicting perspectives. This intricate landscape poses significant challenges for international cooperation, necessitating navigation through a diverse array of distinct visions and interests among various actors. Managing this complexity extends beyond the interaction among conventional states and encompasses the expanding influence of non-state entities like major corporations and non-governmental organizations. The emergence of AI and other disruptive technologies amplifies these challenges and opportunities. The new multilateralism must confront ethical, regulatory, and security concerns intertwined with the global use and governance of AI by embracing broader inclusivity beyond Western values to incorporate Eastern states and diverse perspectives. Addressing these pivotal issues mandates a collaborative and coordinated approach among multiple stakeholders, acknowledging the continuously evolving structures and dynamics that historically shaped the global order. In this evolving landscape shaped by AI, complexities arise that challenge the foundational principles of multilateralism articulated by scholars like Keohane, Ruggie, and Cox.

The theoretical framework provided not only delineates the diverse perspectives on multilateralism but also serves as a lens to understand the complexities of global governance amidst contemporary challenges. Understanding multilateralism’s historical trajectory across distinct periods, from post-Second World War to the present, offers crucial context. These periods illustrate the changing dynamics of power, world orders, and the challenges faced by multilateralism as it navigates a lack of legitimacy. Keohane’s minimalist view, focusing on state-centric coordination, provides a foundational understanding, while Ruggie’s expansion to include qualitative principles highlights the potential limitations of quantitative definitions in complex decision-making scenarios. However, it is Cox’s comprehensive view that incorporates various actors – public, private, and civil society – mediated by states and international bodies, which resonates profoundly within the context of contemporary challenges like AI, requiring a broader spectrum of engagement. This understanding is crucial for organizations like the OECD and UNESCO to navigate the intricate landscape of AI regulation while acknowledging the multiplicity of stakeholders and the evolving dynamics of global power structures.

Nonetheless, it is crucial to note that both the OECD and UNESCO find themselves entangled in the legitimacy crisis of Western multilateral IOs. The research conducted reveals that, while the OECD remains inherently linked to neoliberalism, steadfast in preserving the established order, UNESCO might have more flexibility to deviate from the influences of major powers, as it previously did in its history. This contrast is notably reflected in the OECD's focus on establishing a trustworthy AI market and UNESCO's emphasis on the ethical aspects of AI. Expanding upon the outcomes of this paper, the distinctive traits of these organizations highlight how their individual nature, functioning, and historical backgrounds shape the disparities in their recommendations.

Observations from this research reveal that while the OECD and UNESCO share final goals, their approaches to AI differ significantly. This distinction becomes evident when examining their respective documents and considering the history and characteristics of each organization. While their documents cover similar topics and themes, a deeper analysis highlights nuanced differences in how they assign responsibilities, set priorities, and engage in discourse – an aspect not fully captured by the thematic analytical method, demanding insights derived from a more thorough and comprehensive reading.

As previously mentioned, the OECD includes private actors as directly responsible for certain AI regulations, dedicating an entire section to outline their obligations. The subsequent section focuses on government responsibilities, primarily concerning internal policymaking and international cooperation. On the contrary, while acknowledging the involvement of multiple stakeholders in AI, UNESCO opts for an approach primarily centered on the government agency. For example, when referencing a guideline to be adopted by entities beyond the state, policymakers are emphasized as the key drivers responsible for encouraging and mandating compliance from other stakeholders.

These different focuses can be attributed to the distinct purposes of each organization. The OECD, established to bolster the liberal market economy model, will naturally lean toward actively engaging both the private sector and governments to foster collaboration for economic growth and development. In contrast, UNESCO's purpose aligns more closely with social concerns owing to its affiliation with the UN, which endeavors to promote peace and security through intergovernmental cooperation. Besides, the organization's emphasis on engaging with governments and guiding states in shaping legislation, policies, or other AI-related instruments aligns not only with its objective but also with its internal dynamics. The predominant influence of Anglo-North American ideals has steered UNESCO towards adopting an intergovernmental framework, granting states significant control over its decisions and policies.

Subsequently, the OECD and UNESCO approach AI differently in terms of their ultimate priorities. While the former prioritizes ensuring responsible stewardship and trustworthy AI, the latter emphasizes the ethical use of AI. These distinctions once again stem from their institutional backgrounds and goals. The OECD will, thereby, focus on a sustainable AI market, which is reliant on a trustworthy innovation, while UNESCO will stress the need for AI to adhere to ethical principles for its broader benefit, spanning from individuals to ecosystems.

Finally, the distinct approaches of the OECD and UNESCO are reflected not just in their focus and priorities but also in the discourse within their documents and structural styles. The OECD's emphasis on a market-oriented stance leads to a direct and objective approach in its documentation. This could be attributed to its more selective membership, where Members adhere to specific requirements, fostering alignment on various topics and enabling more direct methods for document production. Conversely, UNESCO's socially oriented discourse stems from its role as a multilateral organization striving for universality. Due to the diverse cultures, developmental levels, and social contexts among its Members, UNESCO's documents might tend to necessitate a more detailed social approach to address the diverse situations and needs encountered within its membership.

In summary, delving into these nuances not only sheds light on the content of their documents but also underscores the underlying purposes guiding these institutions in addressing global challenges. The attempt to handle AI's broad geopolitical implications is manifested through the establishment of guidelines meant to steer the involved actors. However, the OECD intertwines the necessity for regulation with market-oriented approaches, priorities, and discourses, whereas UNESCO aligns it with socially-oriented perspectives. Overall, the hypothesis introduced in this paper appears to hold true.

Therefore, this research lays an initial foundation for future investigations into this topic. Further in-depth analysis and examination are essential to expand upon the observations made here. By doing so, it could gain deeper insights into the effectiveness of different organizational approaches amidst the challenges faced by multilateralism, especially within a recent thematic area such as Artificial Intelligence.

7. FINAL CONSIDERATIONS

This study was driven by the need to discuss in more depth the connection between AI and international cooperation within IR. In order to achieve the aim proposed and the research question defined, this paper was encompassed by contextual, historical, and conceptual explanations.

This research examined how the OECD and UNESCO address the implications of AI within the evolving landscape of multilateralism and the world order. It tracked different phases of world order and multilateralism, recognizing the emergence of a new multilateral dynamic amid a legitimacy crisis in Western-centric institutions. Employing this theoretical framework, the study underscored the need for inclusive approaches to address AI's multifaceted impact. By highlighting disparities between the OECD and UNESCO, the research identified distinct approaches influenced by three key nuances: how they allocate responsibilities, establish priorities, and engage in discourse. These differences underscore the OECD's market-oriented approach in contrast to UNESCO's socially-oriented perspectives, aligning with the initial hypothesis. This divergence emphasized how these institutions' origins and objectives shape their priorities, preferences, and strategies in navigating global AI challenges.

Acknowledging the extensive effort dedicated to this undergraduate study, it is essential to recognize inherent limitations. Primarily, this paper's scope is limited to the analysis of only two reports produced in the international level, resulting in a degree of both generalization and specificity within the attained results. Consequently, future research should encompass a broader array of documents that propose policies and guidelines for AI, thereby expanding the scope for a more comprehensive analysis.

Furthermore, time presented another limitation as this research was carried out within an academic semester and involved tools such as the MAXQDA software, which also imposed constraints on its usage. Consequently, the depth of analysis and exploration was restricted. An alternative method that would offer more valuable and comprehensive insights would be conducting interviews with professionals employed within these institutions who have experienced the production of these documents. This could lead to a deeper understanding of the intricate influences that underlie their operations within the IOs. However, the difficulty of contacting representatives of these organizations in a short time period hindered this possibility.

Despite these challenges and limitations recognized through an epistemological and reflexive process, this study stands as stepping stone for future research, as it was able to synthesize fundamental non-technical insights into the connection between AI and IR. However, it is crucial to note that the elements explored in this research do not decisively confirm which set of recommendations, either from the OECD or UNESCO, is more likely to garner support from stakeholders amidst the crisis of multilateralism. Therefore, validating this assertion would require a more extensive investigation, setting the stage for a potential future research endeavor.

Considering the novelty and unpredictability of AI, this study illuminates the need for further exploration into emerging facets. After all, the landscape of AI in IR will continue to evolve, and humanities studies will be imperative for addressing the impacts and dynamics that may represent global challenges.

8. REFERENCES

ADIPUTERA, Yunizar. Evaluating the Normative and Structural Explanations of Democratic Peace Theory. *Indonesian Journal of International Studies (IJIS)*, vol. 1, no. 1, p. 21-30, June 2014. Available at: https://doi.org/10.22146/globalsouth.28817. Accessed in: December 18, 2023;

ALMEIDA, Celia; CAMPOS, Rodrigo Pires de. Multilateralismo, ordem mundial e Covid-19: questões atuais e desafios futuros para a OMS. *Saúde em Debate*, Rio de Janeiro, vol. 44, no. especial 4, p. 13-39, December 2020. Available at: https://revista.saudeemdebate.org.br/sed/article/view/4511. Accessed in: November 11, 2023;

BENDETT, Samuel. Roles and Implications of AI in the Russian-Ukrainian Conflict. Center for a New American Security (CNAS), July 20, 2023. Available at: https://www.cnas.org/publications/commentary/roles-and-implications-of-ai-in-the-russian-ukrainian-conflict. Accessed in: November 5, 2023;

BLACK, Jeremy. *Great Powers and the Quest for Hegemony*: The World Order since 1500. 1. ed. Routledge, 2008;

BRANNEN, Kate; FLEMING-DRESSER, Julia; MCANANY, Molly. How AI could upend geopolitics: A Conversation With Ian Bremmer and Mustafa Suleyman. *Foreign Affairs*, September 7, 2023. Available at: https://www.foreignaffairs.com/podcasts/how-ai-could-upend-geopolitics-ian-bremmer-mustafa-suleyman . Accessed in: September 28, 2023;

CAMPOS, Rodrigo Pires; LIMA, J. B. B; LOPES, L. L. A. Os fóruns de alto nível da Organização para Cooperação e o Desenvolvimento Econômico (OCDE): limites e perspectivas da posição brasileira na agenda sobre efetividade da ajuda internacional. *Boletim de Economia e Política Internacional*, IPEA, no. 8, p. 27-40, October/December 2011. Available at: https://repositorio.ipea.gov.br/bitstream/11058/4040/1/BEPI_n08_foruns.pdf. Accessed in: September 29, 2023;

CESAREO, Serena; WHITE, Joseph. The Global AI Index. *Tortoise*, June 2023. Available at: https://www.tortoisemedia.com/intelligence/global-ai/. Accessed in: October 15, 2023;

COLLINS, Ben; ZADROZNY, Brandy. How a fake persona laid the groundwork for a Hunter Biden conspiracy deluge. *NBC News*, October 29, 2020. Available at: https://www.nbcnews.com/tech/security/how-fake-persona-laid-groundwork-hunter-biden-conspiracy-deluge-n1245387. Accessed in: November 5, 2023;

COX, Robert W. Multilateralism and World Order. *Review of International Studies*, vol. 18, no. 2, p. 161-180, April 1992. Available at: http://www.jstor.org/stable/20097291. Accessed in: November 30, 2023;

COPELAND, B. J.. Alan Turing: British mathematician and logician. *Britannica*, December 3, 2023. Available at: https://www.britannica.com/biography/Alan-Turing. Accessed in: Oct. 26, 2023;

DAHL, Robert A. The concept of power. *Behavioral Science*, vol. 2, no. 3, p. 201-215, 1957. Available at: https://doi.org/10.1002/bs.3830020303. Accessed in: December 18, 2023;

ECCLESTON, Richard. The OECD and global economic governance. *Australian Journal of International Affairs*, vol. 65, no. 2, p. 243-255, April 2011. Available at: https://www.tandfonline.com/doi/abs/10.1080/10357718.2011.550106. Accessed in: December 20, 2023;

ELZINGA, Aant. *A Unesco e a política de cooperação international no campo da ciência*. In: Maio, Marcos Chor. Ciência, política e relações internacionais: ensaios sobre Paulo Carneiro. 1. ed. Rio de Janeiro: Fiocruz e Unesco, 2004;

EUROPEAN COMMISSION. *Coordinated Plan on Artificial Intelligence 2021 Review*. April 21, 2021. Available at: https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review. Accessed in: December 18, 2023;

EUROPEAN COMMISSION. *OECD - Organisation for Economic Co-operation and Development*. November 4, 2022. Available at: https://knowledge4policy.ec.europa.eu/organisation/oecd-organisation-economic-co-operation-development_en. Accessed in: September 27, 2023;

EEAS. *Relations with OECD and UNESCO*. September 7, 2021. Available at: https://www.eeas.europa.eu/paris-oecd-unesco/relations-oecd-and-unesco_en?s=64. Accessed in: October 27, 2023;

FITCH, Asa; IP, Greg. Chips Are the New Oil and America Is Spending Billions to Safeguard Its Supply. *The Wall Street Journal*, January 14, 2023. Available at: https://www.wsj.com/articles/chips-semiconductors-manufacturing-china-taiwan-11673650917. Accessed in: November 5, 2023;

FLONK, Daniëlle. Emerging illiberal norms: Russia and China as promoters of internet content control . *International Affairs*, vol. 97, no. 6, p. 1925-1944, November 2021. Available at: https://doi.org/10.1093/ia/iiab146. Accessed in: November 8, 2023;

FRANKE, Ulrike. Artificial Intelligence Diplomacy: Artificial Intelligence Governance as a New European Union External Policy Tool. *Policy Department for Economic, Scientific and Quality of Life Policies*, European Union, p. 3-52, June 2021. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662926/IPOL_STU(2021)662926_EN.pdf. Accessed in: November 5, 2023;

FUNDAÇÃO ALEXANDRE DE GUSMÃO. Memórias do Brasil na OMC: Entrevista - Celso de Tarso Pereira. *Youtube*, November 17, 2022. Available at: https://www.youtube.com/watch?v=cipF8kECBxY. Accessed in: September 27, 2023;

FUNDAÇÃO ALEXANDRE DE GUSMÃO. Memórias do Brasil na OMC: Entrevista - Nilo Ditz Filho. *Youtube*, November 17, 2022. Available at: https://www.youtube.com/watch?v=Jsd77eia3zc. Accessed in: September 27, 2023;

G20 BRASIL 2024. *Perguntas Frequentes*. 2023. Available at: https://www.g20.org/pt-br/sobre-o-g20/faq. Accessed in: December 23, 2023;

GONSALVES, Tad. *Artificial Intelligence*: A Non-Technical Introduction. 1. ed. Japan: Sophia University Press, 2017;

GRANADOS, Oscar M.; PEÑA, Nicolas De la. Artificial Intelligence and International System Structure. *Revista Brasileira de Política Internacional*, vol. 64, no. 1, March 2021. Available at: https://doi.org/10.1590/0034-7329202100103. Accessed in: December 18, 2023;

IPEA. *Conheça os BRICS*: Brasil, Rússia, Índia, China e África do Sul. 2014. Available at: https://www.ipea.gov.br/forumbrics/pt-BR/conheca-os-brics.html. Accessed in: December 23, 2023;

JOHNSON, James. Artificial intelligence & future warfare: implications for international security. *Defense & Security Analysis*, vol. 35, no. 2, p. 147-169, April 2019. Available at: https://doi.org/10.1080/14751798.2019.1600800. Accessed in: November 1, 2023;

KAPETAS, Anastasia. The geopolitics of artificial intelligence. *The Strategist*, December 24, 2020. Available at: https://www.aspistrategist.org.au/the-geopolitics-of-artificial-intelligence/. Accessed in: October. 5, 2023;

KEOHANE, Robert O. Multilateralism: An Agenda for Research. *International Journal*, vol. 45, no. 4, p. 731-764, 1990. Available at: https://www.jstor.org/stable/40202705. Accessed in: December 5, 2023;

KRASNER, Stephen D. Structural Causes and Regime Consequences: Regimes as Intervening Variables. *International Organization*, vol. 36, no. 2, p. 185-205, 1982. Available at: http://www.jstor.org/stable/2706520. Accessed in: December 18, 2023;

LAÏDI, Zaki. Towards a post-hegemonic world: The multipolar threat to the multilateral order. *International Politics*, vol. 51, no. 3, p. 350-365, 2014. Available at: https://sciencespo.hal.science/hal-03460497/. Accessed in: December 18, 2023;

LAZARD. *Geopolitics of Artificial Intelligence*. October 2023. Available at: https://lazard.com/media/cyenforc/lazard-geopolitical-advisory_geopolitics-of-artificial-intelligence_-oct-2023.pdf. Accessed in: November 5, 2023;

LAZAROU, Elena *et al*. The Evolving 'Doctrine' of Multilateralism in the 21st Century. *Mercury*, e-paper no. 3, p. 1-34, February 2010. Available at: https://repositorio.fgv.br/server/api/core/bitstreams/e0773c21-a707-48a8-a020-2cfffc7e8836/content. Accessed in: December 20, 2023;

LIMA, Maria Regina Soares de; ALBUQUERQUE, Marianna. Global Reorganization and the Crisis of Multilateralism. *CEBRI*, Rio de Janeiro, p. 4-20, December 2020. Available at: https://www.cebri.org/en/doc/113/global-reorganization-and-the-crisis-of-multilateralism. Accessed in: November 30, 2023;

LIMA, Maria Regina Soares de; ALBUQUERQUE, Marianna. Instituições Multilaterais e Governança Global: Cenários de Reorganização das Estruturas de Governança Global e Perspectivas do Multilateralismo nas Próximas Décadas. *Fundação Oswaldo Cruz*, Rio de Janeiro, p. 7-22, 2021. Available at: https://saudeamanha.fiocruz.br/wp-content/uploads/2021/05/LIMA-MRS-e-ALBUQUERQUE-M-2021-Institui%C3%A7%C3%B5es-Multilaterais-Governan%C3%A7a-Global-Fiocruz-Saude-Amanha-TD059.pdf. Accessed in: November 10, 2023;

LUTKEVICH, Ben. *What is an expert system?*. *TechTarget*. Available at: https://www.techtarget.com/searchenterpriseai/definition/expert-system. Accessed in: October 17, 2023;

MARR, Bernard. The Most Significant AI Milestones So Far. *Bernard Marr & Co.*, 2021. Available at: https://bernardmarr.com/the-most-significant-ai-milestones-so-far/. Accessed in: October 10, 2023;

MASLEJ, Nestor *et al*. The AI Index 2023 Annual Report. AI Index Steering Committee, *Institute for Human-Centered AI*, Stanford University, April 2023. Available at: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf. Accessed in: September 29, 2023;

MCBRIDE, Keegan. AI, Geopolitics, Regulation, and Digital Innovation: What is actually going on right now?. *Oxford Internet Institute*, June 13, 2023. Available at: https://www.oii.ox.ac.uk/news-events/ai-geopolitics-regulation-and-digital-innovation-what-is-actually-going-on-right-now/. Accessed in: November 8, 2023;

MENEZES, E. O. D; FILHO, P. L. M. Análise De Conteúdo: Contextualização, Operacionalização, Discussões E Perspectivas. *Revista Valore*, vol. 7, e-7047, 2022. Available at: https://revistavalore.emnuvens.com.br/valore/article/view/1043. Accessed in: September 29, 2023;

MIAILHE, Nicolas. The Geopolitics of Artificial Intelligence: The return of Empires?. *Politique Étrangère*, vol. 83, no. 3, p. 105-117, 2018. Available at: https://www.cairn-int.info/load_pdf.php?ID_ARTICLE=E_PE_183_0105&download=1. Accessed in: November 5, 2023;

MURPHY, Margi. Emotive Deepfakes in Israel-Hamas War Further Cloud What's Real. *Bloomberg*, October 31, 2023. Available at: https://www.msn.com/en-us/money/other/emotive-deepfakes-in-israel-hamas-war-further-cloud-what-s-real/ar-AA1jatSW. Accessed in: November 16, 2023;

MEARSHEIMER, John J. *The tragedy of great power politics*. New York: WW Norton & Company, 2001;

NDZENDZE, Bhaso; MARWALA, Tshilidzi. *Artificial Intelligence and International Relations Theories*. Singapore: Palgrave Macmillan US, 2023;

NEWMAN, Edward; THAKUR, Ramesh; TIRMAN, John. *Multilateralism under challenge? Power, international order, and structural change*. Tokyo: United Nations University Press, 2006;

NGUYEN, Andy *et al*. Ethical principles for artificial intelligence in education. *Education and Information Technologies*, vol. 28, p. 4221-4241, February 2022. Available at: https://doi.org/10.1007/s10639-022-11316-w. Accessed in: September 28, 2023;

OECD. Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research. *OECD Publishing*, Paris, June 2023. Available at: https://doi.org/10.1787/a8d820bd-en. Accessed in: October 15, 2023;

OECD. *Convention on the OECD*. Available at: https://www.oecd.org/about/document/oecd-convention.htm. Accessed in: September 27, 2023;

OECD. *Discover the OECD*: Better Policies for Better Lives. Available at: https://www.oecd.org/general/Key-information-about-the-OECD.pdf. Accessed in: September 27, 2023;

OECD. *Recommendation of the Council on Artificial Intelligence*. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Accessed in: September 20, 2023;

OPERA MUNDI. Inteligência Artificial: Tudo O Que Você Precisa Saber - Miguel Nicolelis - Programa 20 Minutos. *Youtube*, June 6, 2023. Available at: https://www.youtube.com/watch?v=pb4b4_MlNwo. Accessed in: November 6, 2023;

PAN, T Nang Seng. What Will AI Mean for ASEAN?. *New America*, September 21, 2023. Available at: https://www.newamerica.org/planetary-politics/blog/what-will-ai-mean-for-asean/#:~:text=Among%20ASEAN's%2010%20members%2C%20only,of%20concrete%20strategy%20or%20regulation. Accessed in: November 6, 2023;

PEDROSO, João; CAPELLER, Wanda; SANTOS, Andreia. Os Efeitos Perversos da Inteligência Artificial: a democracia, o estado de direito e a distribuição de desigualdades e poder no mundo. *Confluências*, vol. 25, no. 3, p. 230-253, August-December 2023. Available at: https://periodicos.uff.br/confluencias/article/view/60052/35341. Accessed in: December 18, 2023;

PINTO, Denis Fontes de Souza. *OCDE*: uma visão brasileira. 1. ed. Brasília: IRBr; FUNAG, 2000;

RUGGIE, John Gerard. Multilateralism: the anatomy of an institution. *International Organization*, vol. 46, no. 3, p. 561-598, 1992. Available at: https://www.cambridge.org/core/journals/internationalorganization/article/multilateralism-the-anatomy-of-aninstitution/AB34548F299B16FDF0263E621905E3B5. Accessed in: December 5, 2023;

RUSSEL, Stuart J.; NORVIG, Peter. *Artificial Intelligence*: A Modern Approach. 4. ed. Pearson Series, 2020;

SENADO FEDERAL. *Manual de Comunicação da Secom - G7*. Available at: https://www12.senado.leg.br/manualdecomunicacao/guia-de-economia/g7-e-g8. Accessed in: December 23, 2023;

SLAUGHTER, Anne-Marie. *The chessboard and the web*: Strategies of connection in a networked world. New Haven, Connecticut: Yale University Press, 2017;

STATE COUNCIL. T*he State Council issued the Notice of the Development Plan for the New Generation of Artificial Intelligence*. July 20, 2017. Available at: https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm. Accessed in: December 20, 2023;

STERLING, Toby. Another ASML tool hit by US export curbs, China at 46% of sales. *REUTERS*, October 19, 2023. Available at: https://www.reuters.com/technology/asml-ceo-says-he-expects-demand-china-will-remain-strong-2023-10-18/. Accessed in: November 9, 2023;

THILLIEN, Dexter *et al*. Seizing the opportunity: the future of AI in Latin America. *The Economist Group*, 2022. Available at: https://impact.economist.com/perspectives/sites/default/files/seizing-the-opportunity-the-future-of-ai-in-latin-america.pdf. Accessed in: November 7, 2023;

THORMUNDSSON, Bergur. Amount of artificial intelligence (AI) companies in major economies worldwide in 2023. *Statista*, September 26, 2023. Available at: https://www.statista.com/statistics/1413456/major-economies-ai-companies-worldwide/#:~:text=Number%20of%20AI%20companies%20in%20major%20economies%20worldwide%202023&text=The%20United%20States%20had%20by,considerably%20behind%2C%20with%20only%206%2C000. Accessed in: October 25, 2023;

UNDERSTANDING Artificial Neural Networks. *ForumIAS*, February 8, 2022. Available at: https://forumias.com/blog/understanding-artificial-neural-networks/. Accessed in: October 10, 2023;

UNESCO. *Constitution of the United Nations Educational, Scientific and Cultural Organization*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000382500. Accessed in: September 28, 2023;

UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000380455#:~:text=AI%20actors%20and%20Member%20States,law%2C%20in%20particular%20Member%20States%27. Accessed in: September 16, 2023;

UNESCO. *Recommendations*. Available at: https://www.unesco.org/en/legal-affairs/standard-setting/recommendations. Accessed in: October 25, 2023;

UNITED NATIONS. *UNESCO: United Nations Educational, Scientific and Cultural Organization*. Available at: https://www.un.org/youthenvoy/2013/08/unesco-united-nations-educational-scientific-and-cultural-organization/. Accessed in: September 17, 2023;

VILLASENOR, John. Artificial intelligence and the future of geopolitics. *Brookings Institution*, United States of America, November 2018. Available at: https://policycommons.net/artifacts/4136544/artificial-intelligence-and-the-future-of-geopolitics/4944335/. Accessed in: October 25, 2023;

WALTZ, Kenneth N. *Theory of international politics*. Long Grove, Illinois: Waveland Press, 2010;

WEISS, Charles. How do science and technology affect international affairs?. *Minerva*, vol. 53, no. 4, p. 411-430, November 2015. Available at: https://doi.org/10.1007/s11024-015-9286-1. Accessed in: December 18, 2023;

WHITE HOUSE. *National Security Strategy of the United States of America*. December 2017. Available at: https://trumpwhitehouse.archives.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905.pdf. Accessed in: December 20, 2023;

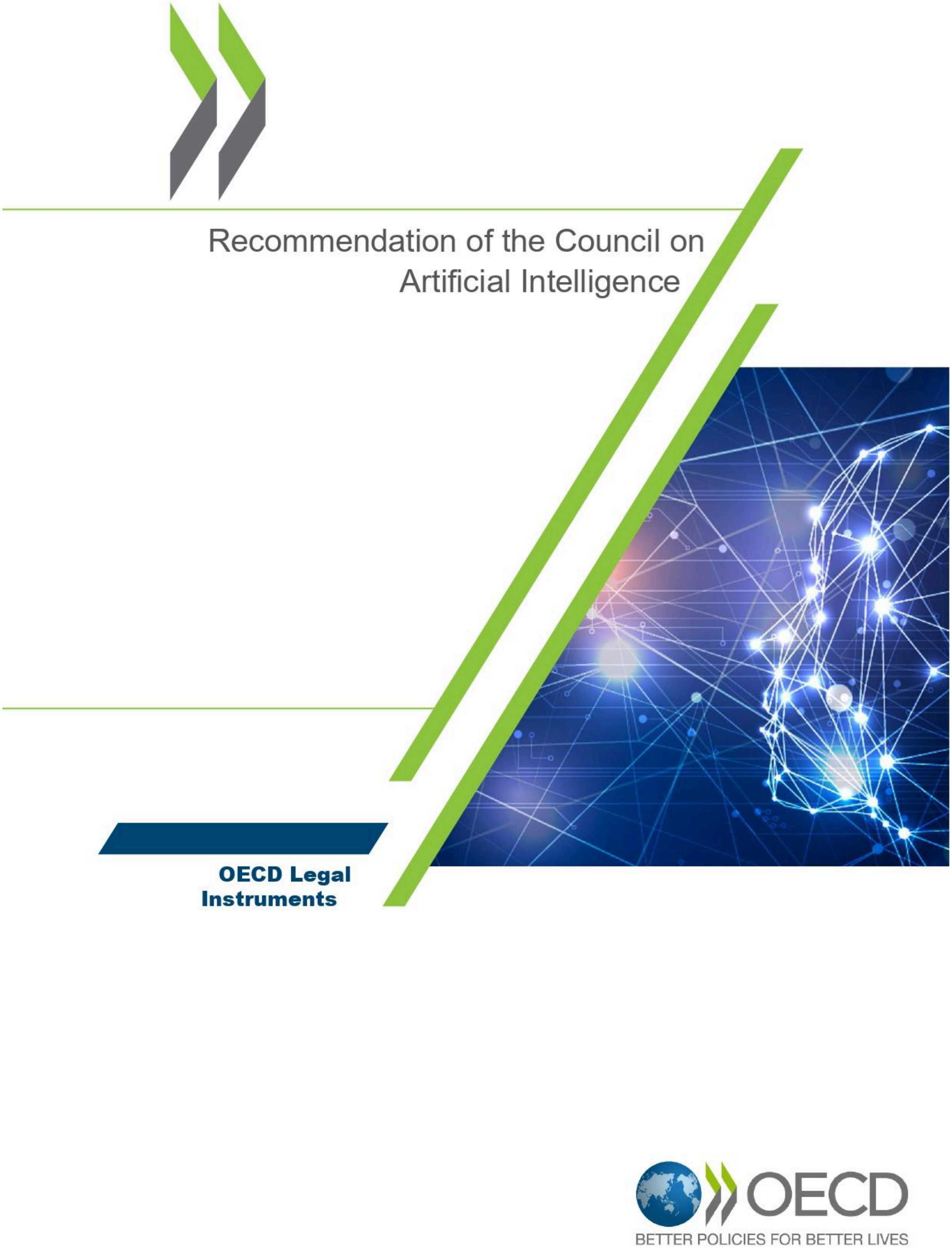
WOHLFORTH, William C. The stability of a Unipolar World. *International Security*, vol. 24, no. 1, p. 5-41, July 1999. Available at: https://doi.org/10.1162/016228899560031. Accessed in: December 18, 2023;

WORLD ECONOMIC FORUM. *Artificial Intelligence*: The Geopolitical Impacts of AI. Available at: https://intelligence.weforum.org/topics/a1Gb0000000pTDREA2/key-issues/a1Gb00000017LCAEA2. Accessed in: November 6, 2023;

YAMASHITA, I. *et al*. Measuring the AI content of government-funded R&D projects: A proof of concept for the OECD Fundstat initiative. *OECD Science, Technology and Industry Working Papers*, Paris, 2021. Available at: https://doi.org/10.1787/7b43b038-en. Accessed in: October 15, 2023.

9. ANNEXES

ANNEX A – OECD’s Recommendation of the Council on Artificial Intelligence



Recommendation of the Council on Artificial Intelligence

OECD Legal Instruments

OECD

BETTER POLICIES FOR BETTER LIVES

This document is published under the responsibility of the Secretary-General of the OECD. It reproduces an OECD Legal Instrument and may contain additional material. The opinions expressed and arguments employed in the additional material do not necessarily reflect the official views of OECD Member countries.

This document, as well as any data and any map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

For access to the official and up-to-date texts of OECD Legal Instruments, as well as other related information, please consult the Compendium of OECD Legal Instruments at http://legalinstruments.oecd.org.

Please cite this document as:
OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449

Series: OECD Legal Instruments

Photo credit: © kras99/Shutterstock.com

Background Information

The Recommendation on Artificial Intelligence (AI) (hereafter the "Recommendation") – the first intergovernmental standard on AI – was adopted by the OECD Council at Ministerial level on 22 May 2019 on the proposal of the Committee on Digital Economy Policy (CDEP). The Recommendation aims to foster innovation and trust in AI by promoting the responsible stewardship of trustworthy AI while ensuring respect for human rights and democratic values. In June 2019, at the Osaka Summit, G20 Leaders welcomed G20 AI Principles, drawn from the Recommendation. The Recommendation was revised by the OECD Council on 8 November 2023 to update its definition of an "AI System", in order to ensure the Recommendation continues to be technically accurate and reflect technological developments, including with respect to generative AI.

The OECD's work on Artificial Intelligence

Artificial Intelligence (AI) is a general-purpose technology that has the potential to: improve the welfare and well-being of people, contribute to positive sustainable global economic activity, increase innovation and productivity, and help respond to key global challenges. It is deployed in many sectors ranging from production, finance and transport to healthcare and security.

Alongside benefits, AI also raises challenges for our societies and economies, notably regarding economic shifts and inequalities, competition, transitions in the labour market, and implications for democracy and human rights.

The OECD has undertaken empirical and policy activities on AI in support of the policy debate over the past two years, starting with a Technology Foresight Forum on AI in 2016 and an international conference on *AI: Intelligent Machines, Smart Policies* in 2017. The Organisation also conducted analytical and measurement work that provides an overview of the AI technical landscape, maps economic and social impacts of AI technologies and their applications, identifies major policy considerations, and describes AI initiatives from governments and other stakeholders at national and international levels.

This work has demonstrated the need to shape a stable policy environment at the international level to foster trust in and adoption of AI in society. Against this background, the OECD Council adopted, on the proposal of CDEP, a Recommendation to promote a human- centric approach to trustworthy AI, that fosters research, preserves economic incentives to innovate, and applies to all stakeholders.

An inclusive and participatory process for developing the Recommendation

The development of the Recommendation was participatory in nature, incorporating input from a broad range of sources throughout the process. In May 2018, the CDEP agreed to form an expert group to scope principles to foster trust in and adoption of AI, with a view to developing a draft Recommendation in the course of 2019. The AI Group of experts at the OECD (AIGO) was subsequently established, comprising over 50 experts from different disciplines and different sectors (government, industry, civil society, trade unions, the technical community and academia) - see http://www.oecd.org/going-digital/ai/oecd-aigo-membership-list.pdf for the full list. Between September 2018 and February 2019 the group held four meetings. The work benefited from the diligence, engagement and substantive contributions of the experts participating in AIGO, as well as from their multi-stakeholder and multidisciplinary backgrounds.

Drawing on the final output document of the AIGO, a draft Recommendation was developed in the CDEP and with the consultation of other relevant OECD bodies and approved in a special meeting on 14-15 March 2019. The OECD Council adopted the Recommendation at its meeting at Ministerial level on 22-23 May 2019.

Scope of the Recommendation

Complementing existing OECD standards already relevant to AI – such as those on privacy and data protection, digital security risk management, and responsible business conduct – the Recommendation focuses on policy issues that are specific to AI and strives to set a standard that is implementable and flexible enough to stand the test of time in a rapidly evolving field. The Recommendation contains five high-level values-based principles and five recommendations for national policies and international co-operation. It also proposes a common understanding of key terms, such as "AI system" and "AI actors", for the purposes of the Recommendation.

More specifically, the Recommendation includes two substantive sections:

1. **Principles for responsible stewardship of trustworthy AI**: the first section sets out five complementary principles relevant to all stakeholders: i) inclusive growth, sustainable development and well-being; ii) human-centred values and fairness; iii) transparency and explainability; iv) robustness, security and safety; and v) accountability. This section further calls on AI actors to promote and implement these principles according to their roles.
2. **National policies and international co-operation for trustworthy AI**: consistent with the five aforementioned principles, this section provides five recommendations to Members and non-Members having adhered to the draft Recommendation (hereafter the "Adherents") to implement in their national policies and international co-operation: i) investing in AI research and development; ii) fostering a digital ecosystem for AI; iii) shaping an enabling policy environment for AI; iv) building human capacity and preparing for labour market transformation; and v) international co-operation for trustworthy AI.

2023 revision to update the definition of an "AI System" and next steps

The Recommendation instructs the CDEP to report to the Council on its implementation, dissemination and continued relevance five years after its adoption and regularly thereafter.

Accordingly, the CDEP, via the AIGO, has begun work towards the preparation of this report to Council. In the context of these discussions, a window of opportunity was identified to maintain the relevance of the Recommendation by updating its definition of an "AI System", and the CDEP approved a draft revised definition in a joint session of the Committee and the AIGO on 16 October 2023. The OECD Council adopted the revised definition of "AI System" at its meeting on 8- November 2023.

The update of the definition included edits aimed at: (i) clarifying the objectives of an AI system (which may be explicit or implicit); (ii) underscoring the role of input which may be provided by humans or machines; (iii) clarifying that the Recommendation applies to generative AI systems, which produce "content"; (iv) substituting the word "real" with "physical" for clarity and alignment with other international processes; (v) reflecting the fact that some AI systems can continue to evolve after their design and deployment.

The CDEP, through AIGO, is now pursuing its work to prepare the report to the Council on the implementation, dissemination and continued relevance of the Recommendation which is expected next year.

For further information please consult: oecd.ai.
Contact information: ai@oecd.org.

Implementation

In addition to reporting to the Council on the implementation of the Recommendation, the CDEP is also instructed to continue its work on AI, building on this Recommendation, and taking into account work in other international fora, such as UNESCO, the European Union, the Council of Europe and the initiative to build an International Panel on AI (see https://pm.gc.ca/eng/news/2018/12/06/mandate-international-panel-artificial- intelligence and https://www.gouvernement.fr/en/france-and-canada-create-new-expert-international- panel-on-artificial-intelligence).

In order to support implementation of the Recommendation, the Council instructed the CDEP to develop practical guidance for implementation, to provide a forum for exchanging information on AI policy and activities, and to foster multi-stakeholder and interdisciplinary dialogue.

To provide an inclusive forum for exchanging information on AI policy and activities, and to foster multi-stakeholder and interdisciplinary dialogue, the OECD launched i) the AI Policy Observatory (OECD.AI) as well as ii) the informal OECD Network of Experts on AI (ONE AI) in February 2020.

OECD.AI is an inclusive hub for public policy on AI that aims to help countries encourage, nurture and monitor the responsible development of trustworthy artificial intelligence systems for the benefit of society. It combines resources from across the OECD with those of partners from all stakeholder groups to provide multidisciplinary, evidence-based policy analysis on AI. The Observatory includes a live database of AI strategies, policies and initiatives that countries and other stakeholders can share and update, enabling the comparison of their key elements in an interactive manner. It is continuously updated with AI metrics, measurements, policies and good practices that lead to further updates in the practical guidance for implementation.

The OECD.AI Network of Experts (ONE AI) is an informal group of AI experts from government, business, academia and civil society that provides AI-specific policy expertise and advice to the OECD. The network provides a space for the international AI community to have in-depth discussions about shared AI policy opportunities and challenges.

THE COUNCIL,

HAVING REGARD to Article 5 b) of the Convention on the Organisation for Economic Co-operation and Development of 14 December 1960;

HAVING REGARD to the OECD Guidelines for Multinational Enterprises [OECD/LEGAL/0144]; Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data [OECD/LEGAL/0188]; Recommendation of the Council concerning Guidelines for Cryptography Policy [OECD/LEGAL/0289]; Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information [OECD/LEGAL/0362]; Recommendation of the Council on Digital Security Risk Management for Economic and Social Prosperity [OECD/LEGAL/0415]; Recommendation of the Council on Consumer Protection in E-commerce [OECD/LEGAL/0422]; Declaration on the Digital Economy: Innovation, Growth and Social Prosperity (Cancún Declaration) [OECD/LEGAL/0426]; Declaration on Strengthening SMEs and Entrepreneurship for Productivity and Inclusive Growth [OECD/LEGAL/0439]; as well as the 2016 Ministerial Statement on Building more Resilient and Inclusive Labour Markets, adopted at the OECD Labour and Employment Ministerial Meeting;

HAVING REGARD to the Sustainable Development Goals set out in the 2030 Agenda for Sustainable Development adopted by the United Nations General Assembly (A/RES/70/1) as well as the 1948 Universal Declaration of Human Rights;

HAVING REGARD to the important work being carried out on artificial intelligence (hereafter, “AI”) in other international governmental and non-governmental fora;

RECOGNISING that AI has pervasive, far-reaching and global implications that are transforming societies, economic sectors and the world of work, and are likely to increasingly do so in the future;

RECOGNISING that AI has the potential to improve the welfare and well-being of people, to contribute to positive sustainable global economic activity, to increase innovation and productivity, and to help respond to key global challenges;

RECOGNISING that, at the same time, these transformations may have disparate effects within, and between societies and economies, notably regarding economic shifts, competition, transitions in the labour market, inequalities, and implications for democracy and human rights, privacy and data protection, and digital security;

RECOGNISING that trust is a key enabler of digital transformation; that, although the nature of future AI applications and their implications may be hard to foresee, the trustworthiness of AI systems is a key factor for the diffusion and adoption of AI; and that a well-informed whole-of-society public debate is necessary for capturing the beneficial potential of the technology, while limiting the risks associated with it;

UNDERLINING that certain existing national and international legal, regulatory and policy frameworks already have relevance to AI, including those related to human rights, consumer and personal data protection, intellectual property rights, responsible business conduct, and competition, while noting that the appropriateness of some frameworks may need to be assessed and new approaches developed;

RECOGNISING that given the rapid development and implementation of AI, there is a need for a stable policy environment that promotes a human-centric approach to trustworthy AI, that fosters research, preserves economic incentives to innovate, and that applies to all stakeholders according to their role and the context;

CONSIDERING that embracing the opportunities offered, and addressing the challenges raised, by AI applications, and empowering stakeholders to engage is essential to fostering adoption of trustworthy AI in society, and to turning AI trustworthiness into a competitive parameter in the global marketplace;

On the proposal of the Committee on Digital Economy Policy:

I. **AGREES** that for the purpose of this Recommendation the following terms should be understood as follows:

- *AI system*: An AI system is a machine-based system 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. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.
- *AI system lifecycle*: AI system lifecycle phases involve: *i)* 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; *ii)* 'verification and validation'; *iii)* 'deployment'; and *iv)* 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.
- *AI knowledge*: AI knowledge refers to the skills and resources, such as data, code, algorithms, models, research, know-how, training programmes, governance, processes and best practices, required to understand and participate in the AI system lifecycle.
- *AI actors*: AI actors are those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI.
- *Stakeholders*: Stakeholders encompass all organisations and individuals involved in, or affected by, AI systems, directly or indirectly. AI actors are a subset of stakeholders.

Section 1: Principles for responsible stewardship of trustworthy AI

II. **RECOMMENDS** that Members and non-Members adhering to this Recommendation (hereafter the "Adherents") promote and implement the following principles for responsible stewardship of trustworthy AI, which are relevant to all stakeholders.

III. **CALLS ON** all AI actors to promote and implement, according to their respective roles, the following Principles for responsible stewardship of trustworthy AI.

IV. **UNDERLINES** that the following principles are complementary and should be considered as a whole.

1.1. Inclusive growth, sustainable development and well-being

Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.

1.2. Human-centred values and fairness

a) AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.

b) To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

1.3. Transparency and explainability

AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:

i. to foster a general understanding of AI systems,

ii. to make stakeholders aware of their interactions with AI systems, including in the workplace,

iii. to enable those affected by an AI system to understand the outcome, and,

iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

1.4. Robustness, security and safety

a) AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.

b) To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.

c) AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.

1.5. Accountability

AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

Section 2: National policies and international co-operation for trustworthy AI

V. RECOMMENDS that Adherents implement the following recommendations, consistent with the principles in section 1, in their national policies and international co-operation, with special attention to small and medium-sized enterprises (SMEs).

2.1. Investing in AI research and development

a) Governments should consider long-term public investment, and encourage private investment, in research and development, including interdisciplinary efforts, to spur innovation in trustworthy AI that focus on challenging technical issues and on AI-related social, legal and ethical implications and policy issues.

b) Governments should also consider public investment and encourage private investment in open datasets that are representative and respect privacy and data protection to support an environment for AI research and development that is free of inappropriate bias and to improve interoperability and use of standards.

2.2. Fostering a digital ecosystem for AI

Governments should foster the development of, and access to, a digital ecosystem for trustworthy AI. Such an ecosystem includes in particular digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate. In this regard, governments should consider promoting mechanisms, such as data trusts, to support the safe, fair, legal and ethical sharing of data.

2.3. Shaping an enabling policy environment for AI

a) Governments should promote a policy environment that supports an agile transition from the research and development stage to the deployment and operation stage for trustworthy AI systems. To this effect, they should consider using experimentation to provide a controlled environment in which AI systems can be tested, and scaled-up, as appropriate.

b) Governments should review and adapt, as appropriate, their policy and regulatory frameworks and assessment mechanisms as they apply to AI systems to encourage innovation and competition for trustworthy AI.

2.4. Building human capacity and preparing for labour market transformation

a) Governments should work closely with stakeholders to prepare for the transformation of the world of work and of society. They should empower people to effectively use and interact with AI systems across the breadth of applications, including by equipping them with the necessary skills.

b) Governments should take steps, including through social dialogue, to ensure a fair transition for workers as AI is deployed, such as through training programmes along the working life, support for those affected by displacement, and access to new opportunities in the labour market.

c) Governments should also work closely with stakeholders to promote the responsible use of AI at work, to enhance the safety of workers and the quality of jobs, to foster entrepreneurship and productivity, and aim to ensure that the benefits from AI are broadly and fairly shared.

2.5. International co-operation for trustworthy AI

a) Governments, including developing countries and with stakeholders, should actively co-operate to advance these principles and to progress on responsible stewardship of trustworthy AI.

b) Governments should work together in the OECD and other global and regional fora to foster the sharing of AI knowledge, as appropriate. They should encourage international, cross-sectoral and open multi-stakeholder initiatives to garner long-term expertise on AI.

c) Governments should promote the development of multi-stakeholder, consensus-driven global technical standards for interoperable and trustworthy AI.

d) Governments should also encourage the development, and their own use, of internationally comparable metrics to measure AI research, development and deployment, and gather the evidence base to assess progress in the implementation of these principles.

VI. **INVITES** the Secretary-General and Adherents to disseminate this Recommendation.

VII. **INVITES** non-Adherents to take due account of, and adhere to, this Recommendation.

VIII. **INSTRUCTS** the Committee on Digital Economy Policy:

a) to continue its important work on artificial intelligence building on this Recommendation and taking into account work in other international fora, and to further develop the measurement framework for evidence-based AI policies;

b) to develop and iterate further practical guidance on the implementation of this Recommendation, and to report to the Council on progress made no later than end December 2019;

c) to provide a forum for exchanging information on AI policy and activities including experience with the implementation of this Recommendation, and to foster multi-stakeholder and interdisciplinary dialogue to promote trust in and adoption of AI; and

d) to monitor, in consultation with other relevant Committees, the implementation of this Recommendation and report thereon to the Council no later than five years following its adoption and regularly thereafter.

About the OECD

The OECD is a unique forum where governments work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to co-ordinate domestic and international policies.

The OECD Member countries are: Australia, Austria, Belgium, Canada, Chile, Colombia, Costa Rica, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Latvia, Lithuania, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Türkiye, the United Kingdom and the United States. The European Union takes part in the work of the OECD.

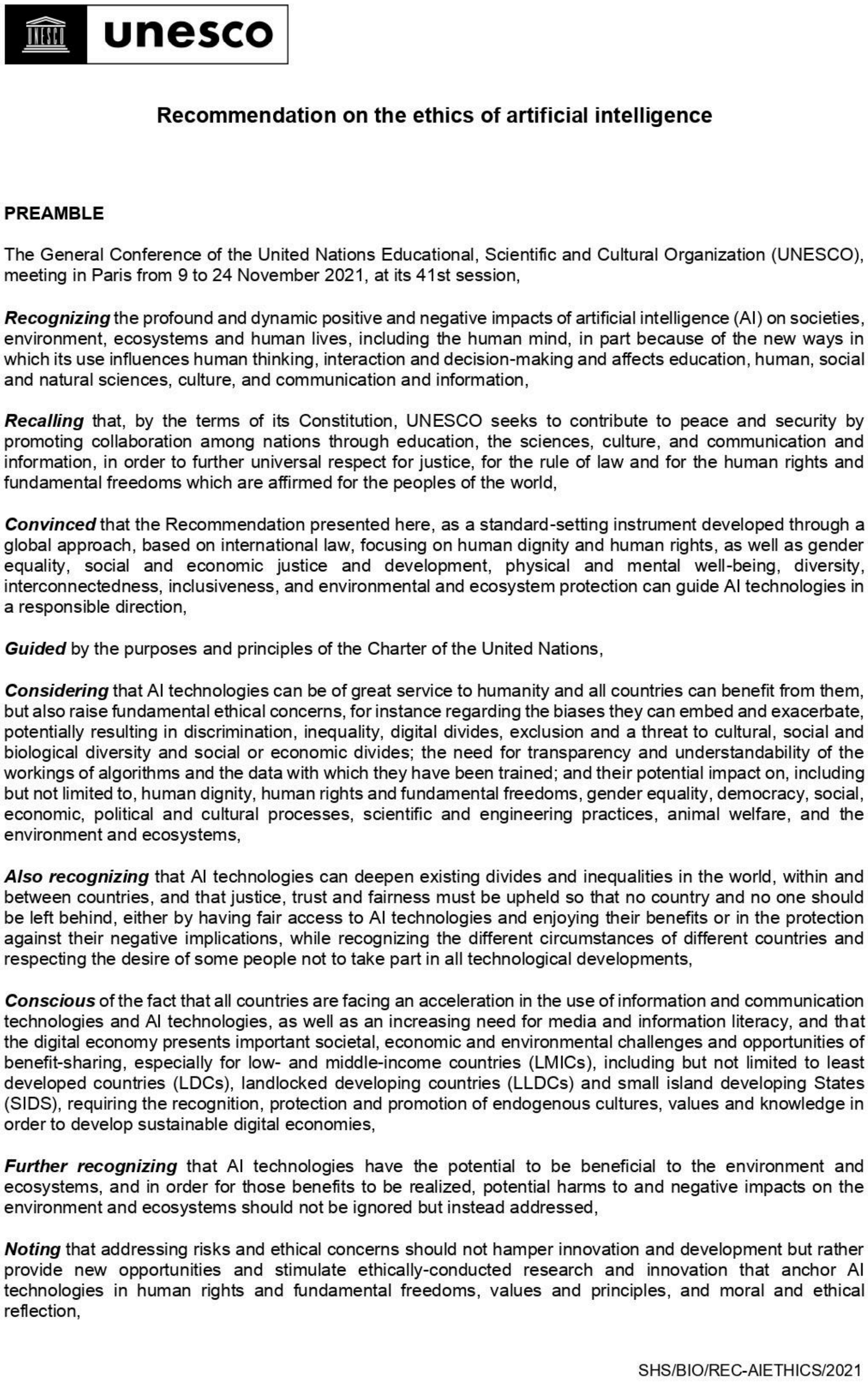
OECD Legal Instruments

Since the creation of the OECD in 1961, around 460 substantive legal instruments have been developed within its framework. These include OECD Acts (i.e. the Decisions and Recommendations adopted by the OECD Council in accordance with the OECD Convention) and other legal instruments developed within the OECD framework (e.g. Declarations, international agreements).

All substantive OECD legal instruments, whether in force or abrogated, are listed in the online Compendium of OECD Legal Instruments. They are presented in five categories:

- **Decisions** are adopted by Council and are legally binding on all Members except those which abstain at the time of adoption. They set out specific rights and obligations and may contain monitoring mechanisms.
- **Recommendations** are adopted by Council and are not legally binding. They represent a political commitment to the principles they contain and entail an expectation that Adherents will do their best to implement them.
- **Substantive Outcome Documents** are adopted by the individual listed Adherents rather than by an OECD body, as the outcome of a ministerial, high-level or other meeting within the framework of the Organisation. They usually set general principles or long-term goals and have a solemn character.
- **International Agreements** are negotiated and concluded within the framework of the Organisation. They are legally binding on the Parties.
- **Arrangement, Understanding and Others**: several other types of substantive legal instruments have been developed within the OECD framework over time, such as the Arrangement on Officially Supported Export Credits, the International Understanding on Maritime Transport Principles and the Development Assistance Committee (DAC) Recommendations.

ANNEX B – UNESCOS's Recommendation on the Ethics of Artificial Intelligence



unesco

Recommendation on the ethics of artificial intelligence

PREAMBLE

The General Conference of the United Nations Educational, Scientific and Cultural Organization (UNESCO), meeting in Paris from 9 to 24 November 2021, at its 41st session,

Recognizing the profound and dynamic positive and negative impacts of artificial intelligence (AI) on societies, environment, ecosystems and human lives, including the human mind, in part because of the new ways in which its use influences human thinking, interaction and decision-making and affects education, human, social and natural sciences, culture, and communication and information,

Recalling that, by the terms of its Constitution, UNESCO seeks to contribute to peace and security by promoting collaboration among nations through education, the sciences, culture, and communication and information, in order to further universal respect for justice, for the rule of law and for the human rights and fundamental freedoms which are affirmed for the peoples of the world,

Convinced that the Recommendation presented here, as a standard-setting instrument developed through a global approach, based on international law, focusing on human dignity and human rights, as well as gender equality, social and economic justice and development, physical and mental well-being, diversity, interconnectedness, inclusiveness, and environmental and ecosystem protection can guide AI technologies in a responsible direction,

Guided by the purposes and principles of the Charter of the United Nations,

Considering that AI technologies can be of great service to humanity and all countries can benefit from them, but also raise fundamental ethical concerns, for instance regarding the biases they can embed and exacerbate, potentially resulting in discrimination, inequality, digital divides, exclusion and a threat to cultural, social and biological diversity and social or economic divides; the need for transparency and understandability of the workings of algorithms and the data with which they have been trained; and their potential impact on, including but not limited to, human dignity, human rights and fundamental freedoms, gender equality, democracy, social, economic, political and cultural processes, scientific and engineering practices, animal welfare, and the environment and ecosystems,

Also recognizing that AI technologies can deepen existing divides and inequalities in the world, within and between countries, and that justice, trust and fairness must be upheld so that no country and no one should be left behind, either by having fair access to AI technologies and enjoying their benefits or in the protection against their negative implications, while recognizing the different circumstances of different countries and respecting the desire of some people not to take part in all technological developments,

Conscious of the fact that all countries are facing an acceleration in the use of information and communication technologies and AI technologies, as well as an increasing need for media and information literacy, and that the digital economy presents important societal, economic and environmental challenges and opportunities of benefit-sharing, especially for low- and middle-income countries (LMICs), including but not limited to least developed countries (LDCs), landlocked developing countries (LLDCs) and small island developing States (SIDS), requiring the recognition, protection and promotion of endogenous cultures, values and knowledge in order to develop sustainable digital economies,

Further recognizing that AI technologies have the potential to be beneficial to the environment and ecosystems, and in order for those benefits to be realized, potential harms to and negative impacts on the environment and ecosystems should not be ignored but instead addressed,

Noting that addressing risks and ethical concerns should not hamper innovation and development but rather provide new opportunities and stimulate ethically-conducted research and innovation that anchor AI technologies in human rights and fundamental freedoms, values and principles, and moral and ethical reflection,

Also recalling that in November 2019, the General Conference of UNESCO, at its 40th session, adopted 40 C/Resolution 37, by which it mandated the Director-General "to prepare an international standard-setting instrument on the ethics of artificial intelligence (AI) in the form of a recommendation", which is to be submitted to the General Conference at its 41st session in 2021,

Recognizing that the development of AI technologies necessitates a commensurate increase in data, media and information literacy as well as access to independent, pluralistic, trusted sources of information, including as part of efforts to mitigate risks of misinformation, disinformation and hate speech, and harm caused through the misuse of personal data,

Observing that a normative framework for AI technologies and its social implications finds its basis in international and national legal frameworks, human rights and fundamental freedoms, ethics, need for access to data, information and knowledge, the freedom of research and innovation, human and environmental and ecosystem well-being, and connects ethical values and principles to the challenges and opportunities linked to AI technologies, based on common understanding and shared aims,

Also recognizing that ethical values and principles can help develop and implement rights-based policy measures and legal norms, by providing guidance with a view to the fast pace of technological development,

Also convinced that globally accepted ethical standards for AI technologies, in full respect of international law, in particular human rights law, can play a key role in developing AI-related norms across the globe,

Bearing in mind the Universal Declaration of Human Rights (1948), the instruments of the international human rights framework, including the Convention Relating to the Status of Refugees (1951), the Discrimination (Employment and Occupation) Convention (1958), the International Convention on the Elimination of All Forms of Racial Discrimination (1965), the International Covenant on Civil and Political Rights (1966), the International Covenant on Economic, Social and Cultural Rights (1966), the Convention on the Elimination of All Forms of Discrimination against Women (1979), the Convention on the Rights of the Child (1989), and the Convention on the Rights of Persons with Disabilities (2006), the Convention against Discrimination in Education (1960), the Convention on the Protection and Promotion of the Diversity of Cultural Expressions (2005), as well as any other relevant international instruments, recommendations and declarations,

Also noting the United Nations Declaration on the Right to Development (1986); the Declaration on the Responsibilities of the Present Generations Towards Future Generations (1997); the Universal Declaration on Bioethics and Human Rights (2005); the United Nations Declaration on the Rights of Indigenous Peoples (2007); the United Nations General Assembly resolution on the review of the World Summit on the Information Society (A/RES/70/125) (2015); the United Nations General Assembly Resolution on Transforming our world: the 2030 Agenda for Sustainable Development (A/RES/70/1) (2015); the Recommendation Concerning the Preservation of, and Access to, Documentary Heritage Including in Digital Form (2015); the Declaration of Ethical Principles in relation to Climate Change (2017); the Recommendation on Science and Scientific Researchers (2017); the Internet Universality Indicators (endorsed by UNESCO's International Programme for the Development of Communication in 2018), including the ROAM principles (endorsed by UNESCO's General Conference in 2015); the Human Rights Council's resolution on "The right to privacy in the digital age" (A/HRC/RES/42/15) (2019); and the Human Rights Council's resolution on "New and emerging digital technologies and human rights" (A/HRC/RES/41/11) (2019),

Emphasizing that specific attention must be paid to LMICs, including but not limited to LDCs, LLDCs and SIDS, as they have their own capacity but have been underrepresented in the AI ethics debate, which raises concerns about neglecting local knowledge, cultural pluralism, value systems and the demands of global fairness to deal with the positive and negative impacts of AI technologies,

Also conscious of the many existing national policies, other frameworks and initiatives elaborated by relevant United Nations entities, intergovernmental organizations, including regional organizations, as well as those by the private sector, professional organizations, non-governmental organizations, and the scientific community, related to the ethics and regulation of AI technologies,

Further convinced that AI technologies can bring important benefits, but that achieving them can also amplify tension around innovation, asymmetric access to knowledge and technologies, including the digital and civic literacy deficit that limits the public's ability to engage in topics related to AI, as well as barriers to access to information and gaps in capacity, human and institutional capacities, barriers to access to technological innovation, and a lack of adequate physical and digital infrastructure and regulatory frameworks, including those related to data, all of which need to be addressed,

Underlining that the strengthening of global cooperation and solidarity, including through multilateralism, is needed to facilitate fair access to AI technologies and address the challenges that they bring to diversity and interconnectivity of cultures and ethical systems, to mitigate potential misuse, to realize the full potential that AI can bring, especially in the area of development, and to ensure that national AI strategies are guided by ethical principles,

Taking fully into account that the rapid development of AI technologies challenges their ethical implementation and governance, as well as the respect for and protection of cultural diversity, and has the potential to disrupt local and regional ethical standards and values,

1. ***Adopts*** the present Recommendation on the Ethics of Artificial Intelligence on this twenty-third day of November 2021;

2. ***Recommends*** that Member States apply on a voluntary basis the provisions of this Recommendation by taking appropriate steps, including whatever legislative or other measures may be required, in conformity with the constitutional practice and governing structures of each State, to give effect within their jurisdictions to the principles and norms of the Recommendation in conformity with international law, including international human rights law;

3. ***Also recommends*** that Member States engage all stakeholders, including business enterprises, to ensure that they play their respective roles in the implementation of this Recommendation; and bring the Recommendation to the attention of the authorities, bodies, research and academic organizations, institutions and organizations in public, private and civil society sectors involved in AI technologies, so that the development and use of AI technologies are guided by both sound scientific research as well as ethical analysis and evaluation.

I. SCOPE OF APPLICATION

1. This Recommendation addresses ethical issues related to the domain of Artificial Intelligence to the extent that they are within UNESCO's mandate. It approaches AI ethics as a systematic normative reflection, based on a holistic, comprehensive, multicultural and evolving framework of interdependent values, principles and actions that can guide societies in dealing responsibly with the known and unknown impacts of AI technologies on human beings, societies and the environment and ecosystems, and offers them a basis to accept or reject AI technologies. It considers ethics as a dynamic basis for the normative evaluation and guidance of AI technologies, referring to human dignity, well-being and the prevention of harm as a compass and as rooted in the ethics of science and technology.

2. This Recommendation does not have the ambition to provide one single definition of AI, since such a definition would need to change over time, in accordance with technological developments. Rather, its ambition is to address those features of AI systems that are of central ethical relevance. Therefore, this Recommendation approaches AI systems as systems which have the capacity to process data and information in a way that resembles intelligent behaviour, and typically includes aspects of reasoning, learning, perception, prediction, planning or control. Three elements have a central place in this approach:

(a) AI systems are information-processing technologies that integrate models and algorithms that produce a capacity to learn and to perform cognitive tasks leading to outcomes such as prediction and decision-making in material and virtual environments. AI systems are designed to operate with varying degrees of autonomy by means of knowledge modelling and representation and by exploiting data and calculating correlations. AI systems may include several methods, such as but not limited to:

(i) machine learning, including deep learning and reinforcement learning;

(ii) machine reasoning, including planning, scheduling, knowledge representation and reasoning, search, and optimization.

AI systems can be used in cyber-physical systems, including the Internet of things, robotic systems, social robotics, and human-computer interfaces, which involve control, perception, the processing of data collected by sensors, and the operation of actuators in the environment in which AI systems work.

(b) Ethical questions regarding AI systems pertain to all stages of the AI system life cycle, understood here to range from research, design and development to deployment and use, including maintenance, operation, trade, financing, monitoring and evaluation, validation, end-of-use, disassembly and termination. In addition, AI actors can be defined as any actor involved in at least one stage of the AI system life cycle, and can refer both to natural and legal persons, such as researchers, programmers, engineers, data scientists, end-users, business enterprises, universities and public and private entities, among others.

(c) AI systems raise new types of ethical issues that include, but are not limited to, their impact on decision-making, employment and labour, social interaction, health care, education, media, access to information, digital divide, personal data and consumer protection, environment, democracy, rule of law, security and policing, dual use, and human rights and fundamental freedoms, including freedom of expression, privacy and non-discrimination. Furthermore, new ethical challenges are created by the potential of AI algorithms to reproduce and reinforce existing biases, and thus to exacerbate already existing forms of discrimination, prejudice and stereotyping. Some of these issues are related to the capacity of AI systems to perform tasks which previously only living beings could do, and which were in some cases even limited to human beings only. These characteristics give AI systems a profound, new role in human practices and society, as well as in their relationship with the environment and ecosystems, creating a new context for children and young people to grow up in, develop an understanding of the world and themselves, critically understand media and information, and learn to make decisions. In the long term, AI systems could challenge humans' special sense of experience and agency, raising additional concerns about, inter alia, human self-understanding, social, cultural and environmental interaction, autonomy, agency, worth and dignity.

3. This Recommendation pays specific attention to the broader ethical implications of AI systems in relation to the central domains of UNESCO: education, science, culture, and communication and information, as explored in the 2019 Preliminary Study on the Ethics of Artificial Intelligence by the UNESCO World Commission on Ethics of Scientific Knowledge and Technology (COMEST):

(a) Education, because living in digitalizing societies requires new educational practices, ethical reflection, critical thinking, responsible design practices and new skills, given the implications for the labour market, employability and civic participation.

(b) Science, in the broadest sense and including all academic fields from the natural sciences and medical sciences to the social sciences and humanities, as AI technologies bring new research capacities and approaches, have implications for our concepts of scientific understanding and explanation, and create a new basis for decision-making.

(c) Cultural identity and diversity, as AI technologies can enrich cultural and creative industries, but can also lead to an increased concentration of supply of cultural content, data, markets and income in the hands of only a few actors, with potential negative implications for the diversity and pluralism of languages, media, cultural expressions, participation and equality.

(d) Communication and information, as AI technologies play an increasingly important role in the processing, structuring and provision of information; the issues of automated journalism and the algorithmic provision of news and moderation and curation of content on social media and search engines are just a few examples raising issues related to access to information, disinformation, misinformation, hate speech, the emergence of new forms of societal narratives, discrimination, freedom of expression, privacy and media and information literacy, among others.

4. This Recommendation is addressed to Member States, both as AI actors and as authorities responsible for developing legal and regulatory frameworks throughout the entire AI system life cycle, and for promoting business responsibility. It also provides ethical guidance to all AI actors, including the public and private sectors, by providing a basis for an ethical impact assessment of AI systems throughout their life cycle.

II. AIMS AND OBJECTIVES

5. This Recommendation aims to provide a basis to make AI systems work for the good of humanity, individuals, societies and the environment and ecosystems, and to prevent harm. It also aims at stimulating the peaceful use of AI systems.

6. In addition to the existing ethical frameworks regarding AI around the world, this Recommendation aims to bring a globally accepted normative instrument that focuses not only on the articulation of values and principles, but also on their practical realization, via concrete policy recommendations, with a strong emphasis on inclusion issues of gender equality and protection of the environment and ecosystems.

7. Because the complexity of the ethical issues surrounding AI necessitates the cooperation of multiple stakeholders across the various levels and sectors of international, regional and national communities, this Recommendation aims to enable stakeholders to take shared responsibility based on a global and intercultural dialogue.

8. The objectives of this Recommendation are:

(a) 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;

(b) 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;

(c) to protect, promote and respect human rights and fundamental freedoms, human dignity and equality, including gender equality; to safeguard the interests of present and future generations; to preserve the environment, biodiversity and ecosystems; and to respect cultural diversity in all stages of the AI system life cycle;

(d) to foster multi-stakeholder, multidisciplinary and pluralistic dialogue and consensus building about ethical issues relating to AI systems;

(e) 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 LDCs, LLDCs and SIDS.

III. VALUES AND PRINCIPLES

9. The values and principles included below should be respected by all actors in the AI system life cycle, in the first place and, where needed and appropriate, be promoted through amendments to the existing and elaboration of new legislation, regulations and business guidelines. This must comply with international law, including the United Nations Charter and Member States' human rights obligations, and should be in line with internationally agreed social, political, environmental, educational, scientific and economic sustainability objectives, such as the United Nations Sustainable Development Goals (SDGs).

10. Values play a powerful role as motivating ideals in shaping policy measures and legal norms. While the set of values outlined below thus inspires desirable behaviour and represents the foundations of principles, the principles unpack the values underlying them more concretely so that the values can be more easily operationalized in policy statements and actions.

11. While all the values and principles outlined below are desirable per se, in any practical contexts, there may be tensions between these values and principles. In any given situation, a contextual assessment will be necessary to manage potential tensions, taking into account the principle of proportionality and in compliance with human rights and fundamental freedoms. In all cases, any possible limitations on human rights and fundamental freedoms must have a lawful basis, and be reasonable, necessary and proportionate, and consistent with States' obligations under international law. To navigate such scenarios judiciously will typically require engagement with a broad range of appropriate stakeholders, making use of social dialogue, as well as ethical deliberation, due diligence and impact assessment.

12. The trustworthiness and integrity of the life cycle of AI systems is essential to ensure that AI technologies will work for the good of humanity, individuals, societies and the environment and ecosystems, and embody

the values and principles set out in this Recommendation. People should have good reason to trust that AI systems can bring individual and shared benefits, while adequate measures are taken to mitigate risks. An essential requirement for trustworthiness is that, throughout their life cycle, AI systems are subject to thorough monitoring by the relevant stakeholders as appropriate. As trustworthiness is an outcome of the operationalization of the principles in this document, the policy actions proposed in this Recommendation are all directed at promoting trustworthiness in all stages of the AI system life cycle.

III.1 VALUES

Respect, protection and promotion of human rights and fundamental freedoms and human dignity

13. The inviolable and inherent dignity of every human constitutes the foundation for the universal, indivisible, inalienable, interdependent and interrelated system of human rights and fundamental freedoms. Therefore, respect, protection and promotion of human dignity and rights as established by international law, including international human rights law, is essential throughout the life cycle of AI systems. Human dignity relates to the recognition of the intrinsic and equal worth of each individual human being, regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds.

14. No human being or human community should be harmed or subordinated, whether physically, economically, socially, politically, culturally or mentally during any phase of the life cycle of AI systems. Throughout the life cycle of AI systems, the quality of life of human beings should be enhanced, while the definition of “quality of life” should be left open to individuals or groups, as long as there is no violation or abuse of human rights and fundamental freedoms, or the dignity of humans in terms of this definition.

15. Persons may interact with AI systems throughout their life cycle and receive assistance from them, such as care for vulnerable people or people in vulnerable situations, including but not limited to children, older persons, persons with disabilities or the ill. Within such interactions, persons should never be objectified, nor should their dignity be otherwise undermined, or human rights and fundamental freedoms violated or abused.

16. Human rights and fundamental freedoms must be respected, protected and promoted throughout the life cycle of AI systems. Governments, private sector, civil society, international organizations, technical communities and academia must respect human rights instruments and frameworks in their interventions in the processes surrounding the life cycle of AI systems. New technologies need to provide new means to advocate, defend and exercise human rights and not to infringe them.

Environment and ecosystem flourishing

17. Environmental and ecosystem flourishing should be recognized, protected and promoted through the life cycle of AI systems. Furthermore, environment and ecosystems are the existential necessity for humanity and other living beings to be able to enjoy the benefits of advances in AI.

18. All actors involved in the life cycle of AI systems must comply with applicable international law and domestic legislation, standards and practices, such as precaution, designed for environmental and ecosystem protection and restoration, and sustainable development. They should reduce the environmental impact of AI systems, including but not limited to its carbon footprint, to ensure the minimization of climate change and environmental risk factors, and prevent the unsustainable exploitation, use and transformation of natural resources contributing to the deterioration of the environment and the degradation of ecosystems.

Ensuring diversity and inclusiveness

19. Respect, protection and promotion of diversity and inclusiveness should be ensured throughout the life cycle of AI systems, consistent with international law, including human rights law. This may be done by promoting active participation of all individuals or groups regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds.

20. The scope of lifestyle choices, beliefs, opinions, expressions or personal experiences, including the optional use of AI systems and the co-design of these architectures should not be restricted during any phase of the life cycle of AI systems.

21. Furthermore, efforts, including international cooperation, should be made to overcome, and never take advantage of, the lack of necessary technological infrastructure, education and skills, as well as legal frameworks, particularly in LMICs, LDCs, LLDCs and SIDS, affecting communities.

Living in peaceful, just and interconnected societies

22. AI actors should play a participative and enabling role to ensure peaceful and just societies, which is based on an interconnected future for the benefit of all, consistent with human rights and fundamental freedoms. The value of living in peaceful and just societies points to the potential of AI systems to contribute throughout their life cycle to the interconnectedness of all living creatures with each other and with the natural environment.

23. The notion of humans being interconnected is based on the knowledge that every human belongs to a greater whole, which thrives when all its constituent parts are enabled to thrive. Living in peaceful, just and interconnected societies requires an organic, immediate, uncalculated bond of solidarity, characterized by a permanent search for peaceful relations, tending towards care for others and the natural environment in the broadest sense of the term.

24. This value demands that peace, inclusiveness and justice, equity and interconnectedness should be promoted throughout the life cycle of AI systems, in so far as the processes of the life cycle of AI systems should not segregate, objectify or undermine freedom and autonomous decision-making as well as the safety of human beings and communities, divide and turn individuals and groups against each other, or threaten the coexistence between humans, other living beings and the natural environment.

III.2 PRINCIPLES

Proportionality and Do No Harm

25. It should be recognized that AI technologies do not necessarily, per se, ensure human and environmental and ecosystem flourishing. Furthermore, none of the processes related to the AI system life cycle shall exceed what is necessary to achieve legitimate aims or objectives and should be appropriate to the context. In the event of possible occurrence of any harm to human beings, human rights and fundamental freedoms, communities and society at large or the environment and ecosystems, the implementation of procedures for risk assessment and the adoption of measures in order to preclude the occurrence of such harm should be ensured.

26. The choice to use AI systems and which AI method to use should be justified in the following ways: (a) the AI method chosen should be appropriate and proportional to achieve a given legitimate aim; (b) the AI method chosen should not infringe upon the foundational values captured in this document, in particular, its use must not violate or abuse human rights; and (c) the AI method should be appropriate to the context and should be based on rigorous scientific foundations. In scenarios where decisions are understood to have an impact that is irreversible or difficult to reverse or may involve life and death decisions, final human determination should apply. In particular, AI systems should not be used for social scoring or mass surveillance purposes.

Safety and security

27. Unwanted harms (safety risks), as well as vulnerabilities to attack (security risks) should be avoided and should be addressed, prevented and eliminated throughout the life cycle of AI systems to ensure human, environmental and ecosystem safety and security. Safe and secure AI will be enabled by the development of sustainable, privacy-protective data access frameworks that foster better training and validation of AI models utilizing quality data.

Fairness and non-discrimination

28. AI actors should promote social justice and safeguard fairness and non-discrimination of any kind in compliance with international law. This implies an inclusive approach to ensuring that the benefits of AI technologies are available and accessible to all, taking into consideration the specific needs of different age groups, cultural systems, different language groups, persons with disabilities, girls and women, and disadvantaged, marginalized and vulnerable people or people in vulnerable situations. Member States should work to promote inclusive access for all, including local communities, to AI systems with locally relevant content

and services, and with respect for multilingualism and cultural diversity. Member States should work to tackle digital divides and ensure inclusive access to and participation in the development of AI. At the national level, Member States should promote equity between rural and urban areas, and among all persons regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds, in terms of access to and participation in the AI system life cycle. At the international level, the most technologically advanced countries have a responsibility of solidarity with the least advanced to ensure that the benefits of AI technologies are shared such that access to and participation in the AI system life cycle for the latter contributes to a fairer world order with regard to information, communication, culture, education, research and socio-economic and political stability.

29. AI actors should make all reasonable efforts to minimize and avoid reinforcing or perpetuating discriminatory or biased applications and outcomes throughout the life cycle of the AI system to ensure fairness of such systems. Effective remedy should be available against discrimination and biased algorithmic determination.

30. Furthermore, digital and knowledge divides within and between countries need to be addressed throughout an AI system life cycle, including in terms of access and quality of access to technology and data, in accordance with relevant national, regional and international legal frameworks, as well as in terms of connectivity, knowledge and skills and meaningful participation of the affected communities, such that every person is treated equitably.

Sustainability

31. The development of sustainable societies relies on the achievement of a complex set of objectives on a continuum of human, social, cultural, economic and environmental dimensions. The advent of AI technologies can either benefit sustainability objectives or hinder their realization, depending on how they are applied across countries with varying levels of development. The continuous assessment of the human, social, cultural, economic and environmental impact of AI technologies should therefore be carried out with full cognizance of the implications of AI technologies for sustainability as a set of constantly evolving goals across a range of dimensions, such as currently identified in the Sustainable Development Goals (SDGs) of the United Nations.

Right to Privacy, and Data Protection

32. Privacy, a right essential to the protection of human dignity, human autonomy and human agency, must be respected, protected and promoted throughout the life cycle of AI systems. It is important that data for AI systems be collected, used, shared, archived and deleted in ways that are consistent with international law and in line with the values and principles set forth in this Recommendation, while respecting relevant national, regional and international legal frameworks.

33. Adequate data protection frameworks and governance mechanisms should be established in a multi-stakeholder approach at the national or international level, protected by judicial systems, and ensured throughout the life cycle of AI systems. Data protection frameworks and any related mechanisms should take reference from international data protection principles and standards concerning the collection, use and disclosure of personal data and exercise of their rights by data subjects while ensuring a legitimate aim and a valid legal basis for the processing of personal data, including informed consent.

34. Algorithmic systems require adequate privacy impact assessments, which also include societal and ethical considerations of their use and an innovative use of the privacy by design approach. AI actors need to ensure that they are accountable for the design and implementation of AI systems in such a way as to ensure that personal information is protected throughout the life cycle of the AI system.

Human oversight and determination

35. Member States should ensure that it is always possible to attribute ethical and legal responsibility for any stage of the life cycle of AI systems, as well as in cases of remedy related to AI systems, to physical persons or to existing legal entities. Human oversight refers thus not only to individual human oversight, but to inclusive public oversight, as appropriate.

36. It may be the case that sometimes humans would choose to rely on AI systems for reasons of efficacy, but the decision to cede control in limited contexts remains that of humans, as humans can resort to AI systems

in decision-making and acting, but an AI system can never replace ultimate human responsibility and accountability. As a rule, life and death decisions should not be ceded to AI systems.

Transparency and explainability

37. The transparency and explainability of AI systems are often essential preconditions to ensure the respect, protection and promotion of human rights, fundamental freedoms and ethical principles. Transparency is necessary for relevant national and international liability regimes to work effectively. A lack of transparency could also undermine the possibility of effectively challenging decisions based on outcomes produced by AI systems and may thereby infringe the right to a fair trial and effective remedy, and limits the areas in which these systems can be legally used.

38. While efforts need to be made to increase transparency and explainability of AI systems, including those with extra-territorial impact, throughout their life cycle to support democratic governance, the level of transparency and explainability should always be appropriate to the context and impact, as there may be a need to balance between transparency and explainability and other principles such as privacy, safety and security. People should be fully informed when a decision is informed by or is made on the basis of AI algorithms, including when it affects their safety or human rights, and in those circumstances should have the opportunity to request explanatory information from the relevant AI actor or public sector institutions. In addition, individuals should be able to access the reasons for a decision affecting their rights and freedoms, and have the option of making submissions to a designated staff member of the private sector company or public sector institution able to review and correct the decision. AI actors should inform users when a product or service is provided directly or with the assistance of AI systems in a proper and timely manner.

39. From a socio-technical lens, greater transparency contributes to more peaceful, just, democratic and inclusive societies. It allows for public scrutiny that can decrease corruption and discrimination, and can also help detect and prevent negative impacts on human rights. Transparency aims at providing appropriate information to the respective addressees to enable their understanding and foster trust. Specific to the AI system, transparency can enable people to understand how each stage of an AI system is put in place, appropriate to the context and sensitivity of the AI system. It may also include insight into factors that affect a specific prediction or decision, and whether or not appropriate assurances (such as safety or fairness measures) are in place. In cases of serious threats of adverse human rights impacts, transparency may also require the sharing of code or datasets.

40. Explainability refers to making intelligible and providing insight into the outcome of AI systems. The explainability of AI systems also refers to the understandability of the input, output and the functioning of each algorithmic building block and how it contributes to the outcome of the systems. Thus, explainability is closely related to transparency, as outcomes and sub-processes leading to outcomes should aim to be understandable and traceable, appropriate to the context. AI actors should commit to ensuring that the algorithms developed are explainable. In the case of AI applications that impact the end user in a way that is not temporary, easily reversible or otherwise low risk, it should be ensured that the meaningful explanation is provided with any decision that resulted in the action taken in order for the outcome to be considered transparent.

41. Transparency and explainability relate closely to adequate responsibility and accountability measures, as well as to the trustworthiness of AI systems.

Responsibility and accountability

42. AI actors and Member States should respect, protect and promote human rights and fundamental freedoms, and should also promote the protection of the environment and ecosystems, assuming their respective ethical and legal responsibility, in accordance with national and international law, in particular Member States' human rights obligations, and ethical guidance throughout the life cycle of AI systems, including with respect to AI actors within their effective territory and control. The ethical responsibility and liability for the decisions and actions based in any way on an AI system should always ultimately be attributable to AI actors corresponding to their role in the life cycle of the AI system.

43. Appropriate oversight, impact assessment, audit and due diligence mechanisms, including whistle-blowers' protection, should be developed to ensure accountability for AI systems and their impact throughout their life cycle. Both technical and institutional designs should ensure auditability and traceability of (the

working of) AI systems in particular to address any conflicts with human rights norms and standards and threats to environmental and ecosystem well-being.

Awareness and literacy

44. Public awareness and understanding of AI technologies and the value of data should be promoted through open and accessible education, civic engagement, digital skills and AI ethics training, media and information literacy and training led jointly by governments, intergovernmental organizations, civil society, academia, the media, community leaders and the private sector, and considering the existing linguistic, social and cultural diversity, to ensure effective public participation so that all members of society can take informed decisions about their use of AI systems and be protected from undue influence.

45. Learning about the impact of AI systems should include learning about, through and for human rights and fundamental freedoms, meaning that the approach and understanding of AI systems should be grounded by their impact on human rights and access to rights, as well as on the environment and ecosystems.

Multi-stakeholder and adaptive governance and collaboration

46. International law and national sovereignty must be respected in the use of data. That means that States, complying with international law, can regulate the data generated within or passing through their territories, and take measures towards effective regulation of data, including data protection, based on respect for the right to privacy in accordance with international law and other human rights norms and standards.

47. Participation of different stakeholders throughout the AI system life cycle is necessary for inclusive approaches to AI governance, enabling the benefits to be shared by all, and to contribute to sustainable development. Stakeholders include but are not limited to governments, intergovernmental organizations, the technical community, civil society, researchers and academia, media, education, policy-makers, private sector companies, human rights institutions and equality bodies, anti-discrimination monitoring bodies, and groups for youth and children. The adoption of open standards and interoperability to facilitate collaboration should be in place. Measures should be adopted to take into account shifts in technologies, the emergence of new groups of stakeholders, and to allow for meaningful participation by marginalized groups, communities and individuals and, where relevant, in the case of Indigenous Peoples, respect for the self-governance of their data.

IV. AREAS OF POLICY ACTION

48. The policy actions described in the following policy areas operationalize the values and principles set out in this Recommendation. The main action is for Member States to put in place effective measures, including, for example, policy frameworks or mechanisms, and to ensure that other stakeholders, such as private sector companies, academic and research institutions, and civil society adhere to them by, among other actions, encouraging all stakeholders to develop human rights, rule of law, democracy, and ethical impact assessment and due diligence tools in line with guidance including the United Nations Guiding Principles on Business and Human Rights. The process for developing such policies or mechanisms should be inclusive of all stakeholders and should take into account the circumstances and priorities of each Member State. UNESCO can be a partner and support Member States in the development as well as monitoring and evaluation of policy mechanisms.

49. UNESCO recognizes that Member States will be at different stages of readiness to implement this Recommendation, in terms of scientific, technological, economic, educational, legal, regulatory, infrastructural, societal, cultural and other dimensions. It is noted that “readiness” here is a dynamic status. In order to enable the effective implementation of this Recommendation, UNESCO will therefore: (1) develop a readiness assessment methodology to assist interested Member States in identifying their status at specific moments of their readiness trajectory along a continuum of dimensions; and (2) ensure support for interested Member States in terms of developing a UNESCO methodology for Ethical Impact Assessment (EIA) of AI technologies, sharing of best practices, assessment guidelines and other mechanisms and analytical work.

POLICY AREA 1: ETHICAL IMPACT ASSESSMENT

50. Member States should introduce frameworks for impact assessments, such as ethical impact assessment, to identify and assess benefits, concerns and risks of AI systems, as well as appropriate risk prevention, mitigation and monitoring measures, among other assurance mechanisms. Such impact assessments should identify impacts on human rights and fundamental freedoms, in particular but not limited

to the rights of marginalized and vulnerable people or people in vulnerable situations, labour rights, the environment and ecosystems and ethical and social implications, and facilitate citizen participation in line with the values and principles set forth in this Recommendation.

51. Member States and private sector companies should develop due diligence and oversight mechanisms to identify, prevent, mitigate and account for how they address the impact of AI systems on the respect for human rights, rule of law and inclusive societies. Member States should also be able to assess the socio-economic impact of AI systems on poverty and ensure that the gap between people living in wealth and poverty, as well as the digital divide among and within countries, are not increased with the massive adoption of AI technologies at present and in the future. In order to do this, in particular, enforceable transparency protocols should be implemented, corresponding to the access to information, including information of public interest held by private entities. Member States, private sector companies and civil society should investigate the sociological and psychological effects of AI-based recommendations on humans in their decision-making autonomy. AI systems identified as potential risks to human rights should be broadly tested by AI actors, including in real-world conditions if needed, as part of the Ethical Impact Assessment, before releasing them in the market.

52. Member States and business enterprises should implement appropriate measures to monitor all phases of an AI system life cycle, including the functioning of algorithms used for decision-making, the data, as well as AI actors involved in the process, especially in public services and where direct end-user interaction is needed, as part of ethical impact assessment. Member States' human rights law obligations should form part of the ethical aspects of AI system assessments.

53. Governments should adopt a regulatory framework that sets out a procedure, particularly for public authorities, to carry out ethical impact assessments on AI systems to predict consequences, mitigate risks, avoid harmful consequences, facilitate citizen participation and address societal challenges. The assessment should also establish appropriate oversight mechanisms, including auditability, traceability and explainability, which enable the assessment of algorithms, data and design processes, as well as include external review of AI systems. Ethical impact assessments should be transparent and open to the public, where appropriate. Such assessments should also be multidisciplinary, multi-stakeholder, multicultural, pluralistic and inclusive. The public authorities should be required to monitor the AI systems implemented and/or deployed by those authorities by introducing appropriate mechanisms and tools.

POLICY AREA 2: ETHICAL GOVERNANCE AND STEWARDSHIP

54. Member States should ensure that AI governance mechanisms are inclusive, transparent, multidisciplinary, multilateral (this includes the possibility of mitigation and redress of harm across borders) and multi-stakeholder. In particular, governance should include aspects of anticipation, and effective protection, monitoring of impact, enforcement and redress.

55. Member States should ensure that harms caused through AI systems are investigated and redressed, by enacting strong enforcement mechanisms and remedial actions, to make certain that human rights and fundamental freedoms and the rule of law are respected in the digital world and in the physical world. Such mechanisms and actions should include remediation mechanisms provided by private and public sector companies. The auditability and traceability of AI systems should be promoted to this end. In addition, Member States should strengthen their institutional capacities to deliver on this commitment and should collaborate with researchers and other stakeholders to investigate, prevent and mitigate any potentially malicious uses of AI systems.

56. Member States are encouraged to develop national and regional AI strategies and to consider forms of soft governance such as a certification mechanism for AI systems and the mutual recognition of their certification, according to the sensitivity of the application domain and expected impact on human rights, the environment and ecosystems, and other ethical considerations set forth in this Recommendation. Such a mechanism might include different levels of audit of systems, data, and adherence to ethical guidelines and to procedural requirements in view of ethical aspects. At the same time, such a mechanism should not hinder innovation or disadvantage small and medium enterprises or start-ups, civil society as well as research and science organizations, as a result of an excessive administrative burden. These mechanisms should also include a regular monitoring component to ensure system robustness and continued integrity and adherence to ethical guidelines over the entire life cycle of the AI system, requiring re-certification if necessary.

57. Member States and public authorities should carry out transparent self-assessment of existing and proposed AI systems, which, in particular, should include the assessment of whether the adoption of AI is appropriate and, if so, should include further assessment to determine what the appropriate method is, as well as assessment as to whether such adoption would result in violations or abuses of Member States' human rights law obligations, and if that is the case, prohibit its use.

58. Member States should encourage public entities, private sector companies and civil society organizations to involve different stakeholders in their AI governance and to consider adding the role of an independent AI Ethics Officer or some other mechanism to oversee ethical impact assessment, auditing and continuous monitoring efforts and ensure ethical guidance of AI systems. Member States, private sector companies and civil society organizations, with the support of UNESCO, are encouraged to create a network of independent AI Ethics Officers to give support to this process at national, regional and international levels.

59. Member States should foster the development of, and access to, a digital ecosystem for ethical and inclusive development of AI systems at the national level, including to address gaps in access to the AI system life cycle, while contributing to international collaboration. Such an ecosystem includes, in particular, digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate.

60. Member States should establish mechanisms, in collaboration with international organizations, transnational corporations, academic institutions and civil society, to ensure the active participation of all Member States, especially LMICs, in particular LDCs, LLDCs and SIDS, in international discussions concerning AI governance. This can be through the provision of funds, ensuring equal regional participation, or any other mechanisms. Furthermore, in order to ensure the inclusiveness of AI fora, Member States should facilitate the travel of AI actors in and out of their territory, especially from LMICs, in particular LDCs, LLDCs and SIDS, for the purpose of participating in these fora.

61. Amendments to the existing or elaboration of new national legislation addressing AI systems must comply with Member States' human rights law obligations and promote human rights and fundamental freedoms throughout the AI system life cycle. Promotion thereof should also take the form of governance initiatives, good exemplars of collaborative practices regarding AI systems, and national and international technical and methodological guidelines as AI technologies advance. Diverse sectors, including the private sector, in their practices regarding AI systems must respect, protect and promote human rights and fundamental freedoms using existing and new instruments in combination with this Recommendation.

62. Member States that acquire AI systems for human rights-sensitive use cases, such as law enforcement, welfare, employment, media and information providers, health care and the independent judiciary system should provide mechanisms to monitor the social and economic impact of such systems by appropriate oversight authorities, including independent data protection authorities, sectoral oversight and public bodies responsible for oversight.

63. Member States should enhance the capacity of the judiciary to make decisions related to AI systems as per the rule of law and in line with international law and standards, including in the use of AI systems in their deliberations, while ensuring that the principle of human oversight is upheld. In case AI systems are used by the judiciary, sufficient safeguards are needed to guarantee inter alia the protection of fundamental human rights, the rule of law, judicial independence as well as the principle of human oversight, and to ensure a trustworthy, public interest-oriented and human-centric development and use of AI systems in the judiciary.

64. Member States should ensure that governments and multilateral organizations play a leading role in ensuring the safety and security of AI systems, with multi-stakeholder participation. Specifically, Member States, international organizations and other relevant bodies should develop international standards that describe measurable, testable levels of safety and transparency, so that systems can be objectively assessed and levels of compliance determined. Furthermore, Member States and business enterprises should continuously support strategic research on potential safety and security risks of AI technologies and should encourage research into transparency and explainability, inclusion and literacy by putting additional funding into those areas for different domains and at different levels, such as technical and natural language.

65. Member States should implement policies to ensure that the actions of AI actors are consistent with international human rights law, standards and principles throughout the life cycle of AI systems, while taking into full consideration the current cultural and social diversities, including local customs and religious traditions, with due regard to the precedence and universality of human rights.

66. Member States should put in place mechanisms to require AI actors to disclose and combat any kind of stereotyping in the outcomes of AI systems and data, whether by design or by negligence, and to ensure that training data sets for AI systems do not foster cultural, economic or social inequalities, prejudice, the spreading of disinformation and misinformation, and disruption of freedom of expression and access to information. Particular attention should be given to regions where the data are scarce.

67. Member States should implement policies to promote and increase diversity and inclusiveness that reflect their populations in AI development teams and training datasets, and to ensure equal access to AI technologies and their benefits, particularly for marginalized groups, both from rural and urban zones.

68. Member States should develop, review and adapt, as appropriate, regulatory frameworks to achieve accountability and responsibility for the content and outcomes of AI systems at the different phases of their life cycle. Member States should, where necessary, introduce liability frameworks or clarify the interpretation of existing frameworks to ensure the attribution of accountability for the outcomes and the functioning of AI systems. Furthermore, when developing regulatory frameworks, Member States should, in particular, take into account that ultimate responsibility and accountability must always lie with natural or legal persons and that AI systems should not be given legal personality themselves. To ensure this, such regulatory frameworks should be consistent with the principle of human oversight and establish a comprehensive approach focused on AI actors and the technological processes involved across the different stages of the AI system life cycle.

69. In order to establish norms where these do not exist, or to adapt the existing legal frameworks, Member States should involve all AI actors (including, but not limited to, researchers, representatives of civil society and law enforcement, insurers, investors, manufacturers, engineers, lawyers and users). The norms can mature into best practices, laws and regulations. Member States are further encouraged to use mechanisms such as policy prototypes and regulatory sandboxes to accelerate the development of laws, regulations and policies, including regular reviews thereof, in line with the rapid development of new technologies and ensure that laws and regulations can be tested in a safe environment before being officially adopted. Member States should support local governments in the development of local policies, regulations and laws in line with national and international legal frameworks.

70. Member States should set clear requirements for AI system transparency and explainability so as to help ensure the trustworthiness of the full AI system life cycle. Such requirements should involve the design and implementation of impact mechanisms that take into consideration the nature of application domain, intended use, target audience and feasibility of each particular AI system.

POLICY AREA 3: DATA POLICY

71. Member States should work to develop data governance strategies that ensure the continual evaluation of the quality of training data for AI systems including the adequacy of the data collection and selection processes, proper data security and protection measures, as well as feedback mechanisms to learn from mistakes and share best practices among all AI actors.

72. Member States should put in place appropriate safeguards to protect the right to privacy in accordance with international law, including addressing concerns such as surveillance. Member States should, among others, adopt or enforce legislative frameworks that provide appropriate protection, compliant with international law. Member States should strongly encourage all AI actors, including business enterprises, to follow existing international standards and, in particular, to carry out adequate privacy impact assessments, as part of ethical impact assessments, which take into account the wider socio-economic impact of the intended data processing, and to apply privacy by design in their systems. Privacy should be respected, protected and promoted throughout the life cycle of AI systems.

73. Member States should ensure that individuals retain rights over their personal data and are protected by a framework, which notably foresees: transparency; appropriate safeguards for the processing of sensitive data; an appropriate level of data protection; effective and meaningful accountability schemes and mechanisms; the full enjoyment of the data subjects' rights and the ability to access and erase their personal data in AI systems, except for certain circumstances in compliance with international law; an appropriate level of protection in full compliance with data protection legislation where data are being used for commercial purposes such as enabling micro-targeted advertising, transferred cross-border; and an effective independent oversight as part of a data governance mechanism which keeps individuals in control of their personal data and fosters the benefits of a free flow of information internationally, including access to data.

74. Member States should establish their data policies or equivalent frameworks, or reinforce existing ones, to ensure full security for personal data and sensitive data, which, if disclosed, may cause exceptional damage, injury or hardship to individuals. Examples include data relating to offences, criminal proceedings and convictions, and related security measures; biometric, genetic and health data; and -personal data such as that relating to race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other characteristics.

75. Member States should promote open data. In this regard, Member States should consider reviewing their policies and regulatory frameworks, including on access to information and open government to reflect AI-specific requirements and promoting mechanisms, such as open repositories for publicly funded or publicly held data and source code and data trusts, to support the safe, fair, legal and ethical sharing of data, among others.

76. Member States should promote and facilitate the use of quality and robust datasets for training, development and use of AI systems, and exercise vigilance in overseeing their collection and use. This could, if possible and feasible, include investing in the creation of gold standard datasets, including open and trustworthy datasets, which are diverse, constructed on a valid legal basis, including consent of data subjects, when required by law. Standards for annotating datasets should be encouraged, including disaggregating data on gender and other bases, so it can easily be determined how a dataset is gathered and what properties it has.

77. Member States, as also suggested in the report of the United Nations Secretary-General's High-level Panel on Digital Cooperation, with the support of the United Nations and UNESCO, should adopt a digital commons approach to data where appropriate, increase interoperability of tools and datasets and interfaces of systems hosting data, and encourage private sector companies to share the data they collect with all stakeholders, as appropriate, for research, innovation or public benefits. They should also promote public and private efforts to create collaborative platforms to share quality data in trusted and secured data spaces.

POLICY AREA 4: DEVELOPMENT AND INTERNATIONAL COOPERATION

78. Member States and transnational corporations should prioritize AI ethics by including discussions of AI-related ethical issues into relevant international, intergovernmental and multi-stakeholder fora.

79. Member States should ensure that the use of AI in areas of development such as education, science, culture, communication and information, health care, agriculture and food supply, environment, natural resource and infrastructure management, economic planning and growth, among others, adheres to the values and principles set forth in this Recommendation.

80. Member States should work through international organizations to provide platforms for international cooperation on AI for development, including by contributing expertise, funding, data, domain knowledge, infrastructure, and facilitating multi-stakeholder collaboration to tackle challenging development problems, especially for LMICs, in particular LDCs, LLDCs and SIDS.

81. Member States should work to promote international collaboration on AI research and innovation, including research and innovation centres and networks that promote greater participation and leadership of researchers from LMICs and other countries, including LDCs, LLDCs and SIDS.

82. Member States should promote AI ethics research by engaging international organizations and research institutions, as well as transnational corporations, that can be a basis for the ethical use of AI systems by public and private entities, including research into the applicability of specific ethical frameworks in specific cultures and contexts, and the possibilities to develop technologically feasible solutions in line with these frameworks.

83. Member States should encourage international cooperation and collaboration in the field of AI to bridge geo-technological lines. Technological exchanges and consultations should take place between Member States and their populations, between the public and private sectors, and between and among the most and least technologically advanced countries in full respect of international law.

POLICY AREA 5: ENVIRONMENT AND ECOSYSTEMS

84. Member States and business enterprises should assess the direct and indirect environmental impact throughout the AI system life cycle, including, but not limited to, its carbon footprint, energy consumption and

the environmental impact of raw material extraction for supporting the manufacturing of AI technologies, and reduce the environmental impact of AI systems and data infrastructures. Member States should ensure compliance of all AI actors with environmental law, policies and practices.

85. Member States should introduce incentives, when needed and appropriate, to ensure the development and adoption of rights-based and ethical AI-powered solutions for disaster risk resilience; the monitoring, protection and regeneration of the environment and ecosystems; and the preservation of the planet. These AI systems should involve the participation of local and indigenous communities throughout the life cycle of AI systems and should support circular economy type approaches and sustainable consumption and production patterns. Some examples include using AI systems, when needed and appropriate, to:

(a) Support the protection, monitoring and management of natural resources.

(b) Support the prediction, prevention, control and mitigation of climate-related problems.

(c) Support a more efficient and sustainable food ecosystem.

(d) Support the acceleration of access to and mass adoption of sustainable energy.

(e) Enable and promote the mainstreaming of sustainable infrastructure, sustainable business models and sustainable finance for sustainable development.

(f) Detect pollutants or predict levels of pollution and thus help relevant stakeholders identify, plan and put in place targeted interventions to prevent and reduce pollution and exposure.

86. When choosing AI methods, given the potential data-intensive or resource-intensive character of some of them and the respective impact on the environment, Member States should ensure that AI actors, in line with the principle of proportionality, favour data, energy and resource-efficient AI methods. Requirements should be developed to ensure that appropriate evidence is available to show that an AI application will have the intended effect, or that safeguards accompanying an AI application can support the justification for its use. If this cannot be done, the precautionary principle must be favoured, and in instances where there are disproportionate negative impacts on the environment, AI should not be used.

POLICY AREA 6: GENDER

87. Member States should ensure that the potential for digital technologies and artificial intelligence to contribute to achieving gender equality is fully maximized, and must ensure that the human rights and fundamental freedoms of girls and women, and their safety and integrity are not violated at any stage of the AI system life cycle. Moreover, Ethical Impact Assessment should include a transversal gender perspective.

88. Member States should have dedicated funds from their public budgets linked to financing gender-responsive schemes, ensure that national digital policies include a gender action plan, and develop relevant policies, for example, on labour education, targeted at supporting girls and women to make sure they are not left out of the digital economy powered by AI. Special investment in providing targeted programmes and gender-specific language, to increase the opportunities of girls' and women's participation in science, technology, engineering, and mathematics (STEM), including information and communication technologies (ICT) disciplines, preparedness, employability, equal career development and professional growth of girls and women, should be considered and implemented.

89. Member States should ensure that the potential of AI systems to advance the achievement of gender equality is realized. They should ensure that these technologies do not exacerbate the already wide gender gaps existing in several fields in the analogue world, and instead eliminate those gaps. These gaps include: the gender wage gap; the unequal representation in certain professions and activities; the lack of representation at top management positions, boards of directors, or research teams in the AI field; the education gap; the digital and AI access, adoption, usage and affordability gap; and the unequal distribution of unpaid work and of the caring responsibilities in our societies.

90. Member States should ensure that gender stereotyping and discriminatory biases are not translated into AI systems, and instead identify and proactively redress these. Efforts are necessary to avoid the compounding negative effect of technological divides in achieving gender equality and avoiding violence such as harassment, bullying or trafficking of girls and women and under-represented groups, including in the online domain.

91. Member States should encourage female entrepreneurship, participation and engagement in all stages of an AI system life cycle by offering and promoting economic, regulatory incentives, among other incentives and support schemes, as well as policies that aim at a balanced gender participation in AI research in academia, gender representation on digital and AI companies' top management positions, boards of directors and research teams. Member States should ensure that public funds (for innovation, research and technologies) are channelled to inclusive programmes and companies, with clear gender representation, and that private funds are similarly encouraged through affirmative action principles. Policies on harassment-free environments should be developed and enforced, together with the encouragement of the transfer of best practices on how to promote diversity throughout the AI system life cycle.

92. Member States should promote gender diversity in AI research in academia and industry by offering incentives to girls and women to enter the field, putting in place mechanisms to fight gender stereotyping and harassment within the AI research community, and encouraging academic and private entities to share best practices on how to enhance gender diversity.

93. UNESCO can help form a repository of best practices for incentivizing the participation of girls, women and under-represented groups in all stages of the AI system life cycle.

POLICY AREA 7: CULTURE

94. Member States are encouraged to incorporate AI systems, where appropriate, in the preservation, enrichment, understanding, promotion, management and accessibility of tangible, documentary and intangible cultural heritage, including endangered languages as well as indigenous languages and knowledges, for example by introducing or updating educational programmes related to the application of AI systems in these areas, where appropriate, and by ensuring a participatory approach, targeted at institutions and the public.

95. Member States are encouraged to examine and address the cultural impact of AI systems, especially natural language processing (NLP) applications such as automated translation and voice assistants, on the nuances of human language and expression. Such assessments should provide input for the design and implementation of strategies that maximize the benefits from these systems by bridging cultural gaps and increasing human understanding, as well as addressing the negative implications such as the reduction of use, which could lead to the disappearance of endangered languages, local dialects, and tonal and cultural variations associated with human language and expression.

96. Member States should promote AI education and digital training for artists and creative professionals to assess the suitability of AI technologies for use in their profession, and contribute to the design and implementation of suitable AI technologies, as AI technologies are being used to create, produce, distribute, broadcast and consume a variety of cultural goods and services, bearing in mind the importance of preserving cultural heritage, diversity and artistic freedom.

97. Member States should promote awareness and evaluation of AI tools among local cultural industries and small and medium enterprises working in the field of culture, to avoid the risk of concentration in the cultural market.

98. Member States should engage technology companies and other stakeholders to promote a diverse supply of and plural access to cultural expressions, and in particular to ensure that algorithmic recommendation enhances the visibility and discoverability of local content.

99. Member States should foster new research at the intersection between AI and intellectual property (IP), for example to determine whether or how to protect with IP rights the works created by means of AI technologies. Member States should also assess how AI technologies are affecting the rights or interests of IP owners, whose works are used to research, develop, train or implement AI applications.

100. Member States should encourage museums, galleries, libraries and archives at the national level to use AI systems to highlight their collections and enhance their libraries, databases and knowledge base, while also providing access to their users.

POLICY AREA 8: EDUCATION AND RESEARCH

101. Member States should work with international organizations, educational institutions and private and non-governmental entities to provide adequate AI literacy education to the public on all levels in all countries

in order to empower people and reduce the digital divides and digital access inequalities resulting from the wide adoption of AI systems.

102. Member States should promote the acquisition of “prerequisite skills” for AI education, such as basic literacy, numeracy, coding and digital skills, and media and information literacy, as well as critical and creative thinking, teamwork, communication, socio-emotional and AI ethics skills, especially in countries and in regions or areas within countries where there are notable gaps in the education of these skills.

103. Member States should promote general awareness programmes about AI developments, including on data and the opportunities and challenges brought about by AI technologies, the impact of AI systems on human rights and their implications, including children’s rights. These programmes should be accessible to non-technical as well as technical groups.

104. Member States should encourage research initiatives on the responsible and ethical use of AI technologies in teaching, teacher training and e-learning, among other issues, to enhance opportunities and mitigate the challenges and risks involved in this area. The initiatives should be accompanied by an adequate assessment of the quality of education and impact on students and teachers of the use of AI technologies. Member States should also ensure that AI technologies empower students and teachers and enhance their experience, bearing in mind that relational and social aspects and the value of traditional forms of education are vital in teacher-student and student-student relationships and should be considered when discussing the adoption of AI technologies in education. AI systems used in learning should be subject to strict requirements when it comes to the monitoring, assessment of abilities, or prediction of the learners’ behaviours. AI should support the learning process without reducing cognitive abilities and without extracting sensitive information, in compliance with relevant personal data protection standards. The data handed over to acquire knowledge collected during the learner’s interactions with the AI system must not be subject to misuse, misappropriation or criminal exploitation, including for commercial purposes.

105. Member States should promote the participation and leadership of girls and women, diverse ethnicities and cultures, persons with disabilities, marginalized and vulnerable people or people in vulnerable situations, minorities and all persons not enjoying the full benefits of digital inclusion, in AI education programmes at all levels, as well as the monitoring and sharing of best practices in this regard with other Member States.

106. Member States should develop, in accordance with their national education programmes and traditions, AI ethics curricula for all levels, and promote cross-collaboration between AI technical skills education and humanistic, ethical and social aspects of AI education. Online courses and digital resources of AI ethics education should be developed in local languages, including indigenous languages, and take into account the diversity of environments, especially ensuring accessibility of formats for persons with disabilities.

107. Member States should promote and support AI research, notably AI ethics research, including for example through investing in such research or by creating incentives for the public and private sectors to invest in this area, recognizing that research contributes significantly to the further development and improvement of AI technologies with a view to promoting international law and the values and principles set forth in this Recommendation. Member States should also publicly promote the best practices of, and cooperation with, researchers and companies who develop AI in an ethical manner.

108. Member States should ensure that AI researchers are trained in research ethics and require them to include ethical considerations in their designs, products and publications, especially in the analyses of the datasets they use, how they are annotated, and the quality and scope of the results with possible applications.

109. Member States should encourage private sector companies to facilitate the access of the scientific community to their data for research, especially in LMICs, in particular LDCs, LLDCs and SIDS. This access should conform to relevant privacy and data protection standards.

110. To ensure a critical evaluation of AI research and proper monitoring of potential misuses or adverse effects, Member States should ensure that any future developments with regards to AI technologies should be based on rigorous and independent scientific research, and promote interdisciplinary AI research by including disciplines other than science, technology, engineering and mathematics (STEM), such as cultural studies, education, ethics, international relations, law, linguistics, philosophy, political science, sociology and psychology.

111. Recognizing that AI technologies present great opportunities to help advance scientific knowledge and practice, especially in traditionally model-driven disciplines, Member States should encourage scientific communities to be aware of the benefits, limits and risks of their use; this includes attempting to ensure that conclusions drawn from data-driven approaches, models and treatments are robust and sound. Furthermore, Member States should welcome and support the role of the scientific community in contributing to policy and in cultivating awareness of the strengths and weaknesses of AI technologies.

POLICY AREA 9: COMMUNICATION AND INFORMATION

112. Member States should use AI systems to improve access to information and knowledge. This can include support to researchers, academia, journalists, the general public and developers, to enhance freedom of expression, academic and scientific freedoms, access to information, and increased proactive disclosure of official data and information.

113. Member States should ensure that AI actors respect and promote freedom of expression as well as access to information with regard to automated content generation, moderation and curation. Appropriate frameworks, including regulation, should enable transparency of online communication and information operators and ensure users have access to a diversity of viewpoints, as well as processes for prompt notification to the users on the reasons for removal or other treatment of content, and appeal mechanisms that allow users to seek redress.

114. Member States should invest in and promote digital and media and information literacy skills to strengthen critical thinking and competencies needed to understand the use and implication of AI systems, in order to mitigate and counter disinformation, misinformation and hate speech. A better understanding and evaluation of both the positive and potentially harmful effects of recommender systems should be part of those efforts.

115. Member States should create enabling environments for media to have the rights and resources to effectively report on the benefits and harms of AI systems, and also encourage media to make ethical use of AI systems in their operations.

POLICY AREA 10: ECONOMY AND LABOUR

116. Member States should assess and address the impact of AI systems on labour markets and its implications for education requirements, in all countries and with special emphasis on countries where the economy is labour-intensive. This can include the introduction of a wider range of “core” and interdisciplinary skills at all education levels to provide current workers and new generations a fair chance of finding jobs in a rapidly changing market, and to ensure their awareness of the ethical aspects of AI systems. Skills such as “learning how to learn”, communication, critical thinking, teamwork, empathy, and the ability to transfer one’s knowledge across domains, should be taught alongside specialist, technical skills, as well as low-skilled tasks. Being transparent about what skills are in demand and updating curricula around these are key.

117. Member States should support collaboration agreements among governments, academic institutions, vocational education and training institutions, industry, workers’ organizations and civil society to bridge the gap of skillset requirements to align training programmes and strategies with the implications of the future of work and the needs of industry, including small and medium enterprises. Project-based teaching and learning approaches for AI should be promoted, allowing for partnerships between public institutions, private sector companies, universities and research centres.

118. Member States should work with private sector companies, civil society organizations and other stakeholders, including workers and unions to ensure a fair transition for at-risk employees. This includes putting in place upskilling and reskilling programmes, finding effective mechanisms of retaining employees during those transition periods, and exploring “safety net” programmes for those who cannot be retrained. Member States should develop and implement programmes to research and address the challenges identified that could include upskilling and reskilling, enhanced social protection, proactive industry policies and interventions, tax benefits, new taxation forms, among others. Member States should ensure that there is sufficient public funding to support these programmes. Relevant regulations, such as tax regimes, should be carefully examined and changed if needed to counteract the consequences of unemployment caused by AI-based automation.

119. Member States should encourage and support researchers to analyse the impact of AI systems on the local labour environment in order to anticipate future trends and challenges. These studies should have an interdisciplinary approach and investigate the impact of AI systems on economic, social and geographic sectors, as well as on human-robot interactions and human-human relationships, in order to advise on reskilling and redeployment best practices.

120. Member States should take appropriate steps to ensure competitive markets and consumer protection, considering possible measures and mechanisms at national, regional and international levels, to prevent abuse of dominant market positions, including by monopolies, in relation to AI systems throughout their life cycle, whether these are data, research, technology, or market. Member States should prevent the resulting inequalities, assess relevant markets and promote competitive markets. Due consideration should be given to LMICs, in particular LDCs, LLDCs and SIDS, which are more exposed and vulnerable to the possibility of abuses of market dominance as a result of a lack of infrastructure, human capacity and regulations, among other factors. AI actors developing AI systems in countries which have established or adopted ethical standards on AI should respect these standards when exporting these products, developing or applying their AI systems in countries where such standards may not exist, while respecting applicable international law and domestic legislation, standards and practices of these countries.

POLICY AREA 11: HEALTH AND SOCIAL WELL-BEING

121. Member States should endeavour to employ effective AI systems for improving human health and protecting the right to life, including mitigating disease outbreaks, while building and maintaining international solidarity to tackle global health risks and uncertainties, and ensure that their deployment of AI systems in health care be consistent with international law and their human rights law obligations. Member States should ensure that actors involved in health care AI systems take into consideration the importance of a patient's relationships with their family and with health care staff.

122. Member States should ensure that the development and deployment of AI systems related to health in general and mental health in particular, paying due attention to children and youth, is regulated to the effect that they are safe, effective, efficient, scientifically and medically proven and enable evidence-based innovation and medical progress. Moreover, in the related area of digital health interventions, Member States are strongly encouraged to actively involve patients and their representatives in all relevant steps of the development of the system.

123. Member States should pay particular attention in regulating prediction, detection and treatment solutions for health care in AI applications by:

(a) ensuring oversight to minimize and mitigate bias;

(b) ensuring that the professional, the patient, caregiver or service user is included as a "domain expert" in the team in all relevant steps when developing the algorithms;

(c) paying due attention to privacy because of the potential need for being medically monitored and ensuring that all relevant national and international data protection requirements are met;

(d) ensuring effective mechanisms so that those whose personal data is being analysed are aware of and provide informed consent for the use and analysis of their data, without preventing access to health care;

(e) ensuring the human care and final decision of diagnosis and treatment are taken always by humans while acknowledging that AI systems can also assist in their work;

(f) ensuring, where necessary, the review of AI systems by an ethical research committee prior to clinical use.

124. Member States should establish research on the effects and regulation of potential harms to mental health related to AI systems, such as higher degrees of depression, anxiety, social isolation, developing addiction, trafficking, radicalization and misinformation, among others.

125. Member States should develop guidelines for human-robot interactions and their impact on human-human relationships, based on research and directed at the future development of robots, and with special attention to the mental and physical health of human beings. Particular attention should be given to the use of robots in health care and the care for older persons and persons with disabilities, in education, and robots for

use by children, toy robots, chatbots and companion robots for children and adults. Furthermore, assistance of AI technologies should be applied to increase the safety and ergonomic use of robots, including in a human-robot working environment. Special attention should be paid to the possibility of using AI to manipulate and abuse human cognitive biases.

126. Member States should ensure that human-robot interactions comply with the same values and principles that apply to any other AI systems, including human rights and fundamental freedoms, the promotion of diversity, and the protection of vulnerable people or people in vulnerable situations. Ethical questions related to AI-powered systems for neurotechnologies and brain-computer interfaces should be considered in order to preserve human dignity and autonomy.

127. Member States should ensure that users can easily identify whether they are interacting with a living being, or with an AI system imitating human or animal characteristics, and can effectively refuse such interaction and request human intervention.

128. Member States should implement policies to raise awareness about the anthropomorphization of AI technologies and technologies that recognize and mimic human emotions, including in the language used to mention them, and assess the manifestations, ethical implications and possible limitations of such anthropomorphization, in particular in the context of robot-human interaction and especially when children are involved.

129. Member States should encourage and promote collaborative research into the effects of long-term interaction of people with AI systems, paying particular attention to the psychological and cognitive impact that these systems can have on children and young people. This should be done using multiple norms, principles, protocols, disciplinary approaches, and assessment of the modification of behaviours and habits, as well as careful evaluation of the downstream cultural and societal impacts. Furthermore, Member States should encourage research on the effect of AI technologies on health system performance and health outcomes.

130. Member States, as well as all stakeholders, should put in place mechanisms to meaningfully engage children and young people in conversations, debates and decision-making with regard to the impact of AI systems on their lives and futures.

V. MONITORING AND EVALUATION

131. Member States should, according to their specific conditions, governing structures and constitutional provisions, credibly and transparently monitor and evaluate policies, programmes and mechanisms related to ethics of AI, using a combination of quantitative and qualitative approaches. To support Member States, UNESCO can contribute by:

(a) developing a UNESCO methodology for Ethical Impact Assessment (EIA) of AI technologies based on rigorous scientific research and grounded in international human rights law, guidance for its implementation in all stages of the AI system life cycle, and capacity-building materials to support Member States' efforts to train government officials, policy-makers and other relevant AI actors on EIA methodology;

(b) developing a UNESCO readiness assessment methodology to assist Member States in identifying their status at specific moments of their readiness trajectory along a continuum of dimensions;

(c) developing a UNESCO methodology to evaluate ex ante and ex post the effectiveness and efficiency of the policies for AI ethics and incentives against defined objectives;

(d) strengthening the research- and evidence-based analysis of and reporting on policies regarding AI ethics;

(e) collecting and disseminating progress, innovations, research reports, scientific publications, data and statistics regarding policies for AI ethics, including through existing initiatives, to support sharing best practices and mutual learning, and to advance the implementation of this Recommendation.

132. Processes for monitoring and evaluation should ensure broad participation of all stakeholders, including, but not limited to, vulnerable people or people in vulnerable situations. Social, cultural and gender diversity should be ensured, with a view to improving learning processes and strengthening the connections between findings, decision-making, transparency and accountability for results.

133. In the interests of promoting best policies and practices related to ethics of AI, appropriate tools and indicators should be developed for assessing the effectiveness and efficiency thereof against agreed standards, priorities and targets, including specific targets for persons belonging to disadvantaged, marginalized populations, and vulnerable people or people in vulnerable situations, as well as the impact of AI systems at individual and societal levels. The monitoring and assessment of the impact of AI systems and related AI ethics policies and practices should be carried out continuously in a systematic way proportionate to the relevant risks. This should be based on internationally agreed frameworks and involve evaluations of private and public institutions, providers and programmes, including self-evaluations, as well as tracer studies and the development of sets of indicators. Data collection and processing should be conducted in accordance with international law, national legislation on data protection and data privacy, and the values and principles outlined in this Recommendation.

134. In particular, Member States may wish to consider possible mechanisms for monitoring and evaluation, such as an ethics commission, AI ethics observatory, repository covering human rights-compliant and ethical development of AI systems, or contributions to existing initiatives by addressing adherence to ethical principles across UNESCO's areas of competence, an experience-sharing mechanism, AI regulatory sandboxes, and an assessment guide for all AI actors to evaluate their adherence to policy recommendations mentioned in this document.

VI. UTILIZATION AND EXPLOITATION OF THE PRESENT RECOMMENDATION

135. Member States and all other stakeholders as identified in this Recommendation should respect, promote and protect the ethical values, principles and standards regarding AI that are identified in this Recommendation, and should take all feasible steps to give effect to its policy recommendations.

136. Member States should strive to extend and complement their own action in respect of this Recommendation, by cooperating with all relevant national and international governmental and non-governmental organizations, as well as transnational corporations and scientific organizations, whose activities fall within the scope and objectives of this Recommendation. The development of a UNESCO Ethical Impact Assessment methodology and the establishment of national commissions for the ethics of AI can be important instruments for this.

VII. PROMOTION OF THE PRESENT RECOMMENDATION

137. UNESCO has the vocation to be the principal United Nations agency to promote and disseminate this Recommendation, and accordingly will work in collaboration with other relevant United Nations entities, while respecting their mandate and avoiding duplication of work.

138. UNESCO, including its bodies, such as the World Commission on the Ethics of Scientific Knowledge and Technology (COMEST), the International Bioethics Committee (IBC) and the Intergovernmental Bioethics Committee (IGBC), will also work in collaboration with other international, regional and sub-regional governmental and non-governmental organizations.

139. Even though, within UNESCO, the mandate to promote and protect falls within the authority of governments and intergovernmental bodies, civil society will be an important actor to advocate for the public sector's interests and therefore UNESCO needs to ensure and promote its legitimacy.

VIII. FINAL PROVISIONS

140. This Recommendation needs to be understood as a whole, and the foundational values and principles are to be understood as complementary and interrelated.

141. Nothing in this Recommendation may be interpreted as replacing, altering or otherwise prejudicing States' obligations or rights under international law, or as approval for any State, other political, economic or social actor, group or person to engage in any activity or perform any act contrary to human rights, fundamental freedoms, human dignity and concern for the environment and ecosystems, both living and non-living.