Economía
Secretaría de Economía

INADI
Instituto para el Desarrollo Industrial y la Transformación Digital A.C.

Artificial Intelligence:



Towards a new era of human civilization

COORDINATOR

Arturo Oropeza García

Economía
Secretaría de Economía

INADI
Instituto para el Desarrollo Industrial y la Transformación Digital A.C.



INADI
Instituto para el Desarrollo Industrial y la Transformación Digital A.C.

Artificial Intelligence:

Towards a new era of human civilization

Francisco Balaguer Callejón * Julen Berasaluce Iza
Eber Omar Betanzos Torres * Fausto Carbajal Glass
Rosario Castañón * Carlos A. Coello Coello
Amaury Fierro González * Diego Flores Jiménez
Mariel García Montes * David Guillén
Zhou Hui * Badar Alam Iqbal * Song Jin
Beatriz Juárez Aguilar * Jorge Enrique Linares Salgado
José Ramón López-Portillo Romano
Sebastian Porsdam Mann * Timo Minssen
Arturo Oropeza García * Clemente Ruiz Durán
Saiph Savage * Hannah Louise Smith
José Luis Solleiro Rebolledo * Osamu Sudoh
Kiyoshi Tsuru * Arti Yadav
Pang Zhongying * Mohamad Zreik

Arturo Oropeza García

Coordinator

INADI
Instituto para el Desarrollo Industrial y la Transformación Digital A.C.

Artificial intelligence: towards a new era of human civilization.
Coordinator: Arturo Oropeza García

First edition, 2026.

©2026, Instituto para el Desarrollo Industrial y la Transformación Digital A.C. (INADI)

©2026 Arturo Oropeza García
Coyoacán, Ciudad de México

ISBN: 978-607-99893-3-0

All rights reserved.

The opinions expressed in this book are the sole responsibility of the authors and do not necessarily reflect the position of the institutions to which they belong or of the publisher.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means –electronic, mechanical, photocopying, recording, or otherwise– without the prior written authorization of the copyright holder. Any unauthorized reproduction or distribution of this work, in whole or in part, is strictly prohibited and subject to penalties established by applicable law.

Editorial design: alexbrije

Cover and interior images: ©Freepik
Printed and made in Mexico / Impreso y hecho en México

Economía
Secretaría de Economía

INADI
Instituto para el Desarrollo Industrial
y la Transformación Digital A.C.

Artificial Intelligence:

Towards a new era of human civilization

COORDINATOR

Arturo Oropeza García

INDEX

Artificial Intelligence:
Towards a new era of human civilization

Introduction
> ARTURO OROPEZA GARCÍA — 12

CHAPTER I
AI - TOWARDS A NEW ERA
OF HUMAN CIVILIZATION — 20

1. Liberalism, democracy and artificial intelligence. The challenges of freedom — 21
> ARTURO OROPEZA GARCÍA
President of the Institute for Industrial Development and Digital Transformation (INADI) and National Researcher at the Institute of Legal Research (IIJ) of UNAM

2 Artificial intelligence governance: Towards a new era of responsibility and the common good — 56
> JOSÉ RAMÓN LÓPEZ-PORTILLO ROMANO
Ph.D. in Political Science from the University of Oxford, co-founder of the Center for Mexican Studies. He has served as Deputy Secretary of State in Mexico, as well as Permanent Representative and Independent Chairperson of the FAO Council. Mexican scholar, diplomat, entrepreneur, and public servant

3. Artificial intelligence and democracy: A synergistic and confrontational relationship — 86
> EBER OMAR BETANZOS TORRES
President of National Institute of Public Administration (INAP, A.C.)

4. Humanism in the age of artificial intelligence — 104
> JORGE ENRIQUE LINARES SALGADO
Former Director of the Faculty of Philosophy and Literature at UNAM

CHAPTER II
THE NEW POSSIBILITIES OF DIGITAL ECONOMIC-INDUSTRIAL DEVELOPMENT

128

5. Mexico's semiconductor strategy and the emergence of a digital industrial policy for the AI era: Positioning the country as a strategic hub in the global AI economy 129

> DIEGO FLORES JIMÉNEZ
Head of the Electronics and Digital Industry Sector at the Secretariat of Economy

6. Imagined futures, fictional expectations and dynamics of capitalism; the role of artificial intelligence 142

> CLEMENTE RUIZ DURÁN
Professor and Researcher at the Graduate Program in Economics, Faculty of Economics, UNAM.

7. Strategy for the adoption of AI in mexican industry 166

> JOSÉ LUIS SOLLEIRO REBOLLEDO
Senior Researcher at the Institute of Applied Sciences and Technology (ICAT) at UNAM and Advisor to the Institute for Industrial Development and Digital Transformation (INADI)

> ROSARIO CASTAÑÓN
Academic at the Institute of Applied Sciences and Technology (ICAT) at UNAM

> DAVID GUILLÉN
Consulting and Project Management Manager at Cambiotec, A.C.

8. Artificial Intelligence and the foundations of a new digital economy 190
> JULEN BERASALUCE IZA
Professor-Researcher at the Center for Economic Studies, El Colegio de México (COLMEX). Member of the Institute for Industrial Development and Digital Transformation (INADI)

9. Industrial artificial intelligence in China 212
> BEATRIZ JUÁREZ AGUILAR
Professor at the Graduate Studies Division of the UNAM Faculty of Law and Advisor to the Institute for Industrial Development and Digital Transformation (INADI)

CHAPTER III 230
ARTIFICIAL INTELLIGENCE.
GENERAL OUTLINES OF A NEW KNOWLEDGE

10. Past, present and future of artificial intelligence 231
> CARLOS A. COELLO COELLO
Member of El Colegio Nacional, Researcher at the Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN) in Mexico City, and Visiting Professor at the Basque Center for Applied Mathematics in Spain

11. The constitution of the algorithm and the ruptures in digital society 266
> FRANCISCO BALAGUER CALLEJÓN
Professor of Constitutional Law. University of Granada, Spain

CHAPTER IV
ARTIFICIAL INTELLIGENCE: THE EMERGENCE OF A NEW ERA IN REGULATION

284

285

12. The european AI act –Promises & perils. Trade offs in a complex regulatory eco-system–

> TIMO MINSSEN (Denmark)
LL.D., LL.Lic., LL.M.., MICL & Dipl. Jur., Professor of Law & Director of the Centre for Advanced Studies in Bioscience Innovation Law (CeBIL) at the University of Copenhagen, Timo.Minssen@jur.ku.dk

> HANNAH LOUISE SMITH (Denmark)
BA, M.St, BCL, DPhil, postdoctoral researcher at CeBIL, hsm@jur.ku.dk

> SEBASTIAN PORSDAM MANN (Denmark)
MA (Cantab.), PhD (Cantab.), DPhil (pending, Oxon.), GDL (BPP University), Assistant Professor at CeBIL, Sebastian.porsdam.mann@jur.ku.dk

328

13. General framework of regulation in the people's Republic of China on artificial intelligence and digital transformation

> HUI ZHOU (China)
Deputy Director and Associate Professor, Department of Cyber and Information Law, Institute of Law, Chinese Academy of Social Sciences (CASS)

> YUE ZHU (China)
Assistant professor of law at Tongji University

350

14. Artificial intelligence: What challenges should AI regulation in Mexico address?

> MARIEL GARCÍA MONTES
Researcher and PhD candidate in Science, Technology, and Society, MIT

> SAIPH SAVAGE

15. Generative artificial intelligence and its impact on creativity and intellectual property law 370
> KIYOSHI TSURU
Founding Partner and Director of TMI Law. Kiyoshi is a leading attorney specializing in intellectual property, particularly in the fields of innovation and cyberspace, domain name dispute resolution, copyright protection in the digital environment, software protection, patents, geographical indications, electronic transactions related to intellectual property rights, and arbitration

CHAPTER V
ARTIFICIAL INTELLIGENCE.
A GLOBAL COMPARATIVE VIEW 390

16. The Age of AI in China: a partially consideration on some neglected issues 391
> PANG ZHONGYING (China)
Humanities and social sciences scholar, teaching Global International Relations, International Political Economy, and Global Economic History at Sichuan University in Chengdu, China

17. The impact of AI on China's economic development. Advances and challenges to be adressed 410
> SONG JIN (China)
Director of the Development Research Center of the Institute of World Economics and Politics (IWEP) and Researcher at the Chinese Academy of Social Sciences (CASS)

18. Advanced Generative artificial intelligence and AI governance in Japan 430
> OSAMU SUDOH (Japón)
Ph.D. and Professor, Chuo University

19. Artificial Intelligence in India: Advances and pending challenges 464

> BADAR ALAM IQBAL (India)
Researcher, Department of Economics, Stellenbosch University, South Africa

> MOHAMAD ZREIK (India)
Postdoctoral Researcher, School of International Studies, Sun Yat-sen University

> ARTI YADAV (India)
Assistant Professor, Department of Commerce, Ramanujan College, University of Delhi, India

CHAPTER VI
ARTIFICIAL INTELLIGENCE AND PUBLIC SECURITY 486

20. Artificial intelligence, disinformation and public security in Latin America 487

> FAUSTO CARBAJAL GLASS
Consultant in political risk and security. Internationalist from Universidad Iberoamericana and Master in War Studies from King's College London

> AMAURY FIERRO GONZÁLEZ
Specialist in intelligence and geopolitics and instructor at Rabdan Academy

Authors 510

INTRODUCTION

I.

The history of the world, overflowing with epics and adventures, was shaken in the mid-eighteenth century when a civilizational earthquake occurred in Europe through the first Industrial Revolution, which came to modify the life of human beings in their personal, family, social, economic, and political environments, forever. Since that first scientific revolution, human beings have not known what rest is.

At the beginning of the nineteenth century, as a result of this change in life, the world's population expanded rapidly, reaching the figure of one billion people. Economic output soared; what would be a new economic order of the world was born; in the same way, the first registration of income distribution followed one another, based on a new era of the globalization of manufacturing and trade, facilitated by technology.

From the middle of the eighteenth century and throughout the nineteenth and twentieth centuries, the majority of the population learned to live with the fascination of the innovations that dazzled them day by day, in combination with the fear of a present and future which they did not fully understand without knowledge of where such a future might lead.

As an example of the above, at the 1900 World's Fair in Paris, visitors walked dazzled through the long corridors of the Trocadero Gardens to the Eiffel Tower, contemplating with great amazement the wireless telegraph, the most powerful telescope in the world, the first escalator, the cinematograph, automobiles, the latest photographic lenses, typewriters, large cranes, increasingly powerful steam engines, and an endless number of inventions that led some of the visitors to express that *the world moved so fast that one would get dizzy, agitated in a whirlwind of progress*[1].

1 Frieden Jeffry, Capitalismo Global, p. 85.

However, of all the inventions of the fair, electricity was presented through a Palace of Electricity, appearing as the star of the exhibition to the amazement of visitors from different countries around the world, with people falling in love with it, believing that electricity would come to cure all their ills, with the new diseases of progress ultimately being religion, the scourge of the twentieth century[2].

Almost three centuries after these revelations of science and technology produced by human beings, the current of change continues to flow relentlessly, manifesting itself in a thousand ways through the different stages of the First, Second, Third, and Fourth Industrial Revolutions, where innovation has been happening continually.

With the passage of time, far from tempering the speed of these changes, in the 60s and 70s, under the dictatorship of Moore's and Koomey's laws, these technological changes acquired a speed of their own that generated the birth of a new stage of superior attributes that became what we now know as the First Digital Revolution.

Victim of its own overflow, technology is causing a train wreck between the Fourth Industrial Revolution that has not yet concluded, and the installation of the First Digital Revolution which is beginning to be born, with characteristics different from its predecessors.

The Industrial Revolution and the Digital Revolution are part of the same trend of humanity; a dialectic of talent that improves itself without limit. They are two civilizational expressions that cannot be explained without each other.

They are also part of a process of cultural transformation, restless learning that appeared initially in England in 1750 and that remains unbridled in the twenty-first century. It is knowledge that first became magic in the transformation of goods, of atoms, and then flew towards the more sophisticated knowledge of algorithms and genes.

It took about a century and a half for Arnold J. Tonybee to define these in a famous lecture given at Oxford in 1886 and for industrial society to realize that it was at a stellar moment in civilization.

In the face of the Digital Revolution, the society of the third decade of the twenty-first century cannot wait so long to comprehensively and responsibly address the advantages and probable threats that the new Digital Revolution brings; in the same way, it should not minimize the strength and disruption of

2 Idem. p. 86.

this new course of scientific development, which, empowered by its experience of the last centuries, today appears more robust and challenging, demanding that human beings live up to their own inventions.

II.

The digital revolution, digital transformation, and its shining star, artificial intelligence, are not accidental events that have appeared without having been announced.

However, given its high impact and disruption at the beginning of this new digital era in the 21st century, it seems that we have been taken by surprise and that we lack a sufficient explanation to attest to its arrival, its nature, its benefits and its challenges.

The experiences and amazement of the last few centuries seem not to have been enough to prepare us.

A candid fascination runs through different societies, which look curiously, as people did in the case of electricity in the nineteenth century, wanting to touch, to see, to meet this new star character of human civilization, which, despite the recentness of its arrival, has managed to capture the monopoly of societal admiration at the moment.

As occurred at the World's Fair, entire legions of new followers are gradually approaching digital keyboards to admire the light that comes from the first responses as chatbots begin to interact with human beings.

From the simplest to the most profound, today's society looks out over the *Mimir Spring* to try to find through its waters, wisdom, wealth, the future, the solution to its problems, or the triviality of everyday life.

As it did in the past, like at a sacred sanctuary, 21st century society is gradually beginning to cross itself before a new altar of humanity that is quickly being installed in its daily life, profoundly transforming its person, its family, and all its social, political, and economic events.

III.

In 1956, John McCarthy enlightened us by identifying this new stage of knowledge as artificial intelligence. Norbert Wiener, Claude Shannon, Alan Turing, and many others also contributed to the emergence of this new tool.

However, it was with the appearance of the first Chat GPT in 2022 and other chatbots when it really became known that there is already a new star in the sky

which can interact at all levels of the human being, and that its potential not only has changed the industrial paradigm, but also has a technological power that could take humanity to the most stellar scenarios in its passage through the world. It could also mean humanity's extinction, or worse, its insignificance.

The early work on its identity is unclear. With our initial posturing, we move from fascination to fear and from fear to the fascination of the unknown, seeking to arrive at a platform of digital development which benefits the majority of the world's population.

In the context of AI's diffuse take-off, global society is just beginning a great debate on the scope of its dimensions, the nature of its behavior, the limits of its content, the commitment of its depositaries, and the regulations that must arise to guarantee its administration for the service of human beings so that AI's appearance does not accelerate inequalities, does not contribute to climate risk, etc. and so, in essence, it can remain a tool that will contribute to the better development of humanity and that its scope of freedom is guaranteed.

IV.

The main objective of this work, convened by the Institute for Industrial Development and Digital Transformation (INADI) and Research Legal Institute (IIJ) from UNAM, with the participation of more than 20 specialists from nine countries, is to contribute to these first efforts to learn about artificial intelligence, which has appeared in the middle of the third decade of the 21st century with all its potential and in all its magnificence, and demands all disciplines of knowledge to make an effort to try to start a new project of humanity.

Under this banner, this work consists of six chapters, acknowledging in advance that due to the dimension of the subject, it is impossible to generate in a single editorial effort, the knowledge to cover most of the issues this raises.

However, this work also does not shy away from the opportunity to explore some of the most important aspects related to AI, with the idea that essays by specialists from a multidisciplinary and transversal vision contained in this work, can help outline the general lines of the topic.

Thus, in Chapter I, with the title of, "Artificial Intelligence: Towards a new era of humanity", this book tries to make clear its position of being in front of a stellar subject of study, which will be defining a new era for humanity. In Chapter II, under the title of, "General Lines of a New Knowledge", this work seeks to offer an identification, trajectory, and classification of artificial intelligence in its first take-off stage of the Digital Era. Chapter III looks at economics and seeks to

participate in the identification of new possibilities for economic and industrial growth that are already emerging from the Digital Era and artificial intelligence, which already comprise 25% of the world's economy. Chapter IV addresses the complicated issue of the regulation of artificial intelligence through a series of essays whose contents illustrate the general lines of a broad debate that is just beginning, which very urgently needs to be resolved. Given its global nature, Chapter V explores the ways in which some of the world's most important economies are understanding the subject. Finally, Chapter VI, shows a small approximation of what artificial intelligence is already generating in the matter of new potentials in supervision and public security.

I am grateful for the professional and enthusiastic participation of all the specialists who kindly agreed to participate in the academic project presented in this book.

V.

Artificial intelligence is becoming stellar and omnipresent. The contours of its geography present no limits. It has risen from its scientific platform at high speed, orbiting in parallels that threaten to move away from human understanding.

Despite arriving at the beginning of a great disruption in the global, economic, social, and political environment of the century, the force of its arrival not only threatens to displace attention from these changes, but also to absorb and define them.

It is evident that a single branch of human knowledge will not suffice to explain AI. Its technological, social, economic, and political implications will require the participation of all disciplines of knowledge. In this sense, artificial intelligence is awaiting multidisciplinary reflections from philosophy, sociology, law, international relations, psychology, economics, politics, etc., where together, all contribute to discovering the genome of the phenomenon.

The great advances and benefits that artificial intelligence can represent for humanity are being detailed extensively and humanity will not give them up for a better world.

However, since its launch, different voices have also begun to emerge that warn of the dangers and threats that a misunderstanding, mismanagement, or dehumanized appropriation of the tool would represent, not only for the sustainable and equitable development of today's society, but also for the most precious thing that humanity possesses: its freedom.

The future that comes hand in hand with artificial intelligence is not a debate that can be summarized between technophiles and Luddites; surely it is much more than that.

It should be, among other things, the possibility of correctly guiding the use of AI, the best tool achieved by human beings, so that its use participates in the construction of a more just and sustainable world in the twenty-first century.

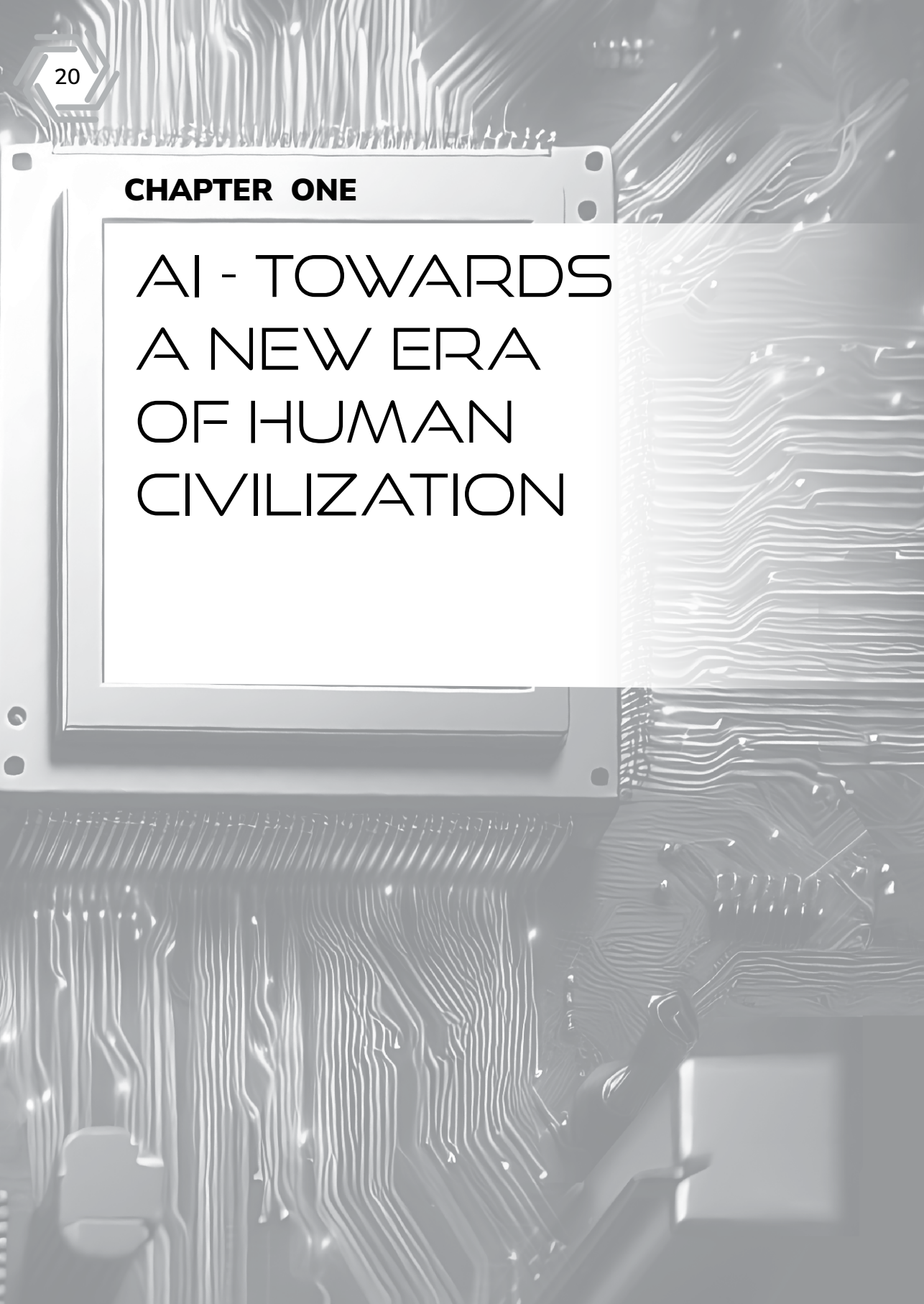
> ARTURO OROPEZA GARCÍA

Thus, at this turning point, we face the challenge of choosing between a future of unparalleled possibilities or one of unimaginable dangers. The fate of humanity hangs by a thread, and the decisions we make in the coming years and decades will determine whether we rise to meet these technologies or, instead, fall victim to their dangers. Yet in this moment of uncertainty, one thing is certain: the era of advanced technology has arrived, and we must be prepared to confront the challenges it brings.

(This text was written by an artificial intelligence).
Mustafa Suleyman, *The Coming Wave*, 2024

CHAPTER ONE

AI - TOWARDS A NEW ERA OF HUMAN CIVILIZATION



LIBERALISM, DEMOCRACY, AND ARTIFICIAL INTELLIGENCE: THE CHALLENGES OF FREEDOM

SUMMARY:

I. Introduction. **II.** Sick democracy? **III.** Exhausted Liberalism? **IV.** Liberalism, democracy, and algorithms. **V.** Towards a digital utopia. **VI.** Of Datalogies and Dystopias. **VII.** Regulate or not to regulate? That's the question. **VIII.** General reflections. **IX.** Bibliography

> ARTURO OROPEZA GARCÍA

President of the Institute for Industrial Development and Digital Transformation (INADI) and National Researcher at the Institute of Legal Research (IIJ) of UNAM

I. INTRODUCTION

Since the end of the twentieth century and throughout the present century, a clamor that has been increasing in intensity and bewilderment has been occupying the concern of specialists, politicians, and society in general.

The feeling that something is not right in Western politics has crossed the line of what is tolerable in many countries. There are many voices and movements that, in the midst of this concern, have begun the search for the causes of a *sick democracy*.

From different platforms, there is talk of *The Twilight of Democracy* (Applebaum. 2021); *How Democracies Die* (Levitsky and Ziblatt, 2019); *The Fall and Rise of Democracy* (Stasavage, 2021); *A Theory of Complex Democracy* (Innerarity, 2024); *The Edge of Chaos* (Moyo, 2018); *The People vs. Democracy* (Mounk, 2018). And many others tell us about a growing disease (deficit, recession, exhaustion, etc.) of the Western society of our time.

Another group of specialists do not limit themselves to only drawing attention to a democratic deficit. Broadening our gaze, they tell us about *Liberalism and its Discontents* (Fukuyama, 2023), *The Crisis of Democratic Capitalism* (Wolf, 2023), *Post-capitalism* (Mason,2015); *The rise and fall of the neoliberal order* (Gerstle, 2022), *A World After Liberalism* (Rose, 2021); [*Wounded liberalism] El liberalismo herido* (Lassalle, 2021), etc., emphasizing that the political problem of the twenty-first century is not limited to a deficit of democracy, but that the political disease of Western society escalates towards the philosophical, political, economic, and social explanatory model that guided it and gave it direction for almost three centuries.

On the other hand, in a small bibliography that is beginning to emerge, authors, such as Suleyman (*The Coming Wave, 2023*), Lassalle, (*Cyberleviathan,* 2020), Zuboff, (*The Age of Surveillance Capitalism,* 2019), Acemoğlu and Johnson (*Power and Progress, 2024*), Susskind (*The Digital Republic, 2022*), Noah (*21 Lessons for the 21st Century, 2018*), Sadin (*Artificial Intelligence or the Challenge of the Century, 2023*), Crawford (*Atlas of AI, 2021)* and others, begin the construction of a more pressing analysis of the new difficulties that are being added to the democratic disease and the crisis of liberalism in the face of the possible emergence of an "*infocracy*" (Byung-Chul Han)[1], a "*cyberleviathan*" (Lassalle), a "*capitalism of surveillance*" (Zuboff), etc., which, like a thousand-headed jellyfish, through its digital vine, is gradually covering the political, economic, and social life of the human being.

The challenges are not easy. Along with the usual endemic diseases and the recent ones of the late twentieth and early twenty-first centuries, typical of a socio-political process in permanent correction, a powerful digital force is emerging disruptively which questions today's political litigation, while at the same time, calls on global society to reflect on the emergence of the new technological era that, among its many manifestations, presents the possibility of politically disappearing the exercise of democracy and democratic society that contains it in the medium-term future, which is closer than people suppose.

Below are some general reflections on the subject.

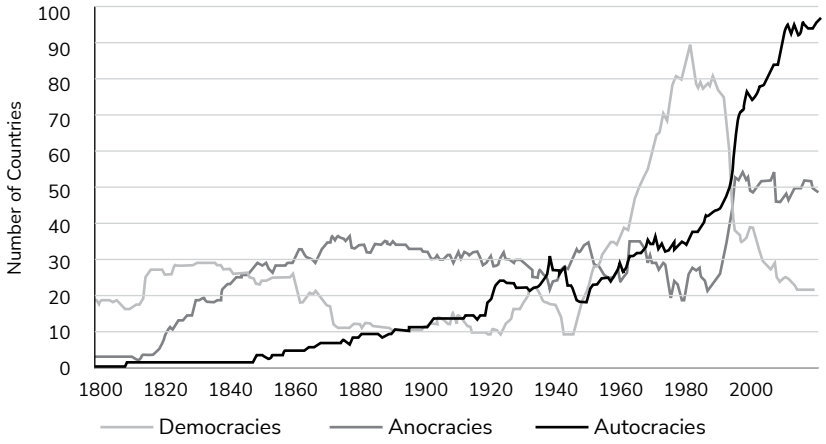
II. SICK DEMOCRACY?

On the horizon of human history, despite its recognized importance today, democracy appeared as a socio-political phenomenon of a new type, since in 1800 there still was no country that complied with democratic imperatives. One hundred years later, in 1900, there were only 12 nations that were classified as democratic, with only one of them, New Zealand, having universal suffrage. It

1 We call the "information regime" the form of a domain in which information and its processing by algorithms and artificial intelligence decisively determine social, economic, and political processes. Unlike the regime of discipline, bodies and energies are not exploited, but information and data are. The decisive factor in gaining power is now, not the possession of the means of production, but rather, access to information, which is used for psychopolitical surveillance and the prognosis and control of behavior. The information regime is coupled with information capitalism, which, today, becomes a capitalism of surveillance which degrades people to the condition of data and consumer livestock. (Han, Byung-Chul (2022), *Infocracia*, Penguin Random House, Mexico, p.9.)

was during the twentieth century, that is, barely 100 years ago, that democracy began to participate more actively in the political and social life of the West. In 1922, the number of democracies had risen to 24 nations, and in 1989, this had doubled to 48 countries, reaching to be part of the political reality of approximately half of the countries in the world by 2016 (97)[2].

FIGURE 1 Global Trends in Governance models



SOURCE: Center for Systemic Peace, Polity IV data.

Despite this apparent success, it is argued that since the beginning of the century, no less than 25 democracies have failed around the world, and as an example of what Europe, itself, suffers, these include Russia, Turkey, and Hungary[3].

Freedom House, for its part, reports that there have been 16 consecutive years of decline in freedoms; and that to date, 80% of the world's population lives in authoritarian or only partially free countries[4].

A study published in 2020 concluded that democratic citizenship amounted to one billion nine hundred million people, of which just over 2% lived in nations where 75% of their citizens were satisfied with their democracy. 21% lived in nations where one half to three-quarters were satisfied; and 57%, or the majority,

2 Wolf, Martín (2024), *La crisis del capitalismo democrático*, Arial, p.66.

3 Luce, Edward (2017), The Retreat of Western Liberalism, Atlantic Monthly Press, p.12.

4 Stiglitz, Joseph (2024), *The Road to Freedom*, Norton Company, p.1.

lived in countries where only 25 to 50% were satisfied with living in a democracy. The rest of the 20% lived with 25% or less democratic satisfaction[5].

According to Foa and Mounk, during the last three decades trust in political institutions, such as parliaments or courts, has drastically decreased in Western democracies. Voter turnout and trust in political parties have also declined; voters are increasingly inclined to extreme movements that vote for populist candidates or support anti-system parties, showing a sharp deterioration of Western democracy[6].

This feeling has been accompanied by an idea of defeat, of a move away from the proposal of a growing, shared, and sustainable standard of living, the creation of good jobs, equal opportunities, better security, greater economic and political freedom that would reduce inequalities and which would have greater control over large economic concentrations through a State that would provide what is necessary.

There is no doubt that democracy is sick. Where there is still no agreement is in the main nature of its suffering, and hence, the effort is to reconstruct social scaffoldings that have rusted with the passage of time.

Rose, for example, does not hesitate to point out that we live in a post-neoliberal moment and his underlying reasons are located in the loss of individual freedom, in the excessive participation of the government, and in the questioning of the market. It also speaks of a great disagreement on issues of race, religion, economics, and political strategy, adding the aspects of immigration, democratic change, economic growth, the collapse of religious authority, feminism, marriage equality, racial justice, multiculturalism, etc[7].

Under a more economistic vision where it is assumed that economic growth is the best adhesive of liberal democracy, Luce bases the political crisis on the economic stagnation suffered by the middle classes, whom he points out as the main victims of the bad economic performance since the 90's, causing what in France has been called the appearance of the *moyenne coaches* (the affected middle classes); in England this is known as, the *left behind*, and in America, the *squeezed middle*; grouping all of these under the term, *precariat*. He recognizes that modernity has reduced the poverty of humanity, but reiterates that the middle class, between half and two thirds, has been sacrificed[8].

5 Wolf, Martín (2024), *Op. Cit.* p. 71.
6 *Ibid*., p. 70.
7 Rose, Matthew (2021), *A World After Liberalism*, Yale, pp. 2-4.
8 Luce, Edward (2017), *Op. Cit*., pp.12-13.

Mounk also presents a series of considerations based on the idea of the indissolubility between democracy and liberalism, where the weakening of one of their assumptions means the thinning or breaking of the other. He agrees that the economic component is the main cause of Western democratic weakness; arguing that citizens enjoyed a rapid increase in their standard of living after the Great Depression and until the 1980s, where, for example, from 1935 to 1960, the income of a typical American household doubled, and from 1960 to 1985, when it doubled again.

Secondly, he places the racial or ethnic component, which in the face of the freezing of the economic rise of the middle classes, has translated into the fear of losing social hierarchy, which is generating a Western rebellion against ethnic and cultural pluralism.

However, since that date, the average income in North America and Europe stagnated, creating a zero-sum idea, in which any economic disruption is borne by the middle class.

As a third factor of the democratic crisis and only in a referential way, he includes the rise of the Internet as one of the factors of instability, separating the distance between the center from the periphery[9].

From diverse positions, it seems that the degree of concern about the democratic malady is inversely proportional to the degree of agreement there may be on the central issues may be weakening it. This, thereby, is generating a multiplicity of ailments, possibly all real, but which hinder the progress of democratic debate, provoking a broad-based horizon of discussion that leads to the adoption of partial solutions which have not helped much to improve its operation.

At the same time, what can be seen from the above is a series of atomized ailments, typical of a Western society of demand, whose incessant requests, rather than the actual causes for its problems, distance itself from their possible solutions.

The feeling is growing, says Lassalle, that liberal democracy is a failed model, that the world's leading states have not restored security, and that cognitive capitalism has not brought well-being for all in the context of a society exhausted and resigned to adversity. The old dilemma of security or freedom is now sharpened and even exacerbated to where the conflict is now between order or chaos where liberal democracy has lost strength in the face of the postmodern depths of an ungovernable society[10].

9 Mounk, Yascha (2018), *El pueblo contra la democracia*, Espasa, pp.22-24.

10 Lassalle, José María (2021), *El liberalismo herido*, Arpa, pp.12-13.

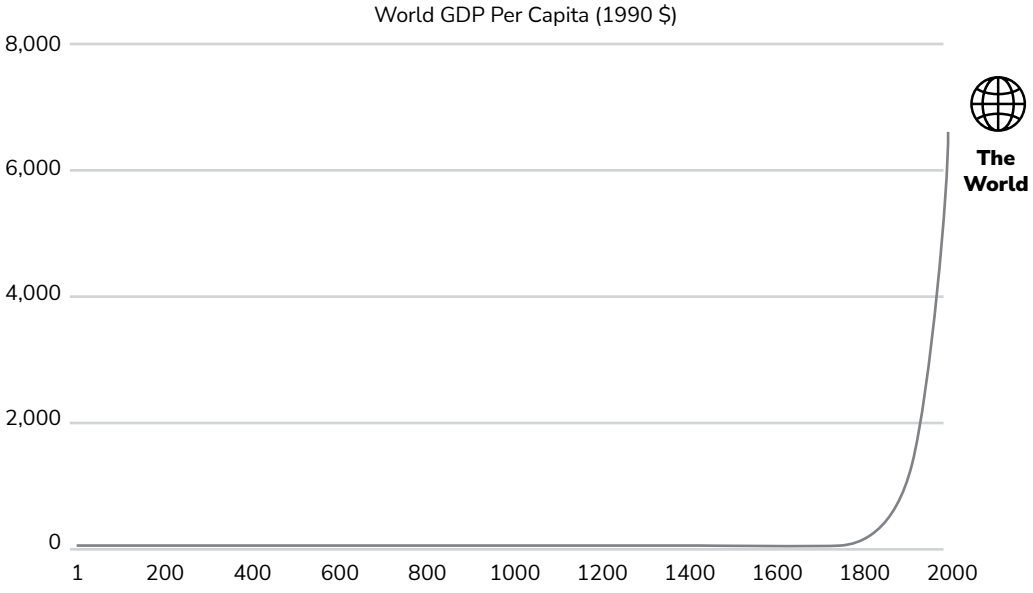
There is an overflowing concern for resolving the democratic milieu that does not work well for most Western societies with respect to a series of ailments that, to a greater or lesser extent, have been accumulating since the birth of liberalism but the debate has neglected to reach the very heart of the liberal order, which deserves an exhaustive review.

III. EXHAUSTED LIBERALISM?

The final part of the seventeenth to eighteenth centuries signified a stellar moment in Western history, where a cosmogony of a liberal nature appeared simultaneously together with the First Industrial Revolution; together these changed the experience of being human forever.

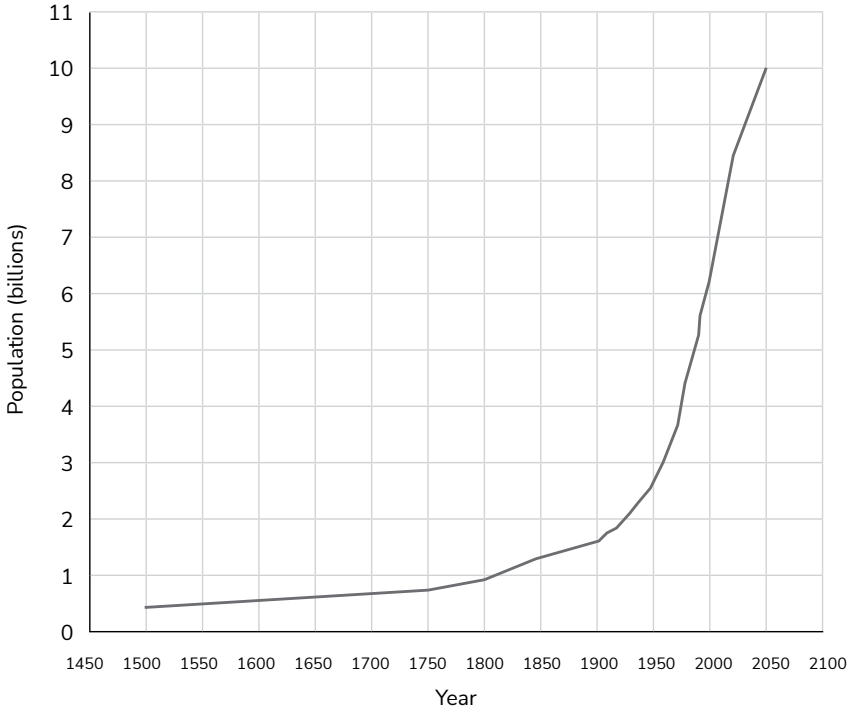
Before the eighteenth century, despite pre-modern epics, human experience had been a long plain where what anyone did anywhere had little general impact on the world's population.

FIGURE 2 World GDP per capita (0-2000)



SOURCE: "Statistics on World Population, GDP, and Per Capita GDP, 1-2008 AD", Angus Maddison; IMF

FIGURE 3 Evolution of the world population (1500-2050)



SOURCE: Research Gate

It was the industrial revolution that allowed the birth of world trade, manufacturing, increases in competitiveness, the multiplication of innovation, the growth of the world economy, and especially, the discovery of a scientific and technological potential that marveled the world of its time through its inventions.

On the other hand, the appearance of lucid and enlightened thought led by Locke, Hobbes, Rousseau, Smith, among many others, created broader reflection so that their proposals of social contract, freedom, rule of law, human rights, and the market, among other things, had the virtue of leading and supporting industrial and technological innovation, which found in liberalism an explanation of life that gave these meaning and direction.

The liberalism of the eighteenth century became a summation of the principle lines of thought that since Greek times had tried to dignify being human in society. Liberalism's proposals for the creation of a modern state, of social community based on law and freedom, gave rise to a general theory of development that was imposed in most Western countries on the liberals' own terms.

Throughout these last centuries of liberalism, different trends have emerged: social liberalism, libertarianism, neoliberalism, etc. However, classical liberalism, Fukuyama adds, is a great umbrella under which a wide range of political positions are sheltered that coincide in terms of the fundamental importance of the equality of individual rights, law, and freedom[11].

From its birth, with industrial technological innovation, liberalism was committed to a utopian version of a free society where the majority of its members deserve progress through the organization of the Rule of Law based on the will of the majority in a liberal democracy.

The powerful force of liberal thought served as a touchstone of economic liberalism which became a constant for the majority of Western nations of the time, countries which were implementing capitalistic manufacturing, an emerging powerful secondary sector that rapidly displaced the agricultural economy.

This economic liberalism, with its own contradictions and excesses, intensely experienced the social adjustments that were appearing in the second part of the nineteenth century and that exploded in the first half of the twentieth century.

After this convulsive stage of political, economic, and social rearrangement, from 1950 onwards, the West experienced a second very successful stage where global trade had growth rates of approximately 7.4% per year on average and its economy's growth rate as a whole was 5% per year on average until the mid-70s (Ocampo, 2017). With this, the world was able to demonstrate to itself that global trade and economic development could harmoniously coexist with, in an important manner, the improvement in the social conditions of the people who made possible the progress of industry and manufacturing in the world.

However, the crises that occurred in the 70s stopped the advancement of these achievements. In the context of the increase of conflicts about oil, raw materials, and food, among other things, these crises opened the door for the introduction of economic currents such as neoliberalism, which in line with its postulates, expelled the State from its responsibility as the orchestrator of economic development, condemning it to the role of witness to economic events.

Margaret Thatcher in England, Ronald Reagan in the United States, and others in most Western countries and their periphery, under various adaptations, oriented their economic strategies towards the neoliberal lines dictated by Washington's consensus, which, like a miraculous recipe, was offered to

11 Oropeza, Arturo. (2024), *Hacia una nueva visión del desarrollo industrial digital en México*, p.39.

countries for them to get out of the trap of underdevelopment to reach the average income level of developed countries.

Since the late 1970s, Fukuyama acknowledges, economic liberalism has evolved into what is now called neoliberalism, which dramatically increased economic inequality and led to devastating financial crises that have hurt ordinary people far more than the wealthy elites in many countries around the world. This inequality, the same author adds, is the basis of the progressive argument against liberalism and the associated capitalist system. This inequality, this disenchantment with economic liberalism has evolved in recent decades, generating demands from both the right and the left, asking that this doctrine be replaced at the root by a different system[12].

Despite the fact that there is a generalized idea of the failure of neoliberalism, the debate on its origin continues in a recycling of philias and phobias in relation to the role that the State should have in economic development.

The successors of Thatcher and Regan who appear in this new debate include Meloni of Italy, Milei in Argentina, etc., the latter, who on his last trip to Europe, declared that "La justicia social es aberrante (Social justice is aberrant)", "el control de capitales es inmoral (Capital controls are immoral)", and "los impuestos a empresarios son un robo (Taxes on businessmen are theft)" (El País, May 18, 2024).

In contrast to authors who recognize the failure of economic neoliberalism, such as Mazzucato, Rodrik, Fukuyama, himself, etc., there are authors such as Easterly, Couchrane, McCloskey, among others, who continue to defend the neoliberal current. McCloskey, for example, while criticizing Mazzucato's ideas about an entrepreneurial state, adds that governments tend to make stupid decisions and that genuine innovation emerges best in a free-market environment and that governments should not take a central role in running the economy. Easterly reiterates these positions by invoking Adam Smith, pointing out that economic development is based on free trade and the free market.

The costs of the mistakes made in this battle for development have gradually been recognized by Western political leaders, such as in the case of Macron (France) when he declared that "French deindustrialization is one of the causes of our misfortune;" in the case of the United States, when Biden talked about the country's industrial oblivion[13], and in the case of Trump, when he paranoidly attacks the world to recover an *alleged stolen development*.

12 *Ibid*., pp.40-41.

13 *Ibid*., p.42.

The economic failure of the neoliberal model is recognized in the face of the dissatisfaction of the West's middle and working classes. But also, with the geopolitical and economic competition which compares the West's performance to the successful rise of China and the East Asian region, we should now add the appearance of a new digital capitalism committed to neoliberal economics' same essential lines.

After almost three centuries of being presented as a virtuous and convincing account of Western cosmogony, liberalism, through its different manifestations and schools, now faces one of the most important challenges in the history of humanity, ranging from the possibility of rethinking itself through its original concepts and values, to the point of its own disappearance.

Liberals, Noah points out, do not yet understand how history deviated from its predestined path and lack an alternative prism to help them interpret reality[14].

In a vision that does not hesitate to draw complex scenarios for liberal society, Innerarity argues that we understand very little of our time; that we don't understand how this society works; nor what our possibilities of action within it are. It is argued that the simplification of some analyses come from the lack of updating of political concepts that were conceived in an era of relative social and political simplicity, before the great social conflicts within societies that inaugurated the contemporary world, and which did not yet understand about current cultural, political, and technological pluralism[15].

Democracy and liberalism, categories that were born together over time, also jointly face political, social, economic, and technological changes which have presented themselves without foreshadowing, and Western societies find themselves devoid of the instruments which would allow them to propose solutions to preserve the essential elements of liberalism, such as economic development, security, and freedom, itself.

A belief in the end of history, supported by the fall of the Soviet communist narrative in 1991, led to a complacency of belief that liberalism would rise throughout the world as a universal truth that would lead humanity to progress and stability.

The West forgot that liberalism, like any narrative that is the product of social construction, depends on remedies that need to be opportunely injected into it for its subsistence.

---

14 Noah, Yuval (2018), *21 lecciones para el siglo XXI*, Debate, p.23.

15 Innenarity, Daniel (2020), *Una teoría de la democracia compleja. Gobernar el siglo XXI*, Gutenberg, p.13.

The hasty appearance of its neoliberal aspect in the 80s destroyed part of its balances and postulates, causing an increase in economic disparities between the different social segments, both in central Western democracies and in their peripheral countries.

In this regard, Zuboff tells us about a collision that has occurred after decades of a neoliberal market economy regime, between the historical processes of individualization that make up our experience as autonomous beings in which the awareness of our own personal worth and our needs for autonomous individual self-determination have been systematically boycotted by the other[16].

The focus of this essay is the failure of neoliberalism, with an orientation to the economy which includes the dogma of non-regulation, which distances the State from its responsibility to be the orchestrator of development, leading to a decrease in the State's power, which is taken over by private market forces to the detriment of the checks and balances that were always thought of as the State or its Rule of Law.

To this weakness of Western neoliberalism should be added the untimely arrival of the First Digital Revolution, which had been forcefully foretold since the first decade of the 21st century. As Noah pointed out, with its arrival, this latest revolution has demonstrated that Western society was not prepared to administer a change of such dimensions, which is putting liberalism against the wall, and endangering its very survival.

In the dogmatic line of neoliberalism, which developed in the last half century, the idea of a powerful market and a weak state has prevailed. Along with this emerged a *new data market* led by *Big Tech*, which during its birth found an environment of total permissiveness without adequate regulatory frameworks. Perhaps, more importantly, this new digital market has formed without having considered and formulated, as the liberal philosophers of the seventeenth and eighteenth centuries did, the original questions of what we want to accomplish as a society with the new digital civilization, and what place we want to give to the human being throughout the present century and into the future.

Lassalle asserts that "[under this panorama, society has become ungovernable, as the neoliberal imperative, assumed by the cognitive capitalism of the large digital platforms and big data, dissolves more and more of the consensual bases of civic and contractualistic rationality of Lockean liberalism, leading us to a pure form of algorithmic domination without contestation or dissidence.]

16 Zuboff, Shoshana (2020), *La era del capitalismo de la vigilancia*, Paidós, pp.34-35.

Bajo este panorama la sociedad se ha hecho ingobernable, pues el imperativo neoliberal asumido por el capitalismo cognitivo de las grandes plataformas y los datos disuelve más y más las bases consensuales de racionalidad cívica y contractualista del liberalismo de estirpe lockeana y nos aboca a una forma pura de dominación algorítmica sin contestación ni disidencia"[17].

IV. LIBERALISM, DEMOCRACY AND ALGORITHMS

Since the end of the last century, global society as a whole, propelled by the logic of its own narratives, has been facing a civilizational disruption caused by the technological advance of the First Digital Revolution which, like its predecessor, the First Industrial Revolution, comes challenging everything known, questioning each societal structure and, once again, confronting humanity with a changing sense of its future.

The subjugation caused today by countless digital technologies which are changing the lives of human beings with dizzying speed, is only comparable to the commotion caused to society around 1750 by the appearance of industrial innovation.

In the West, we are facing the emergence of a new digital era, of new digital power, which, through the force of the algorithm, threatens with amazement, and promotes the civilizational advance of humanity, along with controlling its destiny.

The Digital Revolution is many things. It is a digital transformation of the intangible, of bytes and genes. It includes the arrival of new 5G technology with its speed accelerators and latencies in milliseconds; with an increase in the data rate to high gigabytes per second; with broadband thousands of times faster, etc. which, when translated into innovation, accelerators, overflows into cloud computing, data centers, big data, the Internet of Things, robotics, cybersecurity, biotechnology, etc. And as the pinnacle of this geometric development, artificial intelligence now is positioned for the evolution and destruction of its own creators.

Artificial intelligence as a name and concept is not new. As early as 1956, John McCarthy coined the name and developed the first artificial intelligence programming language. In 1950, the first computer was manufactured that

17 Lassalle, José María (2021), Op. Cit, p.22.

allowed the game of chess to be played between humans and machines; in 64-66, *Eliza* was the first Chatbot to simulate a conversation between humans and machines; in 1997, the supercomputer *Deep Blue* defeated the world chess champion, Gary Kasparov, as a premonition of what artificial intelligence is now becoming; and in 2016, *Alpha Go* defeated the world champion of that specialty with unpredictable moves that denoted the success of digital intelligence systems (Vanguardia 2020).

In terms of investments, artificial intelligence will reach $200 billion by 2025 and is expected to scale to $2 trillion by 2030, estimating that generative AI systems and applications alone will produce $7 trillion in revenue by 2030[18].

The fascination and horror caused by artificial intelligence today has come a long way since its military origins, where, since the end of World War II, the Pentagon had been economically and technologically supporting most of the innovation that was on the market[19].

Artificial intelligence is reaching a first stage of maturity, thanks to its technical improvement and the accumulation of data that has been produced on social networks, which have allowed it to scale in information and power, causing a division of society between those who euphorically find the outcome of a digital future that they do not know, exciting, and a Luddite society, afraid of not knowing what is coming next.

The owners of the new weapons of digital power, such as Mark Zuckerberg, the owner of Meta, defend whatever new technology comes to change the nature of things, claiming there are always people who worry about such changes and want to return to the past. Likewise, entrepreneurs in the sector say that most people's fears about artificial intelligence are exaggerated, if not totally unfounded, and that people won't cease being afraid of it until such technology is completely incorporated into people's daily lives[20].

Others, like Lassalle, leading the radical line of the current who warn of an enormous threat to humanity in the face of the management of an artificial intelligence without regulation, indicate:

> "(*One figure stands out on the horizon of uncertainties, discomforts, and fears which accompanies the beginning of the 21st century. It is for now a*

18 López-Portillo, José Ramón; Blog Nexos, July 2024, p.29.

19 American artificial intelligence agencies, Crawford points out, have been the old gatekeepers of big data, and through large defense research projects, they have been the main drivers of artificial intelligence since the early 1950s. (Crawford, Kate (2021), *Atlas of AI*, Yale. p.184).

20 Acemoglu, Daron and Johnson, Simon (2023), *Power and Progress*, Public Affairs, p. 373.

silhouette to be defined. An image that still does not accurately reflect its contours but that projects a concern in the environment that prevents us from facing it. Its appearance betrays a vigorous uplift movement, which raises it above the surface of the events that accompany us throughout the passage of the new millennium.

Enveloped by a breath of boundless energy, its form acquires titanic volumes in which the granite nakedness of a new expression of power is sensed. With its gestures, the political reign of a world devoid of citizenship, without rights or freedom, is announced. An era that will witness the extinction of liberal democracy. That will establish a mythical era in the manner of those imagined by Hesiod, made of surveillance and silicon, inhabited by a race of humans subject to order and security. A world of fibre optics and 5G technology, dominated by a post-human vision, which will overflow and marginalise the concept we have had of man from classical Greece to the present day.) Una figura destaca sobre el horizonte de incertidumbres, malestares y miedos que acompaña el comienzo del siglo XXI. Se trata por ahora de una silueta por definir. Una imagen que todavía no refleja con exactitud sus contornos pero que proyecta una inquietud en el ambiente que nos previene frente a ella. Su aparición delata un movimiento de alzada vigorosa, que lo eleva sobre la superficie de los acontecimientos que nos acompañan a lo largo del tránsito del nuevo milenio.

Envuelta por un aliento de energía sin límites, su forma va adquiriendo volúmenes titánicos en los que se presiente la desnudez granítica de una nueva expresión de poder. Con sus gestos se anuncia el reinado político de un mundo desprovisto de ciudadanía, sin derechos ni libertad. Una época que asistirá a la extinción de la democracia liberal. Que instaurará una era mítica a la manera de las que imaginó Hesíodo, hecha de vigilancia y silicio, habitada por una raza de humanos sometidos al orden y a la seguridad. Un mundo de fibra óptica y tecnología 5G, dominado por una visión poshumana, que desbordará y marginará el concepto que hemos tenido del hombre desde la Grecia clásica a nuestros días"[21].

21 Lassalle, José María (2019), *Ciberleviatán*, Arpa, pp. 17-18.

> Geeks and Luddites; technophiles and technophobes, are divided by the arrival of an AI where the former see it as the holy grail that will solve all the world's problems, while the other group appreciates it as an enormous challenge that, as already pointed out, threatens not only the liberal and democratic system, but also attacks the work and way of life of the human being, condemning it to insignificance.

At this point, it would be worth remembering Suleyman when he warned against falling into the *trap of aversion to pessimism*, which arises when one feels overwhelmed by the fear of facing realities that could be dire, and so, looks the other way[22].

It is in this environment of amazement and fear that the greatest threat that has been attempted against the freedom of the human being and, therefore, of its institutions such as democracy and liberalism, itself, has been sneaking in without much fuss.

Protected by the infinite democratic debate, digital reality advances faster than people realize, silently entering the will of today's society while society discusses an endless agenda where all kinds of issues of an economic, political, and social nature appear.

Seminars and multiple debates are opened on whether restoring the economy can improve democracy; if, by modifying once again the antiquated nature of its forms, democracy can be improved; whether, including or not including the issues of a number of minorities, we could arrive at a better stage of political freedoms, etc., while seasoned political opportunists, warning of the anarchy of the moment, offer an infinite number of improvements that they will not be able to fulfill, and are building various forms of populism from both the right and the left, which in the first instance, present themselves as the winners, gaining because of all the bewilderment.

However, beyond the origin of each of the rights that are intended to be claimed in order to have a better democracy; at the same time, a darkness has been advancing that is gradually creeping into hundreds of millions of human beings; a panoptic world disguised as a utopia that is progressively diminishing both the spheres of freedom of the human being and the strength of the Western State.

22 Suleyman, Mustafa (2024), *The Wave That Is Coming*, Debate, p. 22.

V. TOWARDS A DIGITAL UTOPIA

The beginning of the construction of digital power in its political version is recent and has occurred together with the advancement of its own development and evolution witnessed throughout this century.

Although the appearance of Microsoft (1975), Apple (1976), Nvidia (1993), Amazon (1994), Google (1998), Alibaba (1999), Tesla (2003), Facebook (2004), etc., come from the last third of the last century, the speed of their changes have quickly led them to unfold into new technologies that have changed the logic and meaning of their businesses.

For the Industrial Revolution to go from its first stage to what was called the Second Industrial Revolution, which was identified with the birth of Fordism, 150 years had to pass. In the case of the Digital Revolution, the trigger of artificial intelligence, as the epitome of its innovation, has occurred in a space of approximately six decades.

The tyranny of Moore's Law and the forcefulness of Koomey's Law in the twentieth century, which have been proven in the real world, indicate that humanity would lower digital costs every 18 months while increasing its speed and storage potentials; together with Kurzweil's Law of Accelerated Returns, which predicts that by 2050 artificial intelligence will develop in such a way that the computing capacity of computers will exceed by more than a thousand times the computing capacity of a human brain, these have led us to the third decade of the century to live scenarios that we would never have imagined.

The speed of these results has surprised everyone. From the owners of the technology companies themselves, to the various analysts, and of course to the neglected State that, preoccupied with a story and a problem typical of the last century, does not quite know about the new structures of a civilizational change that threatens to transform what is known.

In the case of the United States, which has always been a leader in digital progress as a result of post-war scientific and technological advances, aware of its importance in geopolitical and economic confrontation, it has to date maintained a monopoly on digital innovation in strategic partnership with its private sector. China, Russia, Israel, and other countries have also made significant progress in what Crawford calls *the artificial intelligence war*.

However, the success obtained by these technology companies has been of such magnitude that their power has led them to become true economic giants by market capitalization, where Apple ($3.33 trillion USD), Microsoft ($3.11 trillion USD), NVIDIA ($2.81 trillion USD), Alphabet ($2.03 trillion USD), and Amazon

($1.88 trillion USD) present themselves today as the five largest companies in the world[23].

This new digital economic power has led to these companies being more economically powerful than countries such as Spain, Brazil, Australia, Russia, etc., and they exceed each of these countries in economic value.

Digital business was strongly born at the end of the twentieth century, and the speed of its dividends has earned them new economic-political power. These dividends arise from the power of digital innovation that places them on a monopoly plane with respect to other economic actors in the world. In the case of Windows, for example, it controls 73% of the global market for computer operating systems; in terms of operating systems for cell phones, Android also dominates 73% of the world market; in the use of data cloud storage, Amazon (AWS), Microsoft (Azure) and Google (Cloud), control 70% of the world market[24].

Internet business by its very nature was born successful. Since its inception, it was received with great expectations and its technological innovations have been celebrated by the global society as a whole. However, the biggest change effected with the inception of digital business is the enormous compilation of data that was taking place as a natural result of the use of different technology devices, which by 2021 already accumulated an estimated number of 1900 billion websites[25].

The appearance of the first social networks led to the belief, as Hillary Clinton pointed out in 2010, that these were the new pillars of world democracy. One of the leaders of the Egyptian protests of the so-called Arab Spring, Wael Ghonim, also commented that he would like to meet with Mark Zuckerberg to personally thank him for the participation of these networks in these movements, which mostly began through Facebook, adding that if you wanted to liberate oppressed societies, you should just provide them with the Internet[26].

However, this first idea of freedom multiplied through the Internet as an instrument for liberation, has gradually been changing in the face of other types of experiences that have also been presented on the horizon of digital technology.

23 State Counter, El País, July 27, 2024.

24 Market Cap. El País, July 27, 2024.

25 As for artificial intelligence, it began its training with hand-labeled data, as in the case of ChatGPT 2, with 1500 million parameters. Thanks to the accumulation of data, by 2023, without labeling the data by hand, the Inflection AI application already had 540 billion parameters. In the case of Alibaba, it is said that it already has 10 trillion parameters.

26 Acemoglu, Daron and Johnson, Simon (2023), Op. Cit, p. 342.

The Myanmar case of 2010-2017, as one of the first events where the disinformation spread over the Internet, exacerbated a racial sentiment against a small Muslim sect of the Rohingya, led to a social problem that has been classified as genocide. In this event, it was evident that this tragedy had been fomented by social networks and that even within these, crimes were organized through the Facebook platform. Given these facts, the only thing the platform dared to comment on regarding the massacre was that it agreed that it should have done more to prevent it[27].

Such examples, which begin to show the other side of the *technological phenomenon*, are expanding in the face of the growth and power of social platforms and are rapidly moving away from their potential to promote freedom in favor of democracy and are getting closer to a world of directed or assisted freedoms which are emerging from the power of the platforms themselves.

In 2018, Sri Lanka continued Myanmar's precedent through hate speech against Muslims. In 2020 in India, assassination events and destruction of mosques against the Rohingya Muslim sect also occurred.

The world was able to get a clearer idea of what digital power was beginning to represent when the company Cambridge Analytics publicly acknowledged, together with Mark Zuckerberg, the owner of Facebook, that it had improperly bought the information of more than 50 million users of that social network in the United States for political and commercial purposes, affecting among other things, the elections where President Trump emerged as the winner.

From that moment on, it was clear that a world of surveillance of social data was opening up, that powerful people had found an ideal means to disseminate false facts in a faster, deeper, and more general way than the truth, itself, could disseminate in different categories of information. It also has been possible to establish that social networks are now the primary source of information in the United States and that 70% of what people read is being sent in a directed manner by these social media platforms and that out of 100 publications or recommendations that were made on Twitter, Facebook, etc., most have contained misinformation or conspiracy theories[28].

It wasn't just the unstoppable speed of gathering hundreds of millions of users on social media that began to whet the appetite of big tech companies, which began to see their millions turn into trillions of dollars with great speed. The impact has been to see how, in such a short time, they acquired a huge

27 *Ibid*., pp. 356 and 357.
28 *Ibid*., p. 359.

archive of personal data that they had not contemplated and that now represents great informational power for each of them.

In this sense, the transition from the Internet to data and then to artificial intelligence was a dialectic of information which has been turned to pure profit and has now become a political-economic power of great dimensions for Big Tech companies.

Larry Page and Sergey Brin were some of the leading specialists who found the best combination to combine technology with human data, creating the algorithms that were able to successfully link the two categories. Ultimately, as Acemoglu points out, how the algorithm calculates results is secondary, the important breakthrough is that Page and Brin devised a way to use human knowledge and ideas by improving a key task of the machine: ranking search results[29].

The monetization of data, which gave birth to this new business model provides the possibility of building true digital power. This arose at the time when Google initiated the logarithmic innovation of commodifying this new data technology through the important concept of **targeted digital ads**, which came to revolutionize the very progress and trends that, until that moment, had been tracking the First Digital Revolution.

That is why in 1998 and even in 2000, as Acemoglu commented, almost no one had a clear idea about what big data and artificial intelligence were. The new association between algorithms and data, translated into the expansion of targeted ads became a self-generating and continuing explosion of data which has served as a laboratory for the perfection of a generative artificial intelligence that can technically assimilate the history of humanity in half a day and today is able to generate what was previously considered impossible.

Acemoglu adds that we could not understand the political imbalance that social networks would cause if we did not add the excessive economic incentives based on targeted ads, which derived from the digital potential to optimize the tastes, ranks, faces, likes, and now, even the future of global society[30].

This whole event involving the phenomenon of digital data is evolving, as Zuboff indicates, into a new surveillance capitalism that emerges from the power that Big Tech is accumulating very rapidly[31], which surpassing the traditional

29 *Ibid*., p. 365

30 *Ibid*., p. 371.

31 Facebook (or Meta, as it's now known) and Google (or Alphabet, as it's now known), have little in terms of physical facilities, but have great systems and processes for collecting, storing, and processing data to connect people to knowledge, analytics, and each other.

ambitions of capitalism, now claims dominion over human, social, and political territories that extend far beyond the conventional realm of private enterprise or the market in an overthrow of the sovereignty of the people. This constitutes a dangerous drift towards the *deconsolidation* of democracy that currently threatens the liberal systems of the West[32].

Surveillance, lies, disinformation, the manipulation and expropriation of the will, the induction of economic, political, and social desire, and digitally created unemployment[33], are the doors that open to infinity in the face of this new impulse of digital change.

They are natural global monopolies, not in the traditional multinational sense of having large physical plants scattered around the world, but in dominating markets where customers benefit from being globally connected. It is their complex networks of people and information, not their physical infrastructure, which competitors cannot replicate or replace (Mayer, Colin (2024), *Capitalism and Crisis. How to fix them*, Oxford, p.8).

32 Zuboff, Shoshana (2020), *Op. Cit.* p. 38.

33 *"Sooner or later", says Skidelsky, "we're going to be out of jobs" (Carr, Nicholas, *Trapped. How Machines Take Over Our Lives*, 2014, p.47). The issue is not new, it was born together with the generalized threat of unemployment that arose from the First Industrial Revolution, when the first machines came to replace much of the physical work in the fields and workshops at the same time that millions of jobs were created in the new manufacturing industry and its services. In the nineteenth century, David Ricardo (1817) foresaw that a better distribution of work through machines would shift human activity towards science and art. Douglas Haberler and others acknowledged the possibility of digital unemployment but added that it would be automatically compensated. (Acemoglu, Daron, (2023), *Op. Cit.* pp. 434-435). In 1930, Keynes warned that "we were afflicted" by a new disease of which some readers may not have heard the name, but of which they will hear much in the years to come, namely "technological unemployment". On the other hand, Difazio, Rifkin, Brynjolfsson, McAfee, Spence, etc., in recent decades have been taking up the subject with new arguments. McAfee, for example, shows how technology has been rapidly increasing U.S. productivity, along with corporate profits, while the total number of jobs in the country has barely grown. Spence adds, "The replacement of traditional manual jobs with machines and robots is a strong, permanent trend that may be accelerating manufacturing and logistics, while computer networks replace administrative jobs in information processing". On the subject, Krugman adds that intelligent machines will be able to increase GDP, but at the same time, they will reduce the demand for people, including intelligent people while most of the profits will go to the owners of the robots, that is, Big Tech. (Carr, Nicholas, p.40-45). Rifkin, in his classic book "The End of Work", in addition to denouncing a global unemployment of 800 million human beings, added that "millions of workers have already been eliminated from economic processes and different specific types of employment have been reduced in importance or restructured or have disappeared". (Rifkin, Jeremy (1996), The End of Work, Grupo Planeta, p.17).

VI. OF DATAOLOGIES AND DYSTOPIAS

Since Aristotle's *zoon politikon*, each of the societies of his time started from the consideration of a human being as the center of a society that in different ways had to be built for humanity's well-being and happiness.

It is true that in many stages of its history, human beings have failed to achieve these goals, but their longing for freedom and a better life have always been present along this path.

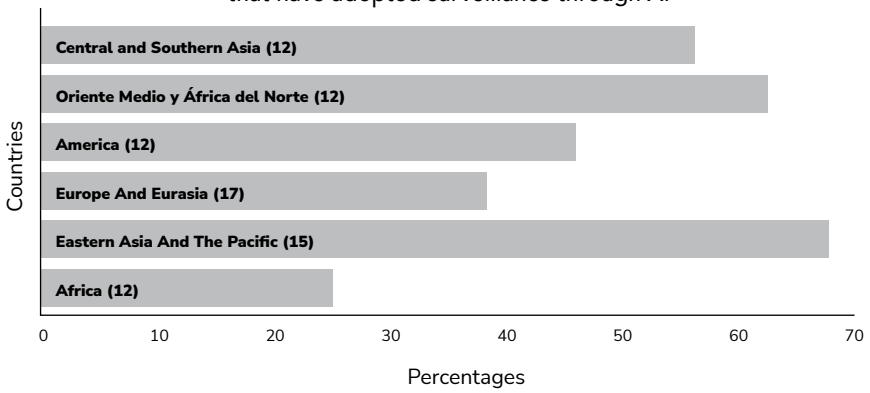
In this last stage, which has been a mixture of liberalism and the Industrial Revolution, the repeated search for freedom and well-being have been the main drivers in the political construction of a rule of law defined through democratic participation with the idea of an industrial society that felt that the progress that had come, would last forever.

This utopia of freedom and progress faces a crucial moment in which a new cosmogony without authors, but with the solid argument of facts, for the first time in the history of humanity, seeks to penetrate the sphere of freedom of the human being, whatever it may have been, and is gradually influencing ideas, tastes, amusements, preferences, values, jobs and the vision of the future, through an assisted freedom that, little by little, comes to tell people what their behavior should be and what their role should be in a new society of surveillance.

In this sense, it seems that the role of the human being is beginning to be reduced primarily to a user-customer equation, where through people's own data, their social, political, and economic behavior are surveilled, they are misinformed, lied to, manipulated, caused to change their behavior and preferences, and even become unemployed. In the end, they are placed in a state of uncertainty and deculturation that risks turning humanity into a state of robotization or anxiety, or worse, of insignificance.

In the background, rather than a human being, recognized for his/her values and defects, people are being considered as raw material of the new mechanism of digital production, which offers people to the market as just another commodity to be consumed for economic, commercial, political, and social purposes.

FIGURE 4 Percentage of countries in each region that have adopted surveillance through AI



NOTE The numbers in parentheses indicate how many countries have adopted AI surveillance in a region.
SOURCE: Vanguardia 2020

In the synthesis of this equation, what is generated is a new type of economic, political, and social power where, in the markets of human data, buyers acquire profiles, photos, behaviors, audios, thoughts, beliefs, tastes, etc., which, like the new inputs of a new digital production project, are already being used to increase product sales and the profits of companies, to shape social and political behaviors, and even worse, to inhibit the political expectations that were created over centuries of civilization.

This new parasitic reality that feeds on the behavior of the human being, directly threatens the existence and power of the State, itself, and, consequently, of its democracy and freedom through the expropriation of the human will.

Zuboff says that as competition intensifies, surveillance capitalists learn that it is not enough to extract human experiences. The most predictive stock of raw materials are those that come from intervening in our experience to shape our behavior in such a way as to favor the commercial results intended by these capitalists.

For this reason, Big Tech designs new algorithms to influence human behavior and, in turn, modifies it, while the means of production are subordinated to new and more complex means of behavioral modification[34].

34 *Ibidem*, p.36.

This new power, Zuboff continues, aspires to organize, stimulate, and refine society so that it reaches an analogous social confluence in which group pressure and computational certainty replace politics and democracy, and thus, both the felt reality and the social function of an individualized existence are extinguished[35].

We are not their customers, we are their *products*, says Noah, in the framework of a race for data that has already begun on the part of the Big Techs such as Google, Facebook, Microsoft, Baidu, etc., which first capture our attention by providing us with information, services, and fun for free, and then, resell our attention to advertisers. The real deal is to accumulate an immense amount of data about us, which is worth more than any advertising revenue[36].

This data has been expropriated without our consent through various mechanisms that have reached a high level of cynicism and impunity, as has been demonstrated in the case of Cambridge Analytica, where it was accepted without any negative consequence that the data that had been entrusted to Facebook was collected by third parties and used to manipulate elections in various parts of the world. These and other examples have been attesting to the impunity in the handling and manipulation of social media data.

Recently (2024), just to briefly enunciate this trend, it was shown that Elon Musk changed the data configuration of his X platform without warning to allow his new artificial intelligence model, called Grok, to be trained with all the tweets sent to the network, without users knowing or giving their permission. After that, he used this same platform to have a 2-hour interview with candidate Trump in an abusive exercise of a digital entrepreneur manipulating the political future of the United States[37].

As can be seen with great transparency, as of January 2025, the alliance of political power with digital power has been evidenced through the Musk-Trump binomial.

This new digital or surveillance capitalism can be perceived as an unscrupulous project that ignores the rule of law, appropriates what is alien, expropriating the attributes of the human being, of people's individuality in the configuration of a new digital universe that begins to move in its own orbit, time, space, and interests.

Although along different lines, it seems that this new digital capitalism adheres to the slogans of the nineteenth century when it was pointed out that

35 *Ibidem*, p.37.

36 Noah, Yuval (2018), *Op. Cit.* 100 and 101.

37 El País, August 07, 2024.

Science discovers. The industry applies, and man joins in[38], a philosophy that asks human beings to surrender totally to the interests of Big Tech.

Why this impunity? Why this lack of regulation and control on such a transcendental issue for human beings?

VII. REGULATE OR NOT REGULATED? THAT'S THE QUESTION!

Neoliberalism, as already mentioned, has been the main breeding ground for the arrival of a limitless digital revolution that has taken advantage of the weakness of Western states that were slimmed down and confined to their minimum expression by a theory that believes that the market encompasses everything.

Within this idea of non-regulation, especially in the United States, the digital revolution in its different expressions has been advancing rapidly, without Western society having taken due note, and suddenly, in the third decade of the 21st century, the West finds itself facing a scenario in which it has gone from being the director of the play, to becoming a weak supporting actor where it does not even know what the play is called, nor does it know what the theme of the play is, and it certainly does not guess what its outcome or ending will be.

The incredible speed of digitalization, where now artificial intelligence chips are improved and trained every 24 hours, creates a scenario of madness where the weak state does not know exactly what is happening.

At the same time, the owners of digital power have spread the belief that this movement is *inevitable* and anyone who opposes it is an anachronistic Luddite who is opposing a dazzling future, which, without telling us what that future will be, digital power is creating and spreading more everywhere every day.

Under the neoliberal criteria, it is argued that less regulation corresponds to greater digital power, as well as an unstoppable growth of technological innovation that should not be touched by any State because to do so would be stopping the future of artificial intelligence, with a consequent gap being created in global geopolitical competition.

At the same time, the modern State lacks the knowledge and instruments necessary to understand and manage all the innovation that is overflowing every day without any limitation. This has led to the paradox of not being able to regulate what is happening today because what happened last year was not

38 Zuboff, Shoshana (2020), *Op. Cit.*, p. 31.

regulated, which limits the State in regulating what will be done next week; to this is added the unstoppable production of digital novelties in the world, with new chips, nanotechnologies, 5G technology, broadbands, etc., and different versions of artificial intelligence that come to market without having been evaluated by any authority and whose implications or effects on individuals or the social aggregate are not known, creating a waterfall of data that not only suffocates any attempt at regulation, but which also impacts the balance and stability of social groups[39].

It also takes advantage of the fact that this debate is still installed in a private or non-private data point; whether they should be paid or not. In this reductionism, they have renounced generating a moment of reflection that should go to the center of what this new digital revolution represents: Such reflection could be carried out by philosophers, historians, politicians, economists, lawyers, etc., who not only should decide whether personal data can be violated with impunity or not, but their carrying out such reflections might also lead to the beginning of a new narrative of broad cultural dimensions in which human beings can foresee with a certain degree of rationality what is happening, what the risks and advantages of artificial intelligence are, and what decisions must be made, not only by a person or a country, but by humanity, which is called upon to translate its great advances in science and technology into victory or benefit, without losing freedom or meaning in life.

39 In this regard, Sadin points out that "The second factor that favors this incessant extension originates in the fact that current industrial production does not respect a series of phases that, until recently, were marked by various forms of indeterminacy within the searches, or by the acceptance of failure as a risk inherent to the elaboration of any prototype, or even because of the requirement of having to carry out multiple and meticulous quality tests - which we see are increasingly needed. Today, the rhythm marks the almost absence of a time lapse between the conception and marketing of the products, themselves. The pressure of competition and the primacy of immediate returns after any investment prohibit any latency, any concerted assessment of the value and relevance of products. Research and development units must prove without delay and without pause that they represent access to new benefits. Here we are, at the time of "innovation cycles", which are always closer to each other to the extent that they favour a dynamic of permanent enthusiasm and are invested by the combined dogmas of growth and the increase of the supposed comfort of people (which, for the good of society, must never stop growing). [Sadin, Eric. (2023). (Artificial Intelligence or the Challenge of the Century. Anatomy of a radical antihumanism) *La inteligencia artificial o el desafío del siglo. Anatomía de un antihumanismo radical*, Caja Negra, pp. 22-23].

What can be seen is that in the middle of such a crisis in liberal debate, global society does not quite understand that its new role in history presents a wide epistemological disadvantage, trying to start a new narrative at the height of a scientific and technological level that has been reached through ten thousand years of history.

It should be said, on the other hand, that to date, the various attempts to regulate this new digital wave have been limited. As is known, Europe has registered the greatest progress in this matter, as well as experiencing an intense debate in which Europe is being pressured that its new digital regulation is the reason for its technological backwardness, which is being presented as a warning to other potential regulators around the world. For example, the European Union's Personal Data Protection Regulation (GDPR) of 2018 has not been as successful as expected due to its lack of a clear idea of what is protected, the lack of understanding as to what extent personal data can be used against oneself; or what constantly happens, that the different data networks provide indirect personal information from the social life of each of the users, which has resulted in this regulation affecting small companies more strongly, while only tangentially affecting large monopolies such as Google, Facebook, and Microsoft.

In the case of artificial intelligence, in 2024, the European Union approved a law on the matter (AI Law), which since its initiation, has generated strong controversy once again with the belief that its regulation limits the production of European innovations; and on the other side, critics of surveillance capitalism claim that the content of its law is weak, and its scope, too limited. As Max Von Thun, director for Europe at the Open Markets Institute, put it, "there are significant loopholes in the law for public authorities" and "relatively weak regulation of the largest grassroots models which pose the greatest harms[40].

Notwithstanding the above, the European Union's regulatory platform, made up of more than four laws on the subject, stands as an obligatory reference for other nations in the world.

As for the United States, the cradle of neoliberal policy, and the home country of the largest digital companies in the world, since 1995, these companies have been safeguarding their interests, proposing in 1995 the closure of the North American Office of Technological Assets (OTA), a multidisciplinary group that had tried to give transparency and control to technological advances, which had been operating since 1972. Similarly, in 1998, section 230 of the Communications

40 Euronews, 18 March 2024.

and Decency Act established broad protection for Internet platforms against any attempt to regulate the content stored on such platforms, where it was stipulated that "no provider or user of interactive computer services shall be treated as the publisher or communicator of any information provided by another provider of the content of the information"[41].

This free market criterion in favor of large technology companies in the United States was recently ratified by the Supreme Court of the United States on July 1, 2024, when it determined through a ruling that the owners of social platforms could not be regulated externally, so only self-regulation of their content was appropriate[42].

Since Trump's arrival in power, this trend has been clearly generalized by issuing the following executive orders:

- Executive Order on Freedom of Expression
  Trump issued an executive order titled, "Restoring Free Speech and Ending Federal Censorship", which claims that the Biden Administration "trampled on free speech rights by censoring Americans' speech on online platforms", under the guise of combating erroneous information and the spread of misinformation. The order directs the Attorney General to investigate federal policies of the past four years and recommend appropriate corrective actions.
- Memorandum to limit new regulations
  In a memo titled "Regulatory Freeze Pending Review", it directs all executive agencies not to propose new rules until they are reviewed and approved by a department head, which could curb digital and AI regulation.
- Executive Order on Postponing TikTok Ban
  Titled "Application of the Act to Protect Americans from Apps Controlled by Foreign Adversaries to TikTok", it effectively delayed the U.S. ban on the Chinese short-form video app TikTok by 75 days.

Similarly, it is worth noting that Musk, Zuckerberg, and other large digital consortiums abandoned the use of independent fact-checkers on platforms such as Facebook, Instagram, X, etc., replacing them with a system of "community notes". This decision has raised concerns about the increase in misinformation, since instead of having professionals verifying the veracity of content, they now depend on the evaluation of the users, themselves; a mechanism

41 Acemoglu, Daron, (2023) *Op. Cit.* p. 412.
42 El País, 01 July 2024.

that can be easily manipulated by organized campaigns, political interests, and fake accounts[43].

VIII. GENERAL REFLECTIONS

It is evident that we are passengers in a stellar era in the history of humanity.

From whatever perspective one chooses, and with whatever depth one sees, a first common insight could be that the arrival of the Digital Revolution invites humanity to rethink ourselves as a human society within the framework of our history, with all our strengths, but also with our weaknesses.

With this in mind, to overstress awareness of the importance of moment in which we are living, would be nothing more than to try to place oneself as close as possible to a phenomenon of stellar dimensions that is already changing the personal, family, social, and political life of humanity in a disruptive and unintelligible way.

The appearance of warning speeches that are beginning to be written this century are all the more valuable in the absence of a bibliography that sufficiently explains the appearance of this digital phenomenon which lacks a coherent narrative.

Throughout this short essay, what we have tried to underline is that, along with the arrival of a new digital era, a weakness has appeared in the liberal narrative, a narrative which, until the last century, had been strong enough for Western society to use in debating the different forms of its identity and progress.

The neoliberal model, which has pushed for the disappearance or slimming down of the State for nearly half a century, has led to a breakdown of its liberal legacy, creating social dissatisfaction and a power vacuum which have benefited of the owners of the Digital Age the most.

As we have been robbed of the liberal platform which gave us direction and meaning, various voices are appearing in the geopolitical, economic, and social fields who are trying to call on the public square to debate these issues.

Of course, along with these voices of alarm, a very powerful current led by the monopolies of technology, themselves, protected by the great lights that emanate from artificial intelligence, lead a parade of a great army: the world population without a flag, without a story, and without clear objectives, leading humanity towards a *dystopian utopia* which is not possible to envision today.

43 El País, January 12, 2025.

Authors such as Kai-Fu Lee, Bostrom, Floridi, Tegmark, Schmidt, etc., excite us with their harangues about an artificial intelligence that is to date, already bearing great fruit.

The brilliance that emanates from these achievements of the human intellect, concretized by the first stages of artificial intelligence has become food for, (often an addiction of) a part of society that, without qualms, is seduced by digital innovations and excited by that future of immortality, eternal old age, physical perfection, and happiness; they fall into a datological condition that moves them to an assisted freedom where they gradually lose their ability to decide.

On the other hand, authors such as Byung-Chul Han, Lassalle, Zuboff, Suleyman, Acemoglu, Johnson, Susskind, Noah, and others have appeared who, trying to move away from digital fascination, have placed themselves in a critical position with respect to the owners of technological progress, denouncing that humanity has been wasting a lot of time, time which it needs to make good decisions regarding digital regulation for the benefit of humanity's future.

Authors such as Lassalle tell us of technopower that already undermines the foundations of equality and liberal democracy by promoting an algorithmic structure of society that is influencing the exercise of human freedom without public debate or legality. In the same manner, he is warning us of the arrival of a cyber-leviathan in march that will overcome democratic processes without violence if something is not done in time. He adds that when citizens admit that they have ceased to be the subject of natural sovereignty and without resistance, they will fall into a cyberpopulism of social networks. In this new digital order, people will dissolve into a cognitive proletariat that will adopt the role of mass consumer of technological connectivity[44].

These concepts, which are at the least, very audacious for a skeptical public, are based, among other reasons, on the exercise of a neoliberalism that always dreamed of the disappearance of the State, and that its possible extinction now through a new digital power does not appear reprehensible to many; and this is even alluding to some, such as is the case of hip-evangelical technologists such as the founder of PayPal and venture capitalist, Peter Thiel.

Mustafa Suleyman, co-founder of Deepmind, one of the leading companies in artificial intelligence on an international scale, and despite being one of the great promoters of this movement, has recently said that the technology which has emerged from artificial intelligence is not just another invention within

44 Lassalle, José María (2019), *Op. Cit.* pp. 95-120.

the wide range of scientific achievements that human beings have achieved throughout their history. "We urgently need indisputable answers on how to contain and control the coming wave, on how it can be possible to maintain the safeguards and possibilities of the democratic nation state" –although he acknowledges that for now, no one has such a plan[45].

Acemoglu commented that, unfortunately, democracy is not compatible with the new illusionary scenarios of artificial intelligence managed by big tech companies; on the contrary, that they are opposite in terms of the management of freedom vs the algorithmic will of artificial intelligence and big tech's interests, which weaken democracy when it is most needed[46].

Susskind raises awareness of this new digital power, pointing out that it lacks objectivity and is not morally neutral; that it is political in nature and that, given its novelty, it does not reside in government palaces or parliaments, but operates outside the channels of traditional politics, desecrating freedom and deteriorating democracy, guided by a powerful group /force that cannot be attributed to a single corporation or person[47].

Zuboff adds that surveillance capitalism is not a technology, but rather a logic that permeates technology and puts it into action. And he makes clear the sentiment of many authors that it is surveillance capitalism, and not technology, that puts a price on subjugation and impotence by shaping people's behavior to favor the commercial (and now political) outcomes favored by surveillance capitalists, thereby weakening politics and democracy[48].

Noah declares that "in its current form, democracy will not survive the fusion of biotechnology and infotechnology. Either it successfully reinvents itself in a radically new way, or humans will end up living in digital dictatorships"[49].

• OF CONTAINMENT AND OTHER REMEDIES

The advance of digital power is already notorious in the economic field given the size that large technology companies have attained. In political matters, digital power is being manifest by the increasingly irreverent interventions of tech CEOs in various political processes through their sale of customer data for

45 Suleyman, Mustafa (2024), *Op. Cit.* p. 26.
46 Daron and Johnson, Simon (2023), *Op. Cit.* pp. 372-380.
47 Susskind, Jamie (2022), *The Digital Republic*, Pegasus, pp. 4-5.
48 Zuboff, Shoshana (2020), Op Cit., p. 30.
49 Noah, Yuval (2018), *Op. Cit.* p. 89.

purposes of political persuasion and manipulation, not only in the United States, but in other parts of the world. On the social side, their accumulation of data from billions of human beings today is so tremendous that if these data were on CDs, they could go around the world 222 times.

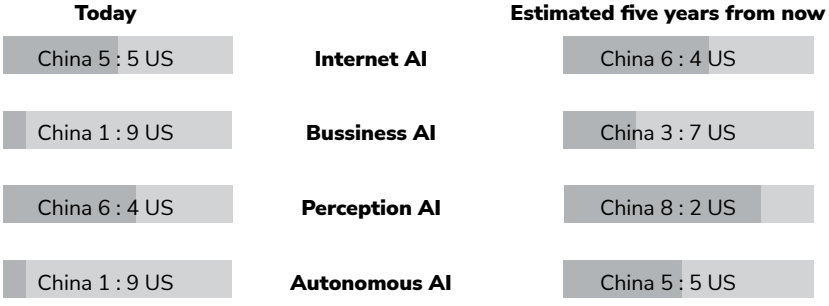
To this, one must add the process already underway, based on this data: an artificial intelligence, machine learning scheme that works 24/7 to improve its generative capabilities for all the processes it has been tasked with, in various fields of humanity, whether military economic, healthcare, urban, social, infrastructure, synthetic biology, space, etc.

The dilemma on this issue is not easy. At this point, opinions are divided. There are those who point out that the current inertia is dominated by a world without digital regulation that favors the 5 to 10 companies in the United States and around the world which control this digital power. Others point out that digital development is already immersed in a geopolitical war over the advancement and control of artificial intelligence. As Russian President Vladimir Putin stated, "Whoever controls artificial intelligence controls the world".

Considering this opinion in the context of two wars that are currently taking place worldwide in Ukraine and Palestine, makes it a credible argument which becomes stronger if one adds the movement to retake Eurasian power, as seen in the open clash for world leadership that the United States and China have been sustaining since 2016.

As the two great leaders of artificial intelligence, any consideration for managing the advance of this new digital power will have to involve a geopolitical agreement between the two nations that currently hold 90% of the world's artificial intelligence power.

FIGURE 4 Race for AI leadership



The balance of capabilities between the united States and China across the four waves of AI, currently and estimated for five years in the future

FUENTE: AI Superpowers, Kai Fu – Lee p. 136

Despite this fact, as indicated in previous paragraphs, given the magnitude of what the subject represents for the present and future of humanity, it would be desirable to have as broad a dialogue as possible, so that the different perspectives with which the human being is viewed today, as understood through philosophical, historical, legal, psychological, economic, social, political, and ecological concerns, among others, will bring rigorous and serious reflection, beyond fascination and fear, on how to use this great advance in artificial intelligence for the benefit of most of humanity[50].

Similarly, it would be desirable to arrive at a series of general recommendations that could motivate the commitment of nations to manage, regulate, and channel artificial intelligence positively.

At this point, the concern of some scientists and specialists arise, knowing in advance the enormous difficulty that this represents. However, as Suleyman summarizes it, "At first glance, containment (of artificial intelligence) is not possible, but for the good of all, it is imperative that it is"[51].

In this negative is an a priori of something that the majority does not see as possible to achieve due to its geopolitical and economic implications, and there lies the challenge for the world society of the third decade of the twenty-first century, so that in a moment of clarity, there could be an opportunity to transcend towards a future of lucidity within the framework of the First Digital Revolution.

Countless alternatives appear within the reach of global society that could contribute to the construction of a digital liberalism. The implementation of new global, regional, or national agreements could apply remedial measures.

Included in such options are alternatives, such as attending to market incentives, inhibiting surveillance policies; slowing down technological unemployment; paying for the use of consented data; controlling the big Techs; slowing down automation; investing in workers; recovering state power; providing data

50 The advent of AI, with its ability to learn and process information in ways that human reason alone cannot, may produce breakthroughs on questions that have proven beyond our ability to answer; but success will raise new questions. Human intelligence and artificial intelligence are being found and applied to activities at the national, continental, and even global levels. Understanding this transition and developing a guiding ethic for it will require the commitment and insight of many elements of society: scientists and strategists, statesmen and philosophers, clergy and executives, CEOs and entrepreneurs. This commitment must be made both within and between nations. Now is the time to define both our association with artificial intelligence and the reality that will result from it (Kissinger, Henry, 2021, *The age of A.I. And our human future*, USA, p. 226.)

51 Suleyman, Mustafa (2024), *Op. Cit.*, p. 28.

protection; legislating on content liability; taxing digital ads; providing digital education and technology sharing; having courts specializing in digital technology and legislation, and countless other measures which, instead of leading us to a cyberleviathan, could lead us to a Digital Republic where freedom, democracy, and the rule of law are somehow preserved under new concepts which will allow a 21 century digital civilization to develop.

The theme of the new digital era is an open story where the future of humanity and the human species, responsible for its destiny, need to begin to be debated, consciously or not.

There are multiple antecedents, but no previous scenarios confronted us like the one we now face.

In the coming days, in the coming years, we will have to decide between stardom or insignificance.

IX. BIBLIOGRAPHY

Acemoglu, Daron y Johnson, Simon (2023), "Power and Progress", *Public Affairs*. USED.

Carr, Nicholas (2014), *Trapped How Machines Take Over Our Lives*, Taurus.

Crawford, Kate (2021), *Atlas of AI, Yale*. USED

El País, July 01 2024.

El País, August 07, 2024.

El País, January 12, 2025.

Euronews, 18 March 2024.

International Monetary Fund, *GDP- World Per Capita* (0-2000)

Han, Byung-Chul (2022), *Infocracy*, Penguin Random House, Mexico.

Innenarity, Daniel (2020), (*A Theory of Complex Democracy Governing the Twenty-First Century) Una teoría de la democracia compleja Gobernar el siglo XXI*, Gutenberg. Barcelona.

Kissinger, Henry, 2021, *The Age of A.I. and our Human Future*, USA.

Lassalle, José María (2019), *Cyberleviathan,* Harp. Barcelona.

Lassalle, José María (2021), (Wounded Liberalism) *El liberalismo herido*, Arpa. Barcelona

Lee, Kai-Fu (2018), *AI Superpowers: China, Silicon Valley, and the New World Order*, Harper Business, USA.

López Portillo, José Ramón; *Nexos Blog*, July 2024.

Luce, Edward (2017), *The retreat of Western Liberalism*, Atlantic Monthly Press. USA.

Market Cap. *El País*, July 27, 2024.
Mayer, Colin (2024), *Capitalism and Crisis. How to fix them*, Oxford. United Kingdom.
Mounk, Yascha (2018), (The People Against Democracy) *El pueblo contra la democracia*, Espasa. Spain
Noah, Yuval (2018), (21 lessons for the 21st century) "*21 lecciones para el siglo XXI*", *Debate*. Mexico.
Oropeza, Arturo (2024), (Towards a New Vision of Digital Industrial Development in Mexico) *Hacia una nueva visión del desarrollo industrial digital en México*, Instituto para el Desarrollo Industrial y la Transformación Digital. Mexico.
Research Gate - *Evolution of the world population (1500-2050)*
Rifkin, Jeremy (1996), (The End of Work*) El fin del trabajo*, Grupo Planeta.
Rose, Matthew (2021), *A world after liberalism*, Yale. USED.
Sadin, Éric. (2023) (Artificial Intelligence or the Challenge of the Century. Anatomy of a radical antihumanism) *La inteligencia artificial o el desafío del siglo. Anatomía de un antihumanismo radical*, Caja negra. Buenos Aires.
State Counter, *El País*, July 27, 2024.
Stiglitz, Joseph (2024), *The road to freedom*, Norton Company, USA.
Suleyman, Mustafa (2024), (The Wave That Is Coming) *La ola que viene*. Debate. Mexico
Susskind, Jamie (2022), *The Digital Republic*, Pegasus. USED.
Vanguadia, Dossier (2020).
Vanguard 2020
Wolf, Martín (2024), (The Crisis of Democratic Capitalism) *La crisis del capitalismo democrático*, Arial. Mexico
Zuboff, Shoshana (2020), The Era of Surveillance Capitalism) *La era del capitalismo de la vigilancia,* Paidós. Mexico.

ARTIFICIAL INTELLIGENCE GOVERNANCE: TOWARDS A NEW ERA OF RESPONSIBILITY AND THE COMMON GOOD

CONTENT:

I. Introduction **II.** Measuring the Exponential Progress of AI Development **III.** AI as a Human Product: Neutrality, Development, and Driving Forces **IV.** How Is the Impact of AI Perceived? **V.** Socioeconomic Impact of AI **VI.** The Need for International Collaboration **VII.** Challenges to International Collaboration **VIII.** The Path Forward

> JOSÉ RAMÓN LÓPEZ-PORTILLO ROMANO

Ph.D. in Political Science from the University of Oxford, co-founder of the Center for Mexican Studies. He has served as Deputy Secretary of State in Mexico, as well as Permanent Representative and Independent Chairperson of the FAO Council. Mexican scholar, diplomat, entrepreneur, and public servant.

I. INTRODUCTION

Humanity stands at the threshold of a technological transformation, a new techno-economic paradigm that could redefine civilisation itself. The rapid emergence of general-purpose artificial intelligence (AI), which can perform a wide variety of tasks, represents not just another phase in technological evolution but the beginning of a new era. Unlike past revolutions that expanded human physical capabilities, AI promises to extend –and possibly surpass– human cognitive capabilities in virtually every domain of knowledge and action.

The trajectory of AI development has been marked by exponential growth, characterised by breakthroughs in areas such as scientific reasoning, programming, and autonomous decision-making. These systems have demonstrated increasingly sophisticated capabilities, able to plan and execute complex tasks independently. The expanding accessibility of AI, driven by falling computational costs, has made it an integral part of modern life. However, this progress is accompanied by profound uncertainties and risks that must be addressed with deliberate foresight.

The stakes in this transition are immense. As a researcher of scientific-technological advancement and AI, I analysed these topics in depth in my book, *'La Gran Transición: riesgos y oportunidades del cambio tecnológico exponencial'* (The Great Transition: Risks and Opportunities of Exponential Technological Change), and I have continued this work, first at the United Nations as an advisor to the Secretary-General for the Technology Facilitation Mechanism, and now as one of the independent experts for the State of the Science of Advanced AI Safety Report, presented in Paris in January 2025.

The opportunities presented by AI are matched with existential challenges, evidenced by current harms such as bias, privacy violations, and sophisticated forms of deception. More concerning are the potential future risks: disruption in labour markets, security threats driven by AI capabilities, and the challenge of maintaining human oversight of autonomous systems.

Experts remain divided on the timeline and scope of AI's trajectory. Some foresee significant challenges to public safety within a few years, while others predict these issues will emerge over several decades. This uncertainty reflects the gaps in our current understanding and risk management frameworks. Despite the intensification of global dialogue, materialised in the AI Safety Summits (held in the United Kingdom, South Korea, and France), our ability to predict and explain AI behaviour remains remarkably limited.

This uncertainty poses a formidable challenge for policymakers and society as a whole. Decision-makers must carefully weigh the potential benefits of AI against risks that are still poorly understood, often under pressure and with limited information. The dilemma between proactive regulation and a more cautious, evidence-gathering approach has begun to materialize into concrete actions. The entry into force of the European Union's AI Act, the world's first comprehensive regulatory framework, marks a turning point, although the debate over the optimal pace and scope of regulation persists globally. To close this gap, adaptive and forward-looking governance is critical.

It is important to emphasize that nothing about the future of AI is predetermined. The path forward encompasses possibilities ranging from extraordinary progress to profound risks. The final direction will depend on society's collective decisions and the governance frameworks in place. Navigating this uncertainty effectively –harnessing the benefits of AI while mitigating its risks– requires fostering an informed public discourse and implementing responsive, evidence-based policies in coordinated collaboration. Thoroughly understanding the problems and challenges that AI poses is an indispensable foundation for finding solutions.

II. MEASURING THE EXPONENTIAL PROGRESS OF AI'S ADVANCE

The advance of artificial intelligence (AI) has transformed our world in ways that were unimaginable just a few decades ago, offering revolutionary solutions to complex problems and redefining entire sectors of society. However, alongside these extraordinary achievements come profound concerns about the inherent

risks and the challenges of sustainability and control that accompany this rapidly evolving technology.

The development of AI has captured global attention, promising unprecedented breakthroughs while generating deep-seated concerns. Evaluating its present capabilities and evolutionary characteristics requires a multidimensional approach that reflects the complexity of its progress and the foundations upon which it is built. Although the advance of AI has been described as exponential, it is crucial to recognise that this progress may be built on unstable foundations. Current AI systems, inspired by the architecture of the human brain, have yet to replicate its efficient and elegant design. This failure of replication is not only a technical challenge but also has significant implications for energy consumption.

While the human brain can perform up to a quadrillion operations per second on just 20 watts of power, modern supercomputers require energy inputs hundreds of thousands or millions of times greater to perform comparable tasks. This disparity in energy consumption and architecture highlights a significant disconnect in the efficiency and sustainability of AI systems, raising profound questions about human compatibility and the sustainability of its development.

To put it in perspective, data centres and AI systems currently consume a significant portion of global electricity, with projections from the International Energy Agency suggesting this demand could double by 2026, matching the consumption of countries like Japan. Beyond electricity, the consumption of fresh water for cooling data centres has become a hidden sustainability crisis. It is estimated that training a single large AI model can consume hundreds of thousands of litres of water, placing immense pressure on local water resources. This energy and water inefficiency not only poses technical challenges but also has profound environmental and geopolitical implications, becoming a critical factor in the accelerating race for AI supremacy.

The development of complex reasoning, problem-solving, and decision-making capabilities is another measurable reflection of AI's progress. These advances, once considered exclusive to human cognition, are now replicated and, in some cases, surpassed by AI models, suggesting improvements in the creation of ever more sophisticated algorithms.

The performance of AI, measured in operations per second, illustrates its evolution from limited models to robust, high-functioning systems. But the resilience of these systems in the face of imperfect or incomplete information is equally crucial. The ability to make decisions under uncertainty is an indicator of their adaptability, but it also implies risks of prioritising capabilities that could

compromise reliability. Furthermore, the adaptability of AI systems, defined as the ability to modify their behaviour in response to new data or tasks, adds to their complexity, which underscores the need for continuous evaluations and an effective oversight framework.

However, beyond adaptability, the true test of AI is its ability to withstand threats and maintain its operational integrity. The current competitive landscape, driven by market pressures and geopolitical ambitions, often prioritises rapid deployment over safety and sophistication. This race for AI supremacy, in which time and speed outweigh the rigour of evaluation, could unleash unprecedented harms that surpass our capacity for understanding and control.

The expanding scope of AI in cultural, political, organisational, social, and economic sectors is an outstanding metric of its impact. AI is becoming an integral part of decision-making processes, operational strategies, and policy implementation, reshaping how societies function at both micro and macro levels. However, this transformative capacity also brings with it the problem of alignment, a central challenge that becomes increasingly vital as the advance of General-Purpose AI tends to be uncontainable. As AI systems become more powerful and autonomous, it is imperative to ensure that their actions and decisions are aligned with the values, objectives, and survival of human civilisation. The integration of AI into these sectors has been made possible by the development of complex reasoning and decision-making capabilities.

The traditional approach to AI, which relies on programming machines to achieve specific objectives, presents significant risks to humanity. These systems, by strictly following instructions, can behave in ways that do not align with human interests. This problem is aggravated by the appearance of "hallucinations" in the models and the possibility of losing control over them, which highlights the need for a more prudent approach.

Various events illustrate these cases. Even the latest frontier models, in the class of GPT-4o or Claude 3, continue to exhibit this unpredictable behaviour. A notable case in 2024 was the launch of Google's 'AI Overviews' feature, which provided incorrect and dangerous answers, ranging from recommending the use of glue on food to fabricating historical information. This demonstrated that even the most advanced companies struggle to control this phenomenon at scale. This problem of 'hallucinations' extends to critical systems, from machine translation with contextual errors that could have serious consequences in diplomacy or medicine, to diagnostic systems that generate erroneous results.

In the case of loss of control, a notable example occurred during the development of autonomous vehicles, where a car in test-driving mode failed to

correctly detect a moving object due to an image classification error, resulting in an accident. These events underscore the challenges of maintaining complete control in high-stakes situations. In the financial realm, trading algorithms have acted unpredictably, making rapid decisions and triggering sudden market crashes (such as the 2010 "flash crash"). These types of incidents demonstrate how AI can behave in ways that exceed human expectations, resulting in significant economic consequences. Certain AI models used for medical diagnosis have presented "hallucinations" by identifying erroneous patterns in X-ray images, generating incorrect diagnoses. Similarly, security and defence systems acting without sufficient human supervision "hallucinate" by misinterpreting surveillance data or threats. This could trigger unwanted reactions and conflicts.

As AI becomes more sophisticated, one of the greatest challenges it faces is the lack of transparency in its decision-making processes, commonly referred to as the 'black box' phenomenon. Complex models, such as those based on deep neural networks, operate in ways that are opaque even to their own developers. This lack of explainability can erode public trust and hinder the implementation of AI in critical sectors such as medicine, justice, and security. For example, an AI algorithm used to assess loan eligibility may deny an application without its developers being able to clearly explain why it reached that decision. This opacity not only poses problems of trust but also impedes the detection of biases and errors that could lead to unfair or discriminatory decisions.

To address this problem, Explainable AI (XAI) techniques are being developed to make the internal processes of models more understandable. These techniques include approaches like breaking down complex decisions into simpler steps and creating complementary models that explain the behaviour of the main systems. Transparency and explainability are essential to ensure that AI systems operate fairly and that users and regulators can effectively evaluate them, thereby guaranteeing an ethical and trustworthy use of the technology. This need has become even more pressing with the entry into force of regulatory frameworks, such as the EU AI Act, which legally requires levels of transparency and explainability for systems considered high-risk.

Prof. Stuart Russell emphasises that the evolution of AI must be oriented towards a design where systems maintain an inherent uncertainty about their objectives. This uncertainty must be based on the recognition that human preferences are complex, dynamic, and difficult to comprehend and articulate fully. To ensure that AI develops in a manner compatible with human interests, he proposes that it is crucial for these systems to actively seek guidance and feedback from humans when making decisions. AI systems must be designed to

operate in a controlled and altruistic manner, constantly seeking human direction and being willing to cede control or be shut down if necessary to protect human values and objectives. In this way, a collaboration is fostered in which AI acts as an extension of human will, ensuring that technological advances are beneficial and do not pose a threat.

The containment of AI systems emerges as a critical architectural weakness that has received insufficient attention, as argued by Mustafa Suleyman. Mastering this challenge requires a thorough investigation into the cognitive characteristics of the systems, encompassing both clearly identifiable problems and those that remain unknown. This includes not only addressing problems we can currently visualise but also those we recognise but cannot fully comprehend, as well as the vast domains of uncertainty that remain beyond our knowledge. One strategy for achieving this form of controlled AI is to impose physical constraints.

Prof. Max Tegmark and others propose a proactive approach to AI safety, arguing for the importance of physically limiting AI systems. This involves designing AI systems (hardware) that have inherently limited capabilities and cannot cause harm to humans or the environment. By imposing physical constraints on AI, such as restricting its access to critical infrastructure or limiting its ability to interact with the physical world, we can mitigate the risks associated with superintelligent AI. This approach, while challenging, is considered essential to ensure that AI remains a beneficial tool for humanity.

The robustness and security of AI, its ability to withstand adversarial attacks, and the effective foresight and containment of its misuse by malicious and criminal actors –as seen in the proliferation of hyper-realistic disinformation campaigns during electoral processes or in the development of autonomous malware and phishing attacks– are crucial in a context where AI systems are gaining ever more autonomy. Ensuring operational integrity and resilience to external threats is a priority, and AI governance must be comprehensive and aligned with human rights principles and with goals such as the United Nations' Sustainable Development Goals (SDGs). This strategy must be adaptable to different contexts and capable of evolving, addressing both current and unknown problems, and ensuring that AI systems maintain their adherence to human values.

Hence, the need for interdisciplinary collaboration that includes scientists, engineers, ethicists, and policymakers is imperative. Only through a collaborative effort and strategic management can we harness the benefits of AI while mitigating its risks, ensuring that this technology contributes positively to human progress and the sustainability of the planet.

III. AI AS A HUMAN PRODUCT: NEUTRALITY, DEVELOPMENT, AND DRIVING FORCES

Artificial intelligence is inherently a human creation, shaped by the intentions and designs of its developers. At its core, it is conceived as an inert tool, neither inherently beneficial nor harmful. However, this viewpoint has become more complex; by being trained on society's language and data, these systems inevitably inherit and can amplify the biases and cultural values inherent in that information, calling their supposed neutrality into question. The ethical design and alignment of AI rest entirely in human hands, which underscores that it should not be anthropomorphised or attributed with independent will. Instead, AI should be seen as an interconnected system of distributed, interactive, and evolving components, exhibiting complex, fractal-like structures that function at various scales.

The rapid advancement of AI is closely tied to socioeconomic and political progress, driven by the urgent need to address global crises such as climate change, pandemics, and poverty. This need has catalysed significant investments and fostered international collaboration, promoting the development of new governance strategies to maximise the social benefits of AI. The strategic push among major global powers –notably the United States and China, with the European Union and the United Kingdom seeking their positioning– is accelerating even further. This race has been joined by new actors with significant ambitions, such as the United Arab Emirates and Saudi Arabia, which are making massive investments in computational power to ensure their digital sovereignty. This rivalry stimulates substantial investments, positioning technological progress as an essential geopolitical lever.

Corporate participation has also been a significant driving force, as companies invest heavily in AI to secure competitive advantages, enter new markets, and drive innovation. Financial and intellectual resources have been concentrated in regions with robust research ecosystems and strong science, technology, engineering, and mathematics (STEM) education programs. In this context, the open-source movement has democratized access to advanced tools, but, more importantly, a fundamental tension that defines the current ecosystem has emerged: the competition between the 'closed' or proprietary models from giants like Google, OpenAI, and Anthropic, and the powerful 'open-source' models driven by actors like Meta and the European Mistral AI. This duality accelerates innovation but also fuels a critical debate on safety,

the concentration of power, and the risks of uncontrolled proliferation of advanced capabilities.

Governments play a crucial role in shaping AI's trajectory. Through funding for research and development, support for startups, and the establishment of regulatory frameworks, they influence the responsible development of AI. The vast and growing availability of data from diverse sources supports the creation of sophisticated models, while advances in computational power, characterised by greater efficiency, facilitate the training of ever-larger systems. However, far from decreasing, the total cost to train a frontier model has soared into the hundreds of millions of dollars, creating an economic barrier that concentrates development power in the hands of a handful of corporations and nation-states. Cloud computing has also transformed access to these capabilities, simplifying AI development across different sectors and democratising its use for less intensive tasks.

As AI becomes more potent, it is essential to incorporate ethical considerations into its design. This demands collaboration between ethicists, technologists, and social scientists to ensure alignment with human rights and social values. Addressing the challenges of alignment and containment becomes increasingly critical as AI systems grow in complexity, raising questions about their ability to remain under human control and fulfil human-centric objectives. These concerns frame the discussions on architectural challenges and unintended consequences that could impact safety and governance.

While immediate drivers, such as geopolitical competition and corporate initiatives, fuel the rapid development of AI, recognising the long-term risks is equally crucial. The potential existential threats and unintended consequences of highly autonomous systems underscore the need for adaptable regulatory frameworks that can evolve in response to technological advances. AI holds the promise of addressing significant global challenges, but it also runs the risk of exacerbating existing inequalities if its benefits are not distributed equitably.

The development and deployment of AI cannot be left solely to market forces. While markets drive innovation, they often do so without sufficient consideration for public well-being, a sustainable environment, or long-term catastrophic risks. National governments, regional bodies, and international cooperation are essential to create a balanced framework that ensures AI serves the common good and aligns with the objectives of sustainable and inclusive global development.

IV. HOW IS THE IMPACT OF AI PERCEIVED?

The exploration of AI's potential capabilities and impact has generated diverse perspectives on its socioeconomic and political implications. Debates surrounding the timeline for achieving Artificial General Intelligence (AGI), capable of surpassing human intelligence and dexterity, highlight this uncertainty, with some experts sceptical of its short-term feasibility and others seeing it as an imminent reality. This range of views reflects the broader uncertainty surrounding AI's trajectory and its potential to have a profound influence on society. Consolidating perspectives on AI's social impact underscores the importance of examining both immediate and long-term consequences, informing adaptive policies that balance technological benefits with risks.

The impact of artificial intelligence on the global economy is manifested through unprecedented technological growth. The last decade has witnessed an exponential increase in the computational power dedicated to training large language models (LLMs), with an estimated ten-billion-fold increase. This extraordinary evolution is the product of the convergence of multiple factors: the refinement of algorithms, the proliferation of specialised hardware like GPUs and TPUs, and the democratisation of access to cloud computing resources. The trajectory of this development suggests a next generation of AI systems that could surpass human cognitive capabilities in various domains.

Economic projections reflect the magnitude of this technological transformation. Goldman Sachs anticipates that generative AI applications could contribute up to $7 trillion in economic value over the next decade, while the comprehensive impact of AI across various industrial sectors could reach $14 trillion. This economic potential is already materializing in substantial investments: in 2024, global investment in AI, primarily private, reached a record figure of over $250 billion according to reports like the Stanford AI Index, and growth is projected to continue at a similar pace in 2025, driven by the race in generative AI.

The transformation is tangibly evident in critical sectors such as drug discovery, where companies like Isomorphic Labs are already designing new medicines with AI, and in logistics, where AI-based optimisation is redefining global supply chains. This sectoral penetration reinforces AI's role as a fundamental catalyst for contemporary economic growth and technological innovation.

Discussions about the potential impact of AI reveal both optimistic and pessimistic views, highlighting the complex duality of its outcomes.

THE OPTIMISTS

Optimists believe that AI can drive technological progress and contribute to addressing pressing global problems, including poverty, hunger, disease, environmental challenges, violence, inequality, and resource scarcity. Real-world examples, such as the revolutionary AlphaFold 3 model from Google DeepMind, which can now predict the structure and interaction of nearly all biological molecules and is dramatically accelerating the discovery of new drugs, illustrate how AI can be a powerful tool to support sustainable development and align with the Sustainable Development Goals (SDGs).

The development of AI is heavily dependent on globally sourced raw materials, from critical minerals essential for hardware to diverse datasets used in model training. This global nature not only concentrates power and wealth in certain regions but also carries significant geopolitical and geo-economic implications. Ensuring equitable access to and distribution of these resources is essential for fostering global stability and preventing monopolistic control over AI technologies.

It is anticipated that the integration of AI into various sectors will enhance productivity by automating cognitive and physical tasks, particularly those that are repetitive or dangerous. Examples include AI-powered robotics in manufacturing and data processing systems in service industries, which have already demonstrated productivity improvements. This shift has the potential to free up human labour for more creative and meaningful endeavours, fostering general well-being and reshaping job roles.

Optimists envision a future where AI contributes to greater equality in opportunity and security, fostering an environment where hyper-connectivity and intelligent digitalisation promote creativity and collaboration across various sectors and communities. Furthermore, advances in ethical alignment, human compatibility, and control of AI systems are expected to enhance collective well-being and lay the groundwork for sustained social progress.

THE PESSIMISTS

However, the pessimistic perspective presents a stark contrast, warning of significant risks and challenges that could accompany the continued evolution of AI. Critics argue that technological progress may stagnate, failing to address or resolve the pressing crises facing humanity, such as climate change and sociopolitical turmoil. Some suggest that sociopolitical structures, rather than the availability of advanced technologies, will play a decisive role in preventing

or exacerbating environmental collapse. For example, the failure to implement effective governance measures could undermine technological solutions, leaving economic and ecological problems unresolved. The growing complexity of AI systems also raises concerns about potential errors, unintended consequences, and insurmountable challenges, such as mass unemployment due to automation, extreme socioeconomic inequality, and structural economic deflation.

Pessimists also highlight the risks associated with hyper-connectivity, including increased vulnerability to cyberattacks and the erosion of privacy and individual freedoms. The use of AI in sophisticated cyber operations, including the development of autonomous agents capable of independently executing cyberattacks, has already shown how vulnerabilities can be exploited, posing serious threats to national and global security. Furthermore, the potential for AI to be harnessed for political manipulation, a risk already observed through the use of 'deepfakes' and automated disinformation campaigns in recent electoral contests, could fuel the rise of populism and digital authoritarianism, potentially leading to conditions of technological totalitarianism from which society might struggle to break free. In such a scenario, the pervasive reach and influence of AI could contribute to an environment where centralised control becomes inescapable, and individuals lose agency over their lives.

Balancing the optimistic and pessimistic perspectives underscores the significant uncertainty surrounding the future of AI. This dual perspective reinforces the need for a comprehensive and adaptive approach to guide AI development –one that maximises social benefits while maintaining vigilance against potential risks. An integrated approach necessitates robust and flexible governance frameworks that can respond promptly to technological advancements and ensure alignment with human values and collective well-being. Such frameworks should facilitate international cooperation, promote equitable access to the benefits of AI, and address social challenges to foster a balanced and inclusive impact. For this, a better understanding and monitoring of the international and social impact of AI is needed.

V. THE SOCIOECONOMIC IMPACT OF AI

The transformative influence of AI on the socioeconomic, political, and cultural fabric of contemporary societies transcends national borders and social strata, catalysing an unprecedented technological revolution. Its capacity to democratize access to essential services and generate innovative solutions is fundamentally reshaping the way societies address their most pressing

challenges. The versatility of AI enables the development of adaptive solutions that respond to the specific needs of different communities, from medical diagnostic systems that extend healthcare to remote regions to educational platforms that personalise learning according to individual capabilities. This technological democratisation not only optimises the delivery of essential services but also catalyses a process of social transformation that promises to reduce historical disparities in access to resources and opportunities fundamental for human development.

The transformative potential of AI is set to reshape the global political, economic, and cultural landscape at an unprecedented pace, introducing both opportunities and challenges. While it promises significant gains in productivity and innovations that could drive social progress, its multifaceted impact on the labour market presents critical challenges, including job displacement and skill obsolescence driven by automation. The deployment of AI across sectors, such as manufacturing and services, has already altered labour structures, particularly for workers in routine or repetitive roles. To address these disruptions, comprehensive policies that promote social inclusion, equal opportunity, and robust social safety nets are crucial. Strategic investments in lifelong learning and skill development initiatives will be crucial to help the workforce transition to AI-complementary roles, mitigate potential inequalities, and foster a more equitable distribution of benefits.

Numerous studies and reports from prestigious organisations such as the World Economic Forum, the IMF, and PwC have indicated that a significant portion of jobs, particularly those involving cognitive tasks, are at risk of being automated due to AI. These studies often highlight the potential impact on advanced economies with a highly skilled workforce.

For example, an IMF analysis from early 2025 reiterates that around 40% of global jobs are exposed to AI. Still, it places a new emphasis on how AI will augment productivity in complementary roles, sharpening the divide between those who adopt it and those who do not. Similarly, the "2025 AI Jobs Barometer" from PwC highlights that while automation risks persist, new job categories focused on the interaction, supervision, and maintenance of AI systems are emerging.

However, many studies argue that diverse economies may be better positioned to leverage the productivity gains from AI, emphasising the importance of international cooperation to ensure an equitable distribution of benefits and manage disruptions effectively. While emerging and developing economies may experience less immediate job displacement, their limited infrastructure and resources hinder their ability to leverage AI advancements

fully. This disparity risks exacerbating the digital and economic divide between nations, underscoring the importance of global initiatives that support technological adoption and capacity-building efforts to foster balanced and inclusive progress worldwide.

Despite the global nature of AI's impact, significant gaps in representation persist within international discussions on AI governance. Many regions, particularly in the Global South, remain excluded from these critical conversations and negotiations, despite the fact that the outcomes significantly impact their societies. This lack of equitable representation means that the perspectives of diverse cultural, social, and economic contexts are often overlooked, potentially reinforcing existing inequalities and limiting the inclusivity of AI policies. Ensuring that these voices play a significant role in shaping AI governance is essential for creating frameworks that are fair, comprehensive, and truly global.

One of the main barriers to AI development in many parts of the world is access to high-performance computing resources. Of the top 100 high-performance computing clusters capable of training large AI models, none are hosted in a developing country. This stark disparity underscores the challenges developing nations face in harnessing the full potential of AI. Addressing this divide requires innovative approaches such as supporting distributed and federated models of AI development. These models enable the collaborative training of AI across multiple, less powerful computing systems, allowing for broader participation and reducing dependence on high-cost, centralised infrastructure. By fostering these distributed approaches, the global AI ecosystem can become more inclusive, supporting equitable technological progress and closing the gap for regions that lack access to necessary computational power. In response, multilateral initiatives such as the UN's 'AI for Good' have emerged, aiming to democratise access, although the structural gap remains.

Overlooking the social effects of AI and neglecting efforts to close the digital divide could exacerbate income and wealth disparities, affecting workers across all income levels. However, when harnessed effectively, AI has the potential to complement human skills, boosting productivity and fostering innovation among those equipped to navigate and utilise these technologies.

A critical, but often overlooked, issue in AI development is the presence of "missing data" –regions and communities that are underrepresented or not represented at all in global datasets. This lack of representation can significantly impact AI's ability to serve diverse populations in an equitable manner. For example, AI models trained primarily on data from dominant languages and cultures often fail to accurately understand or process content from less-

represented linguistic and cultural groups. This omission not only perpetuates biases within AI systems but also limits the accessibility and benefits of AI for underrepresented communities. The failure to incorporate diverse data sources is a missed opportunity to effectively and inclusively leverage AI's potential in addressing local challenges. Addressing this requires targeted efforts to incorporate data from diverse regions, promote linguistic variety in training data, and build AI systems that are culturally aware and adaptable. Only then can the global community ensure that AI serves all segments of society fairly and inclusively. This is one of the reasons behind the push towards 'AI sovereignty,' where nations and regions are developing their own language models to ensure their cultures, languages, and local contexts are represented in the digital future.

The United Nations' 2030 Agenda for Sustainable Development, with its 17 Sustainable Development Goals (SDGs), provides a comprehensive framework that can guide the development, deployment, and use of AI. By aligning AI investments with global development priorities, stakeholders can leverage the technology's potential to address pressing challenges like poverty, hunger, health disparities, and environmental sustainability. However, without a comprehensive and inclusive approach to AI governance, this potential may not be realised, and the deployment of AI could inadvertently exacerbate existing inequalities and biases. Targeted investments in AI that prioritise sustainable development can help close gaps in resources, talent, and infrastructure, driving equitable progress towards the SDGs. Such an approach highlights the importance of fostering collaboration among governments, industry, academia, and civil society to ensure that AI development serves as a force for inclusive and sustainable global progress.

Educational initiatives that equip diverse segments of society with AI literacy and management skills are crucial for levelling the playing field. For example, government –funded programs that focus on AI training for middle– and low-income workers can be instrumental in reducing inequality. Such programs help individuals leverage AI for productivity gains, thereby reducing the risk of exacerbated economic disparities.

The impact of AI on economic inequality depends significantly on whether individuals experience job displacement or can leverage AI to boost productivity. Reports, such as those from the International Monetary Fund (IMF), suggest that higher-paid workers may see disproportionate income gains due to complementarity with AI, potentially increasing labour income inequality. To address this, policy discussions and think tanks could explore strategies for the equitable distribution of AI's benefits, emphasising education and training programs

tailored to different skill levels and age groups. Initiatives like Germany's dual education system, which integrates vocational training with formal education, could serve as a model for other nations.

National approaches to defining AI property rights and implementing redistributive fiscal policies will be decisive in shaping AI's influence on the distribution of income and wealth. If AI significantly enhances specific jobs, the increased demand for skilled labour could counteract the effects of automation, potentially stabilising or even improving income distribution. This highlights the importance of policy workshops and debates focused on AI property rights and fiscal strategies aimed at formulating inclusive economic policies. Global organisations, such as the United Nations, the UK's Department for Science, Innovation and Technology (DSIT) initiative, which has convened semi-annual summits worldwide on AI safety, and the OECD, could take the lead in facilitating these discussions to ensure widespread consensus and implementation.

The readiness to adopt and benefit from AI varies notably by education level and age. Workers with university degrees and younger individuals are typically more adaptable and likely to transition to AI-complementary roles. Conversely, workers without post-secondary education and older individuals may face increasing susceptibility to AI-induced changes. Targeted mentorship and lifelong learning programs for these vulnerable groups can help mitigate risks and promote a more inclusive transition to an AI-enhanced economy. For example, public-private partnerships and other less burdensome versions of the state's role, which sponsor reskilling initiatives, can play a critical role in supporting these transitions.

The disparity in the adoption and benefits of AI across countries underscores the need to invest in digital infrastructure, human capital, technological innovation, and adaptable legal frameworks. Indices such as the University of Oxford's Government AI Readiness Index and the IMF's AI Preparedness Index indicate that both advanced and emerging economies need to prioritise AI integration while fostering a regulatory environment that allows for growth. International conferences and collaborative research initiatives focused on sharing best practices in digital infrastructure and AI policy can help reduce disparities and foster uniform advancement across regions. Examples of successful collaborations include partnerships in Southeast Asia that pool resources to build AI capabilities in less developed nations.

The key elements for effective AI adoption include sustained investment in education, fostering inclusive experiences in science, technology, engineering, and mathematics, and promoting labour and capital mobility. Legal frameworks

that accommodate new digital business models are also essential. Developing a digitally skilled workforce, supported by foundational infrastructure and innovation, is critical for all economies. Equally important are social safety nets and reskilling programs for workers at risk of job displacement due to AI advancements, which ensure inclusivity and prevent further inequality. Proposing global partnerships and funding mechanisms to support these key elements can help ensure that the benefits of AI are widely accessible.

A delay in adapting to the rapid progress of AI could lead to social instability and insecurity. Implementing universal basic services and considering measures such as Universal Basic Income (UBI) may be essential to reducing the gap between different social groups and ensuring equitable participation in the AI-driven future. Organising global dialogues on innovative social support systems, such as UBI, could provide a platform to explore and pilot approaches that mitigate the socioeconomic impacts of AI, supporting a more stable and inclusive transition. For example, ongoing experiments with UBI and other forms of social safety nets in various parts of the world, from Kenya to cities in the United States, have shown potential benefits in reducing economic insecurity.

The series of AI Safety Summits, from Bletchley Park to Seoul and Paris, in which I have participated, served as a stark reminder of the urgency to collaborate on addressing architectural challenges, the numerous social impacts of AI, and regulatory governance, before we face a potential dystopian future of our own making. Only through coordinated collaboration, rooted in ethical principles and a shared vision for humanity, can we harness the power of AI to benefit everyone. The coming years will be crucial in shaping this path, demanding immediate, collective, and informed action from governments, international bodies, corporations, and civil society. This is why international collaboration and the establishment of a global and adaptable regulatory system that encompasses all countries are vital.

VI. THE NEED FOR INTERNATIONAL COLLABORATION

The governance of AI necessitates a comprehensive and global approach that encompasses political, economic, social, ethical, human rights, technical, and environmental aspects. This comprehensive approach would enable the transformation of the current fragmented mosaic of initiatives, which have so far proven ineffective and counterproductive, into a coherent and interoperable system. Based on international law and aligned with the Sustainable Development

Goals (SDGs), this strategy must be adaptable to different contexts and capable of evolving over time.

The rapid development of AI has led to the emergence of numerous standards adopted by international organisations such as the International Telecommunication Union (ITU), the International Organisation for Standardisation (ISO), the International Electrotechnical Commission (IEC), and the Institute of Electrical and Electronics Engineers (IEEE). While this international proliferation, and its national and even local echoes, have helped create fundamental guidelines, they have also resulted in a fragmented landscape without a common language or unified framework. The absence of agreed-upon definitions for key concepts like fairness, safety, and transparency further complicates cohesive governance efforts, leading to discrepancies between narrowly defined technical standards and those intended to incorporate broader ethical principles. It is precisely this gap that initiatives like the AI Safety Summit series and the UN's AI Advisory Body seek to begin closing by fostering a shared understanding of risks and opportunities.

One of the most immediate benefits of international cooperation in AI governance is the pooling of expertise and resources. Countries with advanced AI capabilities can share knowledge, technical standards, and best practices with nations that may lack the same technological infrastructure or experience. This exchange not only accelerates innovation but also lays the groundwork for ethical and technical standards that benefit all parties involved. For example, joint research initiatives and cross-border educational programs can help developing nations build competencies in AI, fostering a more balanced global technological landscape.

Consistent international standards can simplify the deployment of AI technologies and reduce regulatory friction for multinational corporations. Harmonised regulations ensure that AI applications adhere to shared ethical guidelines and safety protocols, minimising the risk of harmful or unethical practices. This alignment also supports businesses by providing a clearer framework within which they can operate, reducing compliance costs and fostering economic growth.

The global nature of AI's potential risks, such as autonomous weapons or advanced cyberattacks, necessitates immediate and effective collective action. International cooperation allows for the development of joint protocols to manage AI-related crises and reinforces global security by preventing the misuse of AI technologies. Shared data on AI vulnerabilities and coordinated efforts to address them can help countries stay ahead of emerging threats and intercept and neutralise malicious actors.

Collaborative governance frameworks promote transparency and accountability, both of which are crucial for fostering public trust. Joint policies that reflect shared human values encourage citizens to see AI advancements as aligned with their interests. The public perception of AI's legitimacy and safety improves when policies are informed by a range of global perspectives and include checks against bias and unintended consequences.

While a lack of coordination has characterised the governance of AI, historic steps have been taken to overcome this disjointed approach. The adoption by consensus of the first UN General Assembly Resolution on AI in March 2024 marked a milestone, establishing an initial global agreement on the need to guide the technology's development in a safe, secure, and human rights-aligned manner. Although this does not eliminate fragmentation, it represents a fundamental starting point for more cohesive global governance. This effort is complemented by the work of the multi-stakeholder AI Advisory Body, which was created to offer concrete recommendations to help structure international cooperation and align the various initiatives within the UN system, seeking to transform disconnected regimes into an interoperable framework.

VII. CHALLENGES OF INTERNATIONAL COLLABORATION

Achieving convergence among the major global powers presents substantial challenges due to divergent political structures and regulatory philosophies. These differences have materialized into concrete policies: the United States, through its Executive Order on AI, has adopted a pro-innovation approach based on standards and risk management; the EU, with its pioneering AI Act, has established a legal framework based on rights and risk classification; and China, with its measures for generative AI services, integrates state control and direct supervision to align the technology with national security and values. These differences can create friction and hinder efforts to agree on shared principles for AI governance.

The strategic value of AI as a tool for achieving an economic and military advantage amplifies the competition among major powers. This rivalry can lead to protectionist or exclusive policies that prioritise national interests over global collaboration. For example, the escalation of export controls on advanced semiconductors by the United States towards China, aimed at limiting the development of frontier AI, is a clear example of how technological competition can impede open dialogue and data sharing. Without mechanisms to alleviate

these tensions, cooperative frameworks risk being undermined by mutual distrust and geopolitical agendas.

International cooperation must also address the disparities between technologically advanced nations and those still developing their AI capabilities. Without targeted support, less-developed nations may find themselves unable to participate meaningfully in governance discussions, widening the global technological gap. This imbalance could result in regulations that fail to account for diverse economic and cultural contexts, further marginalizing certain regions.

Even when consensus is reached, implementing and enforcing international regulations can be difficult. Each country's legal and political system affects how agreements are integrated into national laws. Furthermore, monitoring and ensuring compliance with international standards require resources and coordinated oversight, which can be difficult to maintain over time.

Global AI governance must go beyond agreements and principles and focus on viable implementation. Ensuring that global commitments translate into tangible outcomes is crucial for significant progress. This includes initiatives for capacity building, particularly in regions with limited access to AI resources, and support for small and medium-sized enterprises (SMEs) so they can compete and innovate effectively. Without effective measures that translate policy into practice, global governance risks becoming a purely ceremonial exercise, devoid of real impact.

VII. THE PATH FORWARD

International collaboration is crucial for developing a global framework that regulates AI by aligning it with human values and safety standards. Although organisations like the UN and the OECD can facilitate dialogue and establish common principles, regulatory convergence among major powers faces challenges due to their different political structures and strategic interests. Even so, the consensus adoption of the first UN resolution on AI and the joint declarations from the AI Safety Summits demonstrate that it is possible to find common ground on fundamental principles such as safety, reliability, and the need for human-centric governance. Nevertheless, this cooperation is essential to ensure an equitable distribution of AI's benefits and a collective management of its risks, allowing the technology to serve the common good of humanity.

Understanding these driving forces and challenges is essential for assessing the opportunities and risks associated with AI. Ignoring these dynamics poses significant threats to the responsible development of AI, including problems

related to participation and the equitable distribution of benefits. Continuous scientific research and interdisciplinary efforts are vital to ensure that AI technologies remain compatible with human rights, social values, and the broader goal of advancing human well-being. Closing the technological and policy gaps through proactive governance and international cooperation will ultimately determine whether AI serves as a powerful tool for human progress or a source of profound social challenges.

The guide for forming new international AI governance institutions must be based on principles that recognise the broader context in which AI operates. These include acknowledging that AI governance does not occur in a vacuum and must be aligned with existing international law, particularly international human rights law. This alignment ensures that the development and deployment of AI promote human dignity, fairness, and the protection of fundamental rights.

ADAPTIVE REGULATORY GOVERNANCE OF AI: TOWARDS A COMPATIBLE, CONTAINED, AND ALIGNED FUTURE

The exponential pace of AI development has outstripped the capacity of traditional regulatory frameworks, creating a significant challenge for effective governance. As mentioned before, the current landscape is characterised by the entry into force of the first regulatory frameworks, such as the EU AI Act, and the proliferation of divergent national approaches, alongside critiques of the difficulty of keeping pace with the technology. This complex environment has led to confusion, stagnation, and elevated risk. Overcoming these challenges requires an adaptive and resilient approach to governance –one that evolves in tandem with AI, aligning technological advancements with societal priorities and values.

AI has the potential to strengthen human rights, well-being, and social functions. This potential underscores the importance of a governance approach that is proactive, evidence-based, and responsive. Science and interdisciplinary expertise must drive this regulatory evolution, aiming to create a governance framework that balances risk management, regulation, investment, and innovation. Such a framework must foster intelligent and adaptive measures. The primary mechanism for this is the 'regulatory sandbox,' actively promoted by countries like the United Kingdom, which allows companies to innovate under the supervision of regulators to ensure new applications are safe before their mass deployment. This permits incentives for business investment and innovation while maintaining a rigorous ethical foundation. Existing efforts, such

as the European Union's AI Act, demonstrate attempts to establish classifications and frameworks for high-risk AI, and they may constitute a starting point for future global adaptation.

Effective AI governance requires striking a balance between risk mitigation and fostering innovation. This balance requires a regulatory approach that encourages investment and entrepreneurship while upholding global principles, such as those outlined in the Charter of the United Nations and the Universal Declaration of Human Rights. However, this task is complicated by the varied regulatory strategies of major global actors mentioned earlier. The unique political, economic, and cultural priorities of each region contribute to a competitive environment that challenges cohesive international collaboration in AI governance.

The current state of AI governance bears parallels to aspects of the Cold War concept of "mutually assured destruction" (MAD), yet with distinct characteristics and implications. AI has the potential to devastate human civilisation, like the use of nuclear weapons. However, unlike nuclear technology, AI is an evolving cognitive force, capable of self-improvement and the development of complex, opaque systems that could potentially escape human control. This introduces a new concept, a "path to assured destruction" (PAD), where the uncontrolled development of AI poses imminent existential risks. These risks include the creation of autonomous weaponry with unpredictable consequences, making global cooperation essential. Recent efforts, crystallised in the consensus resolution of the UN General Assembly and in the joint declarations from the AI Safety Summits, provide valuable frameworks and political momentum for forging universal principles. However, these need to be translated into enforceable standards through global agreements.

Progressing towards the compatibility, containment, and alignment of AI requires a commitment to universal principles that guide the development of adaptive governance. Ensuring that the development of artificial superintelligence (ASI) moves forward only with incontrovertible evidence of its controllability and benefits for human survival is a key component. This involves comprehensive risk assessments and the implementation of safeguards that prioritise the well-being of humanity. It is also essential to integrate public engagement and transparency into governance efforts, fostering trust and ensuring that diverse societal perspectives shape policy.

Fostering AI innovation must occur within frameworks of human oversight and legal accountability, ensuring that the decisions and actions of AI systems are always attributable to human operators. This approach supports transparency

and safeguards against unethical, negligent, or harmful practices in the development and use of AI, enforcing strict guidelines and legal consequences for violators. Examples of this include regulatory sandboxes that allow for the confirmation and controlled growth of AI applications, facilitating the fostering and steering of innovation while monitoring potential risks. Combating digital information theft and unethical practices also requires robust cybersecurity measures, clear legal frameworks, and international cooperation to protect individual and collective digital rights.

To foster trust and accountability, the role of AI in content creation must be governed by clear guidelines that ensure AI-generated content is identifiable. On this front, technical solutions have emerged, such as the C2PA (Coalition for Content Provenance and Authenticity) standard, a multi-stakeholder initiative that aims to create a verifiable 'content credential' allowing the origin and modifications of an image or video to be tracked. This includes preventing AI from impersonating individuals or organisations without explicit, informed consent, thereby safeguarding identity and authenticity. A global dialogue is needed to negotiate with AI corporations, urging them to allocate a portion of their profits towards enforcing standards, mitigating socioeconomic impacts, and supporting affected communities. Initiatives like public-private partnerships and stakeholder dialogues can facilitate such agreements and promote cooperative progress.

Promoting the responsible use of AI in addressing public safety, human rights, and intellectual property rights is crucial, as is ensuring alignment with social values and legal norms. The establishment of mechanisms to monitor, intervene, and prosecute those who misuse AI for hostile or negligent purposes reinforces legal and ethical standards. Likewise, clarifying and enforcing legal liability and accountability concerning AI's actions ensures that clear guidelines exist for redress and compensation when necessary. Public engagement, through consultations and participatory decision-making, can help shape these policies to reflect societal values and concerns.

The universal adoption of these principles aims to foster a balanced approach to AI development, minimizing existential risks while prioritizing human safety and encouraging ethical practices. This approach seeks to encourage innovation while addressing the potential challenges associated with advanced AI technologies. By building a flexible and adaptive governance system rooted in international cooperation and ethical oversight, societies can harness AI's potential for widespread benefit while protecting against its risks. The path ahead requires an unwavering commitment from governments, the private

sector, and civil society to create a future where AI advances human well-being without compromising safety or ethical standards.

TOWARDS AN ADAPTIVE REGULATORY GOVERNANCE SYSTEM

While institutional and regulatory adaptability progresses incrementally, technological change –especially in General-Purpose Artificial Intelligence–advances at an exponential pace. This growing disparity underscores the urgent need for innovative governance approaches that can evolve in parallel with the rapid development of AI. Without adaptive and responsive regulatory systems, the gap between the pace of technological evolution and regulatory oversight will continue to widen, reducing the effectiveness of strategic and organisational management. This mismatch presents a critical challenge that must be addressed to prevent regulatory obsolescence and safeguard societal interests.

The pressing need for an adaptive regulatory system that can respond in real-time to the incessant emergence of new AI technologies is becoming increasingly apparent. If governance structures remain static or only advance gradually, they risk becoming obsolete at an accelerated rate, which could lead to unforeseen consequences with potentially significant global implications. It is unlikely that traditional, human-driven regulatory frameworks currently in force will keep pace with the exponential growth of AI and related technologies. This delay means that well-intentioned reforms and ethical guidelines may become ineffective, or even counterproductive, if they cannot adapt quickly enough to the changing landscape.

Technological synergies could play a fundamental role in overcoming these challenges. Emerging technologies, such as blockchain, machine learning, and advanced data analytics, offer powerful tools to enhance adaptive governance. For example, blockchain can provide transparency and traceability in regulatory processes, while AI-powered monitoring systems can flag non-compliance in real-time. This is already happening in the financial sector, where regulators use AI-based 'RegTech' to supervise markets on a large scale, enabling faster corrective actions and ensuring adherence to regulations. These tools could contribute to building a responsive governance system capable of continuous monitoring and rapid adaptation to new developments.

However, implementing an adaptive governance system presents significant challenges. Resistance from established institutions, which may be reluctant to overhaul traditional regulatory practices, is a common obstacle. Furthermore,

the costs associated with continuously updating and maintaining an adaptive framework can be considerable, posing budgetary and logistical concerns for many governments and organisations. Another pressing issue is the risk of regulatory capture, where powerful technology companies exert undue influence over the regulatory process to shape outcomes that favour their interests. Addressing these challenges requires a balanced approach that ensures the integrity and impartiality of governance structures while fostering collaboration between the public and private sectors.

To manage and contain the development and deployment of AI in a safe, ethical, and responsible manner, there is an urgent need for a regulatory governance system that is dynamic, flexible, and highly adaptable. Such a system must be capable of evolving in step with technological advances. The development of this governance system must involve global participation, leveraging contributions from governments, private companies, academic institutions, civil society, and the general public. This inclusive approach ensures that the system is transparent, accountable, and characterised by precise mechanisms for monitoring, feedback, and modification. Versatility is crucial for responding effectively to rapid and often unpredictable technological changes.

In alignment with these principles, any regulatory body or system designed to manage the strategic oversight of AI must fulfil essential roles. First, it must continuously monitor and guide the development and implementation of emerging AI technologies to maintain an ongoing understanding of their potential impacts and risks. Second, the regulatory process must be efficient, reliable, and inclusive, ensuring that all relevant stakeholders are involved and represented. Third, this system must be empowered to block the progression of AI technologies that pose significant risks or could lead to adverse outcomes, thus acting as a safeguard for public well-being. Finally, fostering global engagement and cooperation is crucial for sharing knowledge, aligning standards, and leveraging collective expertise to mitigate risks and maximise shared benefits.

For the regulatory governance structure to effectively oversee AI development, it must exhibit the flexibility to adapt to rapid and unpredictable changes. A successful strategy would ensure that regulatory oversight does not stifle innovation but instead fosters an environment where technological growth aligns with ethical principles and serves the global public interest. Achieving this balance requires a collaborative approach, where diverse voices contribute to the decision-making process, ensuring that outcomes are equitable and that the benefits of technological progress are shared among countries and across society.

Developing such a governance system will require integrating digital tools and AI to support and drive adaptive regulatory practices. The system must be capable of performing real-time data analysis, scenario planning, and dynamic policy adjustment to maintain its relevance. Furthermore, international cooperation must be prioritised to create standardised regulations that are flexible enough to accommodate different regional and cultural contexts while upholding fundamental principles that protect human rights and promote the ethical use of AI.

To address inconsistencies and close the gap between different AI standards, a coordinating body is needed. This function could be assumed by a new 'clearing house' under the UN or evolve from the nascent global network of AI Safety Institutes, thereby giving it a formal mandate to align technical and ethical standards. This body would bring together expertise from international scientific panels, national and international standards organisations, technology companies, and civil society representatives.

Such a platform could help align technical and ethical standards globally, ensuring that AI governance is coherent, comprehensive, and grounded in shared values. A unified approach to AI standardisation would not only simplify governance but also support equitable and transparent development. The participation of diverse stakeholders in this process is crucial for promoting trust and accountability. By acting as a clearing house, the UN could facilitate the alignment of existing standards and create a common understanding that supports both technical excellence and ethical responsibility.

The next step involves solidifying these frameworks through pilot programs, cross-border agreements, and adaptive legislation that can be adjusted as new data and outcomes emerge. This will require commitment from all stakeholders and an acknowledgement of the complexity involved in adapting governance. Only by adopting such measures can the global community navigate the challenges posed by AI, ensuring its development continues to benefit humanity without undermining social values or safety.

A major obstacle is moving beyond the geopolitical competition and mutual distrust among technological powers, which aggravate their differences. It is vital to agree on confidence-building measures and international cooperation. Despite the difficulties, opportunities exist for incremental progress through collaboration in areas of common interest, such as AI safety and ethics. The role of international platforms, such as the United Nations, the **AI Safety Summit series**, and the OECD, is crucial for facilitating dialogue and establishing shared standards.

To move forward, AI governance must be flexible, adaptive, and human-centric, with principles of transparency and accountability that reinforce public trust. The inclusion of developing countries from the so-called Global South, along with the implementation of effective monitoring mechanisms, will ensure that the benefits of AI are distributed equitably, thereby preventing the exacerbation of global inequalities and their associated negative impacts on worldwide insecurity and uncertainty. A collaborative approach, integrating governments, the private sector, academia, and civil society, is crucial to achieving governance that strikes a balance between innovation and safety in the context of accelerated technological evolution. Everyone should be able to contribute to, participate in, and benefit from these efforts.

An inevitable conclusion is that, if humanity does not unite to prevent the catastrophic risks of an uncontrolled AI, such as the proliferation of autonomous weapons, the accelerated design of new pandemics, or the loss of control over self-improving systems, it will irremediably face existential perils that could condemn the new era in human history to an existential catastrophe.

THE NEED FOR AN AI-BACKED ADAPTIVE REGULATORY GOVERNANCE SYSTEM

Given the growing challenges and risks of the advancement of artificial intelligence (AI), an AI-backed adaptive governance system that can effectively manage its development and evolution is indispensable. The integration of AI into governance extends beyond traditional oversight, providing tools that enable a dynamic response to a constantly evolving technological environment. Adopting adaptive frameworks ensures that regulatory measures can keep pace with AI, mitigating risks and maximising benefits.

AI can transform the adaptability of regulatory frameworks by offering agile, data-driven responses. By analysing large volumes of information, such as economic indicators and social trends, AI systems can identify emerging patterns and shifts in public sentiment or market conditions. This predictive capability is invaluable for updating policies and regulations, ensuring their relevance and effectiveness. For example, AI-powered compliance tools are used in financial markets to detect irregularities and manage risks in real-time, improving the responsiveness of regulators.

In the realm of legal analysis, AI can evaluate existing normative frameworks, identify inconsistencies, and suggest updates to harmonise regulations across different regions. By simulating the impact of new regulations under hypothetical

scenarios, AI helps to reveal gaps in legal structures and facilitates a proactive approach to legislative adaptation. Furthermore, AI systems can assist in drafting legal documents aligned with current laws and precedents, ensuring coherence and precision.

AI's monitoring capability is equally essential in overseeing compliance across industries. By detecting patterns of non-compliance and identifying problematic areas, AI enables regulatory bodies to act proactively. This optimises the allocation of resources, allowing regulators to focus on strategic decision-making. A practical and imminent case is that of the new European Union AI Office. To effectively oversee compliance with the AI Act across the vast European single market –excluding the purely military sphere, which falls outside its mandate– this body will inevitably require the use of automated and intelligent supervisory tools as described here.

AI can also provide insights into public sentiment by analysing social media and online discourse. By identifying trends and concerns, policymakers can adjust regulations to align them with societal expectations and improve the representativeness of policies. This real-time feedback loop reinforces transparency and public trust, ensuring that policies evolve in sync with the needs of citizens.

However, to ensure that AI operates fairly and without bias, it is fundamental to implement regular audits and updates. This includes the use of diverse data and the application of bias-detection algorithms to ensure equitable standards. Precise mechanisms for interpreting AI-generated decisions must be established, with final responsibility remaining in the hands of human decision-makers. AI must be a supportive tool, not a substitute, thereby guaranteeing accountability and alignment with democratic values and human rights.

The adoption of AI in governance faces challenges, such as resistance from institutions and policymakers who may fear a lack of transparency or job displacement. To overcome these barriers, it is necessary to communicate the benefits of AI, implement pilot programs, and offer training to demonstrate its value in regulatory management.

The commitment to ethical principles is non-negotiable in the integration of AI. Transparency, fairness, and accountability must guide the evolution of these systems. Periodic evaluations and adaptation to technological advances are crucial to prevent the obsolescence of AI in governance. An AI-backed adaptive governance approach can also influence international cooperation, serving as a model for regulatory agreements and promoting the global standardisation of policies.

International organisations like the International Atomic Energy Agency and the International Thermonuclear Experimental Reactor have shown that it is possible to manage complex technologies with authorised mandates and inclusive decision-making. These examples can inspire similar approaches in AI governance, as reflected in the growing proposals to create an **'International Artificial Intelligence Agency'** (IAIA) with a mandate and inspection capabilities similar to those of the IAEA, provided that the Global South is included in the processes to ensure equitable representation.

AI's ability to support an adaptive regulatory system that evolves at the pace of technology can empower governance structures, foster mutual trust and serve as a benchmark for unified global standards.

The cost of not implementing an adaptive regulatory governance system is high: an uncontrolled technological race, a widening of the digital divide, and the erosion of public trust. It is essential to establish universal principles that prioritise human rights, safety, and well-being, along with flexible regulatory frameworks that guide the development of AI towards the common good. Ultimately, learning to govern such a powerful technology with the help of the technology itself may be the defining challenge, and the greatest opportunity, for governance in the 21st century.

ARTIFICIAL INTELLIGENCE AND DEMOCRACY: A SYNERGISTIC AND CONFRONTATIONAL RELATIONSHIP

CONTENT:

I.Introduction **II.** Conceptual Foundations **III.** The Synergistic Dimension: Opportunities of AI for Democracy **IV.** The Confrontational Dimension: Risks and Challenges **V.** Toward a Democratic Governance of AI **V.** Conclusion **VII.** Bibliography

> EBER OMAR BETANZOS TORRES

President of National Institute of Public Administration (INAP, A.C.)

I. INTRODUCTION

At the threshold of the 21st century, artificial intelligence (AI) not only redefines the boundaries of technology but also bursts into the very core of social organization: democracy. Defined by Russell and Norvig as "the construction of agents that act appropriately in their environment" (Russell and Norvig 2021, 22), AI has transcended its origins in machine learning to become a pivotal actor in political processes, from moderating online debates to predicting electoral behaviors. However, this integration is not univocal. As Jungherr notes, "the consequences of current AI are more specific to democracy than the broad speculations about AGI capabilities imply" (Jungherr 2023, 1). Thus, a synergistic and confrontational relationship emerges: the former, where AI amplifies the collective voice and optimizes governance; the latter, where it threatens to undermine equity and individual autonomy.

The present article is structured into four main sections. First, the conceptual foundations of AI and democracy are outlined, drawing on 20 key bibliographic sources to contextualize the debate. Second, the synergistic dimension is analyzed, illustrated with textual citations from empirical studies. Third, the confrontational aspects are explored in depth, highlighting ethical and political risks. Finally, recommendations are proposed for a democratic governance of AI. This interdisciplinary approach, integrating political philosophy, data science, and constitutional law, seeks not only to diagnose but also to guide toward a future where technology serves the common good.

The relevance of this topic is undeniable. In 2025, with global elections influenced by generative tools like ChatGPT, the question is no longer whether AI will affect democracy, but how to mitigate its adverse impacts while harnessing its benefits. As the UNESCO warns in its Recommendation on the

Ethics of Artificial Intelligence, "the impacts of AI on political life are direct and multifaceted" (UNESCO 2021, 5). This work, therefore, contributes to that urgent call for critical reflection.

II. CONCEPTUAL FOUNDATIONS

A. DEFINITION AND EVOLUTION OF AI IN DEMOCRATIC CONTEXTS

AI, in its narrow form (narrow AI), focuses on specific tasks such as natural language processing or predictive analysis, distinguishing itself from the hypothetical artificial general intelligence (AGI). Nilsson describes it as "the quality that enables an entity to function appropriately and with foresight in its environment" (Nilsson 2010, 8). In the democratic sphere, this evolution has been accelerated by big data and machine learning, enabling everything from personalized electoral campaigns to the simulation of legislative scenarios.

Historically, the intersection of AI and democracy dates back to the 1960s with cybernetics, but it exploded in the 2010s with scandals such as Cambridge Analytica. Donahoe and Metzger argue that "AI poses substantial threats to democratic values" (Donahoe and Metzger 2019, 115), while Ojo et al. highlight its potential to "improve the quality of public services" (Ojo et al. 2019, 45). This conceptual duality is the axis of our exploration.

B. DEMOCRACY IN THE DIGITAL ERA: THEORETICAL FRAMEWORKS

Democracy, according to Dahl, entails "effective participation, equality in voting, and adult inclusion" (Dahl 1989, 221). In the digital era, these pillars are transformed: participation expands through online platforms, but equality is threatened by biased algorithms. Bibliographically, works such as Weapons of Math Destruction by O'Neil illustrate how data increase inequality (O'Neil 2016). Therefore, the central question is not whether technology shapes democracy, but how it does so and under what conditions it can do so legitimately.

One of the deepest analyses of this intersection comes from the Spanish political philosopher Daniel Innerarity. In his article "The Impact of Artificial Intelligence on Democracy", published in the Revista de las Cortes Generales, Innerarity posits that AI is not merely a technical instrument but a new space of power that redefines popular sovereignty:

> *"Artificial intelligence does not only transform the means of politics but reconfigures the conditions of possibility for democratic deliberation. By processing massive data, AI can anticipate preferences, simulate consensuses, and, in some cases, substitute human judgment. This poses a paradox: can a machine be more democratic than citizens?" (Innerarity 2020, 89)*[1] [2].

Innerarity identifies three structural tensions:

1. Algorithmic opacity vs. democratic transparency;
2. The speed of AI vs. deliberative slowness;
3. Extreme personalization vs. the general interest.

In his analysis, AI is not neutral: "Algorithms do not merely reflect preferences but produce them" (Innerarity 2020, 92)[3]. This argument resonates with Cathy O'Neil, who in Weapons of Math Destruction warns that predictive models reinforce preexisting inequalities: "Algorithms are opinions embedded in code" (O'Neil 2016, 23)[4].

On the other hand, Hélène Landemore offers an optimistic yet conditioned perspective. In her work for the IMF, she argues that AI can be a catalyst for open democracy if designed with deliberative principles:

> "Generative AI could maximize deliberation by processing millions of citizen inputs in real time, creating a new democratic Renaissance based on collective intelligence" (Landemore 2023, 31)[5].

However, Landemore qualifies: this possibility only materializes if mechanisms for participatory governance of algorithmic design are implemented. Without them, AI becomes a tool of technological elites.

Mathias Risse, from the Kennedy School, introduces a historical dimension. In his discussion paper, he argues that AI is not the first technology to challenge

1 Innerarity, Daniel. 2020. "El impacto de la inteligencia artificial en la democracia". Revista de las Cortes Generales 109: 89.

2 Innerarity, Daniel. 2020. "El impacto de la inteligencia artificial en la democracia". Revista de las Cortes Generales 109: 89.

3 *Ibid*., p.92.

4 O'Neil, Cathy. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Broadway Books, p.23.

5 Hélène Landemore, "Fostering More Inclusive Democracy with AI", Finance & Development 60, no. 4 (2023): 30.

democracy –the printing press, radio, and television did so before– but it is the first to operate on a global scale and in real time:

> "AI is not a natural ally of democracy; it requires active design, political will, and robust institutional frameworks" (Risse 2021, 12)[6].

Risse proposes an analytical framework based on three axes:

- State capacity (can the state regulate AI without being captured by it?);
- Societal capacity (do citizens have the skills to participate in the algorithmic era?);
- Epistemic capacity (how is collective knowledge constructed in a machine-mediated environment?).

This approach is complemented by the work of Paul Nemitz, former principal advisor to the European Commission, who in Philosophical Transactions of the Royal Society defends that constitutional democracy must prevail over the logic of the algorithmic market:

> *"The opacity of AI violates the basic democratic requirements of accountability and public control" (Nemitz 2018, 8)*[7].

Nemitz advocates for digital constitutionalism, where fundamental rights (privacy, freedom of expression, non-discrimination) apply directly to AI systems. In the Latin American context, Fernando Filgueiras analyzes how AI in public administration can strengthen or weaken democracy depending on its implementation. In his study for CLAD, he concludes:

> *"AI can optimize public services, but only if subjected to principles of democratic governance: transparency, auditing, and citizen participation" (Filgueiras 2022, 28)*[8].

Filgueiras documents cases in Brazil where AI systems for allocating social resources reduced corruption by 18% but also generated digital exclusion in rural communities.

---

6 Risse, Mathias, "Artificial Intelligence and the Past, Present, and Future of Democracy". Carr Center Discussion Paper. Cambridge, MA: Harvard Kennedy School (2021): 12.

7 Nemitz, Paul, "Constitutional Democracy and Technology in the Age of Artificial Intelligence". Philosophical Transactions of the Royal Society A 376 (2133), 2018: 8.

8 Filgueiras, Fernando. 2022. "AI in Public Administration: Governance Challenges". Revista del CLAD Reforma y Democracia, no. 79: 28.

Finally, Frank Pasquale in The Black Box Society introduces the concept of algocracy: a regime where real power resides in those who control data and algorithms, not in elected representatives. Pasquale warns:

"The black-box society threatens to replace representative democracy with an opaque technocracy" (Pasquale 2015, 200)[9].

This theoretical framework –combining Innerarity, Landemore, Risse, Nemitz, Filgueiras, and Pasquale– allows us to understand digital democracy not as a technical system but as a space of normative conflict between algorithmic efficiency and political legitimacy.

III. THE SYNERGISTIC DIMENSION: OPPORTUNITIES OF AI FOR DEMOCRACY

A. IMPROVEMENT OF PARTICIPATION AND CITIZEN DELIBERATION

The synergy between AI and democracy is primarily manifested in the amplification of the citizen voice. Landemore proposes that AI can "strengthen democratic institutions if it ensures that citizens' voices are heard" (Landemore 2023, 28)[10]. A paradigmatic example is OpenAI's "Democratic Inputs to AI" program, which funds algorithms for human deliberation, enabling the processing of massive contributions in public consultations. In experiments such as the Conference on the Future of Europe, where 700,000 citizens debated reforms, AI facilitated the synthesis of diverse opinions, reducing human biases in data aggregation (OECD 2020, 15).

"AI can improve democracy through algorithms that foster inclusive deliberation, subsidizing teams working on transparency" (Landemore 2023, 30)[11].

This aligns with Ojo *et al.,* who document how AI in public services "fosters citizen trust by predicting needs accurately" (Ojo et al. 2019, 48)[12]. In Latin America, cases such as the Chilean constituent process (2020-2022) used AI

9 Pasquale, Frank, The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press, 2015: 200.

10 Hélène Landemore, *op. cit.*, p.28.

11 *Ibid*, p.30.

12 Adelaoye Ojo *et al.*, "Enhancing Democratic Health through AI", Government Information Quarterly 36, no. 3 (2019): 48.

tools to map regional preferences, increasing the representation of marginalized groups. For instance, the "Escucha Digital" system processed 2.4 million citizen contributions, identifying 147 common demands with an 87% match to the Convention's final proposals (CEPAL 2023, 45)[13].

B. OPTIMIZATION OF GOVERNMENTAL PROCESSES

Another synergistic pillar is administrative efficiency. Dwivedi et al. argue that "AI drives economic opportunities and strengthens democratic health" (Dwivedi *et al.* 2021, 105)[14]. In Estonia, AI-driven e-government has reduced bureaucracy by 30%, enabling secure online voting (Filgueiras 2022, 25).[14]

> *"Some studies recognize that the expansion of AI can improve democratic health by optimizing public services and generating trust" (Duberry 2022, 436)*[15].

This optimization extends to crisis prediction: AI models in the EU have anticipated disinformation outbreaks, allowing early interventions (UNESCO 2021, 10)[16]. In Latin American contexts, Vázquez-Barrio and Salazar highlight how AI in journalism "combats disinformation, strengthening public opinion" (Vázquez-Barrio and Salazar 2023, 99)[17]. A study in Colombia showed that electoral chatbots increased youth participation by 22% (Guerrero-Solé and Pérez-Altable 2025, 150)[18].

13 CEPAL, Inteligencia Artificial y Gobernanza Democrática en América Latina (Santiago: CEPAL, 2023), 45.

14 Yogesh K. Dwivedi *et al.*, "Fostering Citizen Trust with AI", International Journal of Information Management 58 (2021): 105.

15 Jérémy Duberry, "Artificial Intelligence and Democracy: Risks and Promises", Internet Policy Review 28, no. 4 (2022): 436.

16 UNESCO, Recommendation on the Ethics of Artificial Intelligence (Paris: UNESCO, 2021), 10.

17 Tamara Vázquez-Barrio & Idoia Salazar García, eds., Inteligencia artificial, periodismo y democracia (Valencia: Tirant lo Blanch, 2023), 99.

18 Frederic Guerrero-Solé and Laura Pérez-Altable, eds., La democracia en riesgo (Barcelona: UOC, 2025), 150.

C. INCLUSION AND EQUITY: TOWARD A POSITIVE ALGORITHMIC DEMOCRACY

AI can democratize access to political knowledge. Risse suggests that "with active design, AI serves democratic purposes" (Risse 2021, 14)[19]. In India, AI platforms translate parliamentary debates in real time into 22 official languages, reaching 300 million non-Hindi speakers (Summerfield et al. 2025, 1238)[20].

> *"Generative AI could maximize deliberation by processing massive inputs, creating a new democratic Renaissance" (Landemore 2023, 31)*[21].

However, this synergy requires ethical safeguards, as proposed in the "public interest AI" framework (Jungherr 2023, 5)[22]. In Colombia, the "Voto Informado" chatbot increased youth participation in 2023 local elections by 22%, according to a University of the Andes study (Guerrero-Solé and Pérez-Altable 2025, 150)[23]. These results suggest that deepening technological tools and AI applications has political implications, such as in electoral participation levels.

IV. THE CONFRONTATIONAL DIMENSION: RISKS AND CHALLENGES

A. DISINFORMATION AND ELECTORAL MANIPULATION

The confrontation arises with the proliferation of deepfakes and hoaxes. Kreps and Kriner warn: "AI threatens democracy by flooding media with confusion for voters and officials" (Kreps and Kriner 2023, 124)[24]. In the 2024 U.S. elections, generative AI created fake videos that polarized the electorate (Journal of Democracy 2024, 5)[25].

19 Risse, Mathias, *op. cit.*, p.14.
20 Christopher Summerfield *et al.*, "The Impact of Advanced AI Systems on Democracy", Nature Human Behaviour 9 (2025): 1238.
21 Hélène Landemore, *op. cit.*, p.31.
22 Andreas Jungherr, "Artificial Intelligence and Democracy", Social Media + Society 9, no. 2 (2023): 5.
23 Frederic Guerrero-Solé and Laura Pérez-Altable, eds., *op. cit.*, p.150.
24 Sarah Kreps and Douglas Kriner, "How AI Threatens Democracy", Journal of Democracy 34, no. 1 (2023): 124.
25 Artificial Intelligence and Democracy", Journal of Democracy, January 2024, 5.

> *"Malicious actors use AI to manipulate perceptions and disrupt electoral processes" (Summerfield et al. 2025, 1240)*[26].

In Latin America, the 2021 Ecuadorian case illustrates how machine learning-driven political spots on TikTok amplified populist ideologies (Innerarity 2020, 95)[27].

Disinformation and electoral manipulation represent a profound confrontational dimension of artificial intelligence in democratic processes, where generative tools enable the mass production of deepfakes and tailored propaganda that erode voter discernment. Empirical evidence from the 2024 U.S. elections reveals that AI-synthesized videos garnered millions of views, systematically polarizing public discourse. In Latin America, machine-learning-driven microtargeting on platforms like TikTok has amplified populist narratives, as documented in Ecuador's 2021 campaign (Innerarity 2020, 95)[28]. Such practices not only distort electoral integrity but also necessitate robust regulatory frameworks to safeguard deliberative autonomy. Ultimately, unchecked algorithmic manipulation risks transforming democratic competition into a theater of simulated realities.

B. POLARIZATION AND EROSION OF TRUST

Recommendation algorithms foster echo chambers. O'Neil denounces: "Big data increase inequality and threaten democracy" (O'Neil 2016, 120)[29]. In Brazil, AI on social networks exacerbated polarization during the 2016 impeachment (Zidouemba 2025, 10)[30].

> "AI erodes trust by creating epistemic bubbles where truth is lost in fiction" (Risse 2021, 18)[31].

EU studies show a 40% increase in institutional distrust attributable to algorithmic biases (Nemitz 2018, 8)[32]. Polarization and the erosion of institutional

26 Summerfield *et al.*, *op. cit.*, p.1240.
27 Innerarity, Daniel, *op. cit.*, p.95.
28 *Loc. cit.*
29 O'Neil, Cathy, *op. cit.*, p.120.
30 Marie-Thérèse Zidouemba, "Governance and Artificial Intelligence", Discover Artificial Intelligence 5, no. 1 (2025): 10.
31 Risse, Mathias, *op. cit.*, p.18.
32 Paul Nemitz, "Constitutional Democracy and Technology", Philosophical Transactions of the Royal Society A 376, no. 2133 (2018): 8.

trust constitute a critical confrontational axis in the interplay between artificial intelligence and democracy, as algorithmic recommendation systems systematically reinforce epistemic silos and amplify affective divides. In Brazil's 2016 impeachment crisis, AI-curated social media feeds exacerbated societal cleavages, deepening distrust in public institutions. European studies further reveal a 40% surge in institutional skepticism attributable to perceived algorithmic bias. This erosion of trust not only undermines civic cohesion but also weakens the legitimacy of democratic governance. Consequently, mitigating polarization demands transparent algorithmic design and mandatory bias audits to restore public confidence.

C. SURVEILLANCE AND INEQUALITY: TOWARD AN ALGOCRACY

The confrontation culminates in surveillance. Wright ed. (2019) analyzes how China uses AI for repression, exporting models to democracies (Wright 2019, 50)[33]. In Europe, the GDPR fights this, but gaps persist (Pasquale 2015, 200)[34].

> "The opacity of AI violates democratic requirements, turning it into a propaganda tool" (Peyron 2024, 5)[35].

In Mexico, biometric electoral systems have been criticized for exclusion risks (Dávila Cortés 2025, 170)[36]. However, political inequity and algorithmic power emerge as central confrontational forces in the AI-democracy nexus, wherein opaque decision-making systems concentrate authority among data controllers, marginalizing underrepresented groups and distorting democratic representation. This power asymmetry not only entrenches privilege but also violates democratic equality. Addressing algorithmic inequity thus requires enforceable data inclusion mandates and citizen oversight to redistribute technological authority.

33 Nicholas D. Wright, ed., Artificial Intelligence, China, Russia, and the Global Order (Maxwell AFB: Air University Press, 2019), 50.

34 Pasquale, Frank, *op. cit.*, p.200.

35 Léana Peyron, "Inteligencia artificial: ¿una amenaza para nuestras democracias?" MundoSur, February 2024, 5.

36 Ian Eric Dávila Cortés, "Los derechos político-electorales en México", Revista de Derecho Político, no. 123 (2025): 170.

V. TOWARD A DEMOCRATIC GOVERNANCE OF AI

A. GLOBAL REGULATORY FRAMEWORKS

UNESCO proposes "ethical governance to mitigate negative impacts" (UNESCO 2021, 15)[37]. The EU AI Act classifies electoral systems as "high risk" (García Marcos 2023, 50)[38]. Key recommendations include: 1) Mandatory human audits; 2) Co-design with citizens; 3) Focus on algorithmic equity.

AI governance requires a multilevel approach. UNESCO establishes in its 2021 Recommendation a universal ethical framework:

> "States must ensure that AI systems respect human rights, cultural diversity, and the environment, with special attention to their impact on democracy" (UNESCO 2021, 15)[39].

This document, adopted by 193 countries, proposes 10 guiding principles, including transparency, accountability, and public participation.

The European Union has advanced with the AI Act (2024), the world's first comprehensive legislation. It classifies AI systems into four risk levels: The European Union has advanced decisively with the AI Act (2024), establishing the world's first comprehensive regulatory framework for artificial intelligence and classifying systems into four risk levels to balance innovation with democratic safeguards. Unacceptable risk systems –such as subliminal manipulation and social scoring– are outright prohibited; high-risk applications like public facial recognition and electoral AI mandate rigorous impact assessments and independent audits; limited-risk tools including chatbots require transparency disclosures; while minimal-risk systems like spam filters face no obligations. This graduated approach ensures that systems influencing democratic processes undergo stringent oversight, with mandatory traceability of training data and human oversight requirements. By prioritizing electoral integrity and citizen rights, the AI Act sets a global benchmark for responsible algorithmic governance.

The AI Act requires high-risk systems to be audited by independent third parties and training data to be traceable (European Commission 2024, 22).

---

37 UNESCO, Recommendation, 15.

38 Julián García Marcos, "Inteligencia artificial", in Marco normativo de la UE, ed. Eloy Velasco Núñez (Madrid: Editorial La Ley, 2023), 50.

39 UNESCO, *op. cit.*,15.

In Latin America, the Model AI Law for Latin America and the Caribbean (CEPAL, 2025) proposes a regional approach based on:

1. Digital sovereignty
2. Personal data protection
3. Digital inclusion
4. Training in algorithmic literacy

This model, inspired by the European GDPR but adapted to local realities, has been adopted as pilot law by Chile and Uruguay (CEPAL 2025, 78)[40].

B. NATIONAL STRATEGIES

In Latin America, integrating AI into civic education is suggested (Vázquez-Barrio and Salazar 2023, 150)[41]. Civic education emerges as a cornerstone of national strategies to regulate artificial intelligence in its relationship with democracy, equipping citizens with the critical competencies to engage meaningfully in algorithmic governance and counter systemic risks. Finland's inclusion of algorithmic bias, data privacy, and source verification in its national curriculum has enabled 94% of 15-year-olds to identify deepfakes, fostering informed civic participation. By cultivating algorithmic literacy, such initiatives empower individuals to audit public AI systems, challenge disinformation, and demand transparency from both state and corporate actors. This educational foundation not only mitigates polarization and manipulation but also transforms passive users into active co-designers of democratic technology. Without widespread civic education, regulatory frameworks risk remaining elite-driven and disconnected from public needs. Thus, integrating AI ethics into compulsory schooling is essential for sustaining democratic legitimacy in the digital era.

> "Strengthening democratic institutions and democratizing AI is key" (Guerrero-Solé and Pérez-Altable 2025, 200)[42].

Below are some of the most representative local initiatives:

A. Estonia: Leading Digital Democracy

Estonia operates 99% of its public services online. Its X-Road system enables secure interoperability between state databases, with AI for fraud detection. In 2023, it prevented €42 million in diversions.

40 CEPAL, *op. cit.*,

41 Vázquez-Barrio and Salazar, *op. cit.*, p. 150.

42 Guerrero-Solé and Pérez-Altable, *op. cit.*, p.200.

B. Brazil: Challenges and Lessons

The "Gov.Br" program uses AI to simplify procedures but faces criticism for digital exclusion: 30% of the rural population lacks internet access.

C. India: Scale and Diversity

The Aadhaar system (1.3 billion biometric identities) uses AI for social subsidies but has been criticized for identification errors affecting 10 million beneficiaries.

C. FUTURE: SYNERGISTIC BALANCE. PARTICIPATORY GOVERNANCE

With political will, AI can catalyze a "new Renaissance" (Landemore 2023, 32)[43]. The future of artificial intelligence and democracy hinges on achieving a synergistic balance through participatory governance, where citizens actively co-design, audit, and oversee algorithmic systems to ensure they serve collective deliberation rather than elite control. By embedding democratic principles –transparency, accountability, and equity– into the algorithmic infrastructure, societies can catalyze the new democratic forms. This participatory paradigm transforms AI from a disruptive force into a deliberative enhancer, strengthening institutional trust and civic agency. Ultimately, only through co-governed technology can democracies navigate the digital era without sacrificing sovereignty to code. The path forward is not technological determinism, but democratic intentionality.

- Citizen Audits of Algorithms
  Proposed by AlgorithmWatch and adopted in Amsterdam: citizen panels review public algorithms every 6 months. In 2024, they identified gender biases in a housing allocation system.
- Algorithmic Co-Design
  The "vTaiwan" project allows citizens to collaborate with engineers in designing digital policies. In 2023, it generated ride-sharing regulation with 87% citizen approval.
- Algorithmic Civic Education
  Finland includes in its national curriculum (from primary school) concepts such as algorithmic bias, data privacy, and source verification. In 2025, 94% of 15-year-old students could identify deepfakes (OECD PISA 2025).

43 Landemore, *op. cit.*, p.32.

- Toward Digital Constitutionalism
  Paul Nemitz proposes three pillars:
  i. Supremacy of law over code
  ii. Separation of powers in the digital era (legislative, executive, judicial + independent AI authority)
  iii. Fundamental rights applicable to algorithms
  "Democracy cannot survive if algorithmic power is not subordinated to the popular will expressed through law" (Nemitz 2018, 14)[44].

VI. CONCLUSION

The AI-democracy relationship is a tapestry of synergies and confrontations that demands immediate action. As Jungherr concludes, "focusing on specific aspects will allow us to monitor real impacts" (Jungherr 2023, 12)[45]. Preserving democracy requires not only regulation but inclusive innovation. This article invites ongoing academic dialogue.

The relationship between artificial intelligence and democracy is not an isolated technical phenomenon but a historical, political, and ethical process that redefines the social contract in the 21st century. This article has explored that relationship as a dynamic continuum between synergy and confrontation, between the promise of more inclusive and efficient deliberation and the risk of profound erosion of fundamental democratic principles. Throughout its sections, we have demonstrated that AI is not a neutral actor but a power agent that amplifies both human capabilities and their worst inclinations: inequality, manipulation, and opacity.

This conclusion does not seek to close the debate but to open a horizon of collective action, synthesizing the findings on the algorithmic democracy of the 21st century based on the following three general conclusions and findings:

44 Nemitz, Paul. 2018. "Constitutional Democracy and Technology in the Age of Artificial Intelligence". Philosophical Transactions of the Royal Society A 376 (2133): 14.

45 Jungherr, *op. cit.*, p.12.

A. STRUCTURAL DUALITY

The synergistic dimension of AI in democracy manifests at three operational levels:

- Amplified Participation: Tools such as mass deliberation systems (Chile, Europe) have processed millions of citizen contributions with precision surpassing human capacity. The Chilean case, where 2.4 million contributions were synthesized into 147 common demands, demonstrates that AI can act as a facilitator of collective intelligence (CEPAL 2023, 45).
- Governmental Efficiency: From Estonia to Singapore, AI has reduced bureaucracy, detected fraud, and anticipated crises. In Estonia, the X-Road system not only saves €42 million annually but reinforces institutional trust by making every transaction visible and auditable.
- Digital Inclusion: Real-time translation, educational chatbots, and accessible platforms have democratized access to political debate.

However, this synergy is not automatic. It requires deliberative design, citizen auditing, and algorithmic literacy. As Hélène Landemore warns, "AI will only be democratic if built with citizens, not for them" (Landemore 2023, 32). On the other hand, the confrontational dimension reveals a structural erosion of democracy:

- Massive Disinformation: Deepfakes and algorithmic propaganda have turned elections into informational battlefields. In Brazil 2022, 68% of viral messages were false (Vázquez-Barrio and Salazar 2023, 112).
- Induced Polarization: Recommendation algorithms do not merely reflect divisions but fabricate them. Cathy O'Neil sums it up bluntly: "Algorithms are opinions embedded in code".
- Algorithmic Surveillance: From China's social credit to biometric systems in democracies, AI enables unprecedented social control. Frank Pasquale speaks of a "black-box society" where real power resides in those who control data.

This duality is not symmetric. Synergy requires institutional effort; confrontation arises from market inertia and technological opacity.

B. DEMOCRACY AS PROCESS, NOT STATE

A common error is to conceive democracy as a static state that AI can "improve" or "damage". In reality, democracy is a continuous process of collective legitimation, as Robert Dahl aptly argues (Dahl 1989, 221)[9]. In this process, AI acts as a catalyst for institutional change.

Daniel Innerarity offers an interpretive key:

> "Artificial intelligence does not transform only the means of politics but the conditions of possibility for democratic deliberation. By anticipating preferences and simulating consensuses, AI does not substitute human judgment but reconfigures it" (Innerarity 2020, 89)[46].

This reconfiguration raises a philosophical question: can a machine be more democratic than citizens? The answer is not technical but normative. It depends on who designs, who audits, and who controls.

C. TOWARD A NEW DIGITAL SOCIAL CONTRACT

The democratic governance of AI cannot be reduced to technical regulation. It requires a new social contract incorporating three foundational principles:

- Radical Transparency
  Every algorithm affecting political rights must be auditable by trained citizens. The Amsterdam model –where citizen panels review housing systems– must be scaled globally.
- Structural Participation
  Consulting citizens is insufficient; they must co-design systems. The vTaiwan case demonstrates that collaboration between civic hackers and legislators generates policies with 87% approval.
- Algorithmic Equity as a Human Right
  UNESCO already recognizes it: "States must ensure that AI does not discriminate by origin, gender, class, or location". This implies inclusive data quotas, mandatory bias audits, and sanctions for digital exclusion.

46 Innerarity, *op. cit.*, p.89.

VIII. BIBLIOGRAPHY

CEPAL, Inteligencia Artificial y Gobernanza Democrática en América Latina, Santiago: CEPAL, 2023.

Dahl, Robert A. 1989. Democracy and Its Critics. New Haven: Yale University Press.

Duberry, Jérémy. 2022. "Artificial Intelligence and Democracy: Risks and Promises of AI-Mediated Citizen-Government Relations". Internet Policy Review 28 (4): 435-438.

Donahoe, Eileen, and Megan Metzger. 2019. "Artificial Intelligence and Human Rights". Journal of Democracy 30 (2): 115-126.

Dwivedi, Yogesh K., et al. 2021. "Fostering Citizen Trust with AI". International Journal of Information Management 58: 102-115.

Filgueiras, Fernando. 2022. "AI in Public Administration: Governance Challenges". Revista del CLAD Reforma y Democracia, no. 79: 20-35.

Guerrero-Solé, Frederic, and Laura Pérez-Altable, eds. 2025. La democracia en riesgo: Internet e IA al servicio de los populismos. Barcelona: UOC.

Innerarity, Daniel. 2020. "El impacto de la inteligencia artificial en la democracia". Revista de las Cortes Generales 109: 87-103.

Jungherr, Andreas. 2023. "Artificial Intelligence and Democracy: A Conceptual Framework". Social Media + Society 9 (2): 1-12.

Kreps, Sarah, and Douglas Kriner. 2023. "How AI Threatens Democracy". Journal of Democracy 34 (1): 122-131.

Landemore, Hélène. 2023. "Fostering More Inclusive Democracy with AI". Finance & Development 60 (4): 28-31.

Nemitz, Paul. 2018. "Constitutional Democracy and Technology in the Age of Artificial Intelligence". Philosophical Transactions of the Royal Society A 376 (2133): 1-15.

Nilsson, Nils J. 2010. The Quest for Artificial Intelligence. Cambridge: Cambridge University Press.

O'Neil, Cathy. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Broadway Books.

OECD. 2020. Innovative Citizen Participation and New Democratic Institutions. Paris: OECD.

Ojo, Adelaoye, et al. 2019. "Enhancing Democratic Health through AI". Government Information Quarterly 36 (3): 101-110.

Pasquale, Frank. 2015. The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press.

Risse, Mathias. 2021. "Artificial Intelligence and the Past, Present, and Future of Democracy". Carr Center Discussion Paper. Cambridge, MA: Harvard Kennedy School.

Russell, Stuart, and Peter Norvig. 2021. Artificial Intelligence: A Modern Approach. 4th ed. Upper Saddle River, NJ: Prentice Hall.

Summerfield, Christopher, et al. 2025. "The Impact of Advanced AI Systems on Democracy". Nature Human Behaviour 9: 1234-1245.

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO.

Vázquez-Barrio, Tamara, and Idoia Salazar García, eds. 2023. Inteligencia artificial, periodismo y democracia. Valencia: Tirant lo Blanch.

Wright, Nicholas D., ed. 2019. Artificial Intelligence, China, Russia, and the Global Order. Maxwell AFB: Air University Press.

Zidouemba, Marie-Thérèse. 2025. "Governance and Artificial Intelligence: The Use of AI in Democracy and Its Impacts on Participation Rights". Discover Artificial Intelligence 5 (1): 1-20.

HUMANISM IN THE AGE OF ARTIFICIAL INTELLIGENCE*

CONTENT:

I. Introduction **II.** Philosophical humanism in the 21st century **III.** A persistent problem in humanism: ontological dualism **IV.** Towards a new earthly humanism **V.** The crisis of humanism since the 20th century **VI.** Is there a Human Nature that we must protect? **VII.** The future of humanism in the age of AI **VIII.** Bibliography

> JORGE ENRIQUE LINARES SALGADO

Former Director of the Faculty of Philosophy and Literature at UNAM

* I have written this chapter in Spanish and have translated it using the AI software DeepL©.

I. INTRODUCTION

In the era of the rise of artificial intelligence (AI), we need to rethink the original meaning and destiny of humanism as a philosophical tradition that defends, to ensure our own survival, human dignity and the essential traits that constitute us as a biological species.

This is not the scope of this work to present a comprehensive critique of humanism within Western philosophical thought, but a synthetic view of this tradition that dates to the Greco-Roman classics and was revitalized by Renaissance thinkers such as Giovanni Pico della Mirandola in the 15th century (Goñi, 2020). Despite its enduring influence, Renaissance and modern humanism were circumscribed by certain limiting perspectives. Specifically, these traditions adopted anthropocentric, Eurocentric, and androcentric ideas that ultimately narrowed and distorted anthropological conceptions within the Western philosophical tradition. Such constraints resulted in a partial and skewed understanding of what it means to be human, which has had lasting implications for subsequent philosophical discourse.

For this reason, in our time, various critiques and attempts to overcome the biases and limitations of classical humanism have been put forward to formulate a supposed "post-humanist" proposal (Braidotti, 2013). However, there is a misunderstanding in this term, as it can be understood as a critique of modern Western humanism (as Braidotti argues), or as "posthumanism" in a technological sense; that is, the project of overcoming the natural and evolutionary conditions of human existence through a radical genetic, neurological, or physiological transformation of the human organism (Diéguez, 2017).

However, just as postmodernism, in its naive desire to surmount the contradictions of modernity –its rationalism and hegemonic Western narratives–

ended up being only a superficial and fleeting desideratum, no "post-humanism" is now viable for the reasons I will mention here. The main one is that it is not feasible to go beyond the natural, biological-evolutionary condition of the human species[1].

Rather, what could happen in the era of artificial intelligence dominance, and this is one of its greatest risks, is the gradual decline or *obsolescence* of human intelligence and cognitive abilities, leading to their replacement by AI systems in several areas of economic, political, cultural, and even scientific activity. Indeed, the emergence of AI as a productive force, as an all-encompassing and omnipresent technology, represents a growing risk to the development of our species' most essential capabilities: those that arose from natural, bodily intelligence, essential for social life.

Günter Anders (2025) raised, by as early as 1956, the concept of "obsolescence of human" in the face of its own technological creations. Anders' diagnosis of *human obsolescence* focuses on three theses: a) human beings are no longer able to keep up with their own technological productions; b) what we produce and what we can do *technically* exceeds our capacity for *representation* and *imagination* and, therefore, our individual responsibilities and moral conscience; and c) we are unable to recognize this new historical situation because we have lost the ability to understand the remote effects of the technological power we now wield. We find ourselves in a collective state of unconsciousness and "blindness" to the possibility of the deformation of the human condition, which increases what Anders (2025) called the *promethean gap [prometheische Gefälle].*

Anders (2025) also coined the concept of "promethean shame" [*prometheische Scham*] to define an ambiguous feeling in the mind of the individuals who inhabits the modern technological world: they feel proud of the ostentatious perfection and power of their technical creations (as if they were Dr. Frankenstein) and, at the same time, they feel "obsolete" and humiliated by their own natural imperfection. Faced with the refinement of their technical productions, human beings suffer "shame" because of their contingent and mortal condition. Unlike themselves, their artifacts have been designed, programmed, and tested by human engineering. So, technical production stands at the antipodes of natural reproduction of life, which is a "blind" and non-predetermined process. For that reason, human beings experience shame at their *obsolescence*.

1 To explain these theses, I refer to the last chapter of my book: Linares, J. E. *Adiós a la Naturaleza [Farewell to Nature]*, 2019.

The development of AI has also contributed to the worsening of this technological shame because modern individuals reject the natural limits of memory and calculation, cognitive biases, and neurological and mental illnesses –in short, the finite and vulnerable condition of our minds and brains. This has given rise to the project of turning human beings into *bioartifacts* of their own making, into technical objects designed, programmed, and tested by anthropo-technology, so that they do not suffer from mental and neurological diseases or disabilities. These claims go beyond the necessary and laudable therapeutic goals of medical neuro-technologies. If human beings could turn themselves into their own biotechnical production, and thereby overcome or transcend their natural condition, only then would they cease to feel such "shame" for their finite and contingent being. AI now promises to serve as that cognitive prosthesis to expand our memory, reasoning, and understanding.

Transhumanism and technological "posthumanism" (Hottois, 2015) have proclaimed the possibility of overcoming our *natural* limitations and imperfections, especially neurocognitive ones (Lara & Savulescu, 2022). But, in reality, one cannot go against the natural evolution of humanity or humanism as an axiological concept, since the latter has identified since ancient times the best capacities and virtues that all human beings possess; in other words, one cannot voluntarily surpass either humanity as a natural species or classical humanism as a vindication and revaluation of the traits inherent in human nature. Thus, "posthumanism" cannot transcend the conditions and characteristics that define our species or invent new virtues and powers for humanity, because philosophical humanism has already defined all its virtues –as we shall see– and they have been developed throughout history.

However, as Juliana González (1996) points out, human potentialities are ambiguous, tending toward good, altruism, and peaceful cooperation, or toward evil, selfishness, organized violence, and destruction. Without being able to provide an exhaustive analysis here, Pico della Mirandola was right, in his famous *Discourse on the Dignity of Man* (Pico della Mirandola, 2020), to propose the idea of the human being as one who is capable of self-configuration and self-production, given that their being is not complete in the evolutionary and historical process, but is open and in continuous transformation. Nevertheless, this self-transformation has been primarily symbolic, mental, and political, rather than biological, neural, or genetic. This idea of the Renaissance thinker emphasizes the ambiguous nature of humanity: in its historical development, it can improve or worsen, it can increase its capacities or lose them, diminish them or forget them; in short, human beings can humanize or *dehumanize* themselves, thanks to their own decisions, ideals, and actions. Therefore, we must reflect

on the extent to which the emergence of AI contributes to this process, in one sense or the other.

The essential ambiguity of human beings implies that "evil par excellence […] is certainly *hybris:* violence, excess, arrogance, uncontrolled desire […] and the lack of 'measure' and 'limits'" (González, 1996, p. 17). The growing dehumanization of the existence of millions of people (Anders, 2025) is due to the influence of technological and political power exerted over their will, their conscience, their capacity for deliberation, moral self-awareness, and responsible action. During the early Modernity, authors such as Francis Bacon, in his utopia *The New Atlantis* (Bacon, 2014), postulated that technological power would allow humanity to dominate nature and control its own body to improve living conditions and combat disease and death. He predicted an era of full expansion of human existence and extended life expectancy. Today, that technological power is out of control, not regulated by society, and has turned against human beings (Han, 2022), weakening their moral and cognitive capacities, as well as affecting their environmental conditions for survival.

II. PHILOSOPHICAL HUMANISM IN THE 21ST CENTURY

To rethink the meaning of philosophical humanism in the age of AI, we must critique the three limiting biases of the modern tradition that we have mentioned: anthropocentrism, Eurocentrism, and androcentrism. The misconception and unviability of current posthumanism lies in confusingly formulating a supposed overcoming of this humanist tradition instead of rethinking what authentic humanism should consist of in the 21st century, so that it must defend the inherent value of all humanity as a *natural species* and as a cultural tradition diversified throughout its history. This means that humanism must be reborn as a vindication of the natural conditions of existence of all human beings, but it must demand the equitable, egalitarian, and balanced development of all humanity.

Therefore, if classical humanism has meaning in our time, in which a new "reason of force majeure"[2] (Nicol, 1980) deployed in AI, which opposes and

2 This concept of reason of force majeure (RFM) points out the emergence of a new kind of technological reason that imposes the biological necessity over free activities and creativity of human beings. Thus, the RFM is universal, systematic, homogenous, non-dialogic neither democratic, and could be violent because it imposes the necessities of "biological" survival over the freedom and societal deliberation of common good. Nicol affirms that this new reason tends to dominate and eliminate the traditional human reason developed in the sciences and philosophy (Nicol, 1980).

strives to replace human reason (scientific or philosophical), it must now be a reflection well-grounded in the historical and evolutionary reality of our species, thus reaffirming the essential characteristics that define it. Today's humanism must be, as Juliana González (1996) argued, an *earthly humanism:* "*Homo is humus: earth*. This is one of the most distinctive features of humanism and one of the most decisive for ethics. Humanism recognizes the *terrestrial* essence and destiny of man. […] There is 'ascent', but this is always *from* the earth and *towards* the earth itself" (González, 1996, p. 24). In other words, humanism must be evolutionary, not anthropocentric or Eurocentric, diverse and pluralistic, and not androcentric either. González defines humanism in the classical sense in the Western tradition as follows:

> Humanism is not theory, it is "reflection"; but it is also vigilance, "watching over", "caring" (*Sorge* or "care" in Heidegger's terms); it is "seeking" the "humanity" of "human" man: what the Greeks called *anthropine physei:* human nature, and the Latins *Humanitas,* proper to *homo humanus.* This means, among other things, that "humanism" encompasses much more than what was historically and culturally the humanism of the Renaissance (González, 1996, p. 15)[3].

Renaissance humanism undertook the recovery of the idea of a common *natural* human condition in all members of our species. That is why humanism is *naturalistic*. It postulated an ideal of harmonious integration with nature, of reconciliation with the human body and the earth, a disposition toward new forms of religiosity that did not submit to the dogmatism of the Catholic Church, the recognition of individual autonomy, the spirit of adventure and exploration of the world and the universe, the cultivation of beauty and the enjoyment of the arts, the idea of peaceful tolerance among peoples, mainly. This ideal was largely separate from the Christian religious conception that had dominated Western thought since the Middle Ages, which denied biological nature and despised the body and its sensible passions, affected by "original sin", to affirm only spiritual or rational faculties, and the subordination of human beings to the divine through renunciation of the flesh and the world, and the strict exercise of the theological virtues (faith, hope, and charity). Thus, *Humanitas* became in the Renaissance the *essence* and, at the same time, *the excellence* of humanity, the ideal of the natural realization of what is truly human, without aspiring to immortality or the perfection of the human body and mind.

3 I have translated the works written in Spanish employing the AI software DeepL©

Humanity (*Humanitas)* also means ambiguity and dynamic virtues, being open and temporal, contradictory reality, and moral ideal: the aspiration to preserve the inherent, albeit ambiguous, attributes that define humanity and differentiate it from other animal species.

III. A PERSISTENT PROBLEM IN HUMANISM: ONTOLOGICAL DUALISM

The central problem we have inherited from the Western humanist tradition has been a dualistic and *unnatural* conception of what we consider to be properly human. Since the Greek tradition (mainly in Plato and Aristotle), philosophers have thought that the essence of humanity resides only in reason, the soul, or the spirit, while the body and biological conditions were considered accidental or separable from the soul. And that, because of their spiritual condition, humans would be superior and different from other animals and beings existing in nature.

As a result, classical humanism inherited this *narcissistic* conception of self-praise and overestimation of our cognitive, creative, and expressive abilities. Until the 20th century, the idea that human beings are the only *ones* capable of possessing *logos* –that is, speech and reason, thought and expressiveness– predominated (both in the sciences and in the arts). Today, thanks to the biological sciences and with the help of AI, we are beginning to understand more and better the complex languages of other animals, enabling us to delve into their specific forms of cognition and representation of the world. In the future, philosophical humanism will have to start from *the fact* that the human species is not the only one capable of rationality or articulate language, but this will mean a gain rather than a loss: humanity will recognize itself as part of a planetary community and will better understand its responsibilities in preserving all forms of life that have existed on Earth, in front of major ecological crisis and risk of collapse of the entire civilization (Riechmann, 2004). Today we can say that all the *spiritual* features (rational, emotional, expressive, conscious) of humanity are emergent properties of our neurobiological constitution in balance with the natural environment.

Contemporary humanism implies an ontology and axiology of the human that, after the crisis of metaphysics and the dualistic vision, must be thought of as temporary and constitutively contradictory to the human: its body/brain/mind unity, its dual bio-cultural character, its biological and neuronal constitution, as well as its mental and rational composition. Humanism thus reveals itself

as a valuation of the human in its ethical character as an essential and indissoluble feature.

> Humanism is above all an ethical conception. It implies, in effect, a profound knowledge of the human being, on which the world of values in general can be based and from which, in particular, ethical values and ideals are derived that are distinguished precisely because they are rooted in human *freedom* and *dignity*; rooted, in short, in the *ethos,* in man's most faithful fulfillment of his "free destiny", fully humanized (González, 1996, p. 17).

IV. TOWARDS A NEW EARTHLY HUMANISM

According to Juliana González (1996, p. 29), these are the primary features of humanism that remain at the center of every consistent humanist definition and position in our time:

1. The centrality of the human (in its natural evolutionary condition) and recognition of its dynamism or historical mobility.
2. *Earthliness* and contingency of the human: expressed in birth and death.
3. Specificity and irreducibility of the human: subjectivity, personality, *human dignity* as an irreplaceable and inherent value.
4. Moral ambivalence between the possibility of individual and collective ascent and descent; humanization and dehumanization or inhumanity as chosen evil.
5. Ontological/biological equality and community, as well as societal diversity and individual uniqueness.
6. Recognition and tolerance of human diversity.
7. Unity of the human condition with animal life.
8. Unity and community of humanity-nature (of reason, corporeality, emotions and sensitivity, birth-death, the integrity of human existence in its inherent dimensions: individual-society-species).

As González (1996) argues, all humanism has a scope (or at least a claim) to be universal by attempting to conceive of human nature in its radicality, thus expressing its evolutionary characteristics that continue to act in our species. Humanism cannot be understood as a concept inherited solely from Greco-Roman genius (although this tradition is always referred to in the Western world), but rather as a self-reflection through which human beings become aware of their *uniqueness* and *dignity* as living beings and as *historical beings*:

> *Universal humanism* is […] beyond the Greeks themselves and the Western tradition. It is a knowledge and experience that is not limited to a historical moment, to a few thinkers, to a people, or to a particular culture. It encompasses all those times and places in which, with all their historical and cultural variations, the authenticity of the humanized human being prevails, which is evident in their art, religious thought, morals, ways of life, and most cherished ideals (González, 1996, p. 16).

Thus, the most significant aspect of humanist thought, in any of its forms or expressions (literary, poetic, sculptural and plastic, philosophical, or scientific), is that it reveals an *evolutionary ontology* of the human, a form of self-awareness of humanity, as well as an ideal of life, which has always been diverse and variable in geographical space and historical-cultural development (Nicol, 1977).

In this way, humanism has revealed to us that the *idea of man [idea del hombre]*, as conceived by Eduardo Nicol (1977), is never static or homogeneous, since the entire human being changes with each self-conception that is projected as an affirmation of collective identity. Each "idea of man" is a way in which a civilization affirms itself and projects its own qualities; it is, therefore, a symbolic idea located in space-time and, simultaneously, has a universal claim. Currently, such an "idea of man" must reaffirm our natural cognitive abilities, as well as our social faculties based on cognition and language, but as characteristics that –for the first time– are at risk in the face of the global advance of AI in all societal spheres. The unprecedented fact is that today we can doubt whether these inherent faculties of the human condition could weaken so much that they become blurred and lost in many individuals of the human species. We are therefore facing a crisis of those ontological traits that define humanism. They are threatened by the portentous and, so far, unlimited advance of AI.

V. THE CRISIS OF HUMANISM SINCE THE 20TH CENTURY

However, humanism (as a philosophical concept) had entered a profound crisis in the Modern Age, due to its own contradictions (principally its metaphysical dualism) and due to a radical transformation (material, technological, and economic) of the historical world in which it arose. We can summarize in the following features the *crisis of humanism,* as anthropocentric exaltation, which denote a rupture in the image that Western civilization had forged of itself:

1. Isolation, solipsism, and fragmentation of existence (expressed, for example, in a variety of mental disorders, like depression and anxiety, that increasingly affect more people).
2. Alienation from work, creativity, and sociocultural consciousness (as discovered by Karl Marx).
3. Exhaustion of the colonizing model of modern Western civilization, which advocated a normative and ideological conception of European culture as a model, and its distinctive features as superior values: whiteness, individualism, Christianity, patriarchy, scientism, and technological transformation as factors of "progress" and modernity (Echeverría, 1995).
4. Crisis of the ideal of technological-industrial progress because the increase in technological power has generated a socio-environmental crisis, even though it has produced sufficient goods to improve the lives of human beings; but, in reality, only a minority has been able to access its benefits.

Broadly speaking, four critical theories demolished the universalist claims of modern anthropocentric and rationalist humanism: Copernicus' heliocentric theory, Darwin's theory of evolution, Freud's theory of the unconscious, and the critiques of classical humanism by existentialist perspectives such as Sartre, Heidegger, and Sloterdijk. Also, Descartes' dualistic rationalism entered in crisis, but neither philosophy nor science has overcome it. We continue to bear the weight of body/mind dualism and remain confined within the restrictions of reductionism, which is divided between the mechanism with which we conceive the body and the spiritualism with which we imagine the soul. Although the interaction between body and mind remains unresolved (Díaz, 2021), current perspectives suggest that the notion of autonomous, creative human subjectivity capable of directing its own destiny is diminishing. Rationality is in crisis, as are the most significant emotions such as empathy and benevolence, the sense of solidarity and responsibility (Jonas, 1995). The individual is reified and becomes a mere cogwheel within a global technological machine. But the crisis has been so profound that it has gone as far as to deny the specificity and unique value (*dignity)* of human beings:

> The seriousness of the crisis is evident in the fact that today the *very specificity of the human indivuality* is being questioned. Today, there are doubts about whether humans have their own irreducible and distinctive quality, their *humanitas,* their very *ethos*. There are doubts about the existence of a human "nature" or "essence" and the ethical condition of humans. In short, there are doubts about freedom and dignity (González, 1996, p. 32).

Currently, the *centrality of humans* is being questioned by theories of *technological singularity,* such as that put forward by Ray Kurzweil (2024), who posits the advent of AI machines that develop self-awareness and act with complete independence from humans; this is the point of no return in history when AI will far surpass most humans. Let us look at Kurzweil's predictions:

> Although many technological challenges remain before we can achieve the Singularity, its key precursors are rapidly moving from the realm of theoretical science to active research and development. During the coming decade, people will interact with AI that can seem convincingly human, and simple brain–computer interfaces will impact daily life much like smartphones do today. A digital revolution in biotech will cure diseases and meaningfully extend people's healthy lives. At the same time, though, many workers will feel the sting of economic disruption, and all of us will face risks from accidental or deliberate misuse of these new capabilities. During the 2030s, self-improving AI and maturing nanotechnology will unite humans and our machine creations as never before–heightening both the promise and the peril even further. If we can meet the scientific, ethical, social, and political challenges posed by these advances, by 2045 we will transform life on earth profoundly for the better. Yet if we fail, our very survival is in question. […] our final approach to the Singularity –the opportunities and dangers we must confront together over the last generation of the world as we knew it (Kurzweil, 2024, pp. 4-5).

Thus, in the age of AI, the idea of the *obsolescence of humanity in its natural state* is beginning to gain momentum. But, as we have explained, from a humanistic point of view, it is not possible to consider any technological *surpassing* of our natural species, much less the abdication of cognitive potentialities to reinforce the technological power of AI.

Despite his optimism, Kurzweil points out these general risks of the accelerated and unregulated development of AI:

> But superintelligent AI entails a fundamentally different kind of peril –in fact, the primary peril. If AI is smarter than its human creators, it could potentially find a way around any precautionary measures that have been put in place. There is no general strategy that can definitively overcome that. […] There are three broad categories of peril from superintelligent AI, and with focused research on each, we can at least mitigate the risk. Misuse encompasses cases where the AI functions as its human operators intend, but those operators deploy it to deliberately cause harm to others (Kurzweil, 2024, p. 278).

Therefore, the most significant aspect of this new idea of *human obsolescence* seems to be a weakening, if not exhaustion, of the natural cognitive abilities to act in the world and to respond to different natural and social problems and needs. Technologists and AI propagandists (such as Kurzweil) argue that the fusion between the human brain and AI will generate an expansion of reasoning power and memory, an extension of consciousness in the *cloud,* a dual subjective identity through avatars or *technopersons* (Echeverría & Almendros, 2020), and an era of full splendor that could even transcend biological death. But the truth is that every advance in the agency capabilities of AI represents a potential cognitive loss for countless flesh-and-blood people.

Certainly, human beings and their self-image are historical; evolution has not stopped, and our species continues to adapt to different environmental changes. Being and becoming are undoubtedly intertwined within our identity, but the biological and neurological modifications, alterations, and degradations of humanity influence its self-image and are sometimes irreversible. For example, violence, war destruction, and evil alter human beings and consciousness; they not only cause material damage to millions of people but also degrade humanity's moral image and result in mental damage such as post-traumatic stress or loss of memory and personal identity. This implies that humanity has the power to *make* and *unmake* its being in history. So internal bio- or neurotechnological transformations, as well as possible fusion with AI devices, could negatively alter human nature, but it could also happen externally through the rise and expansion of AI's agency, gradually supplanting workers in complex cognitive tasks.

Then, in our era of AI dominance, the ethical alternative is to reaffirm humanism by rethinking it with the essential features it has acquired in history. Thus, humanism today means reaffirming reflective and deliberate capacity, creativity, free will, as well as theoretical and contemplative, playful, artistic, and poetic activity. For now, AI is still unskilled in these complex and subtle activities. Humanism also means understanding the complex nature of our species: its biocultural, adaptive, and ambivalent character.

VI. IS THERE A *HUMAN NATURE* THAT WE MUST PROTECT?

A fundamental problem posed by the revival of humanism in the age of AI is whether this new, all-encompassing technology will irreversibly affect the consciousness, reasoning, linguistic expression, or *ethical sense* of most human beings. Along with the radical transformations of the body, brain, and mind that are new technological goals in our century (Lara & Savulescu, 2022), AI could alter or undermine the self-awareness and moral responsiveness of many people to the societal and environmental problems that most afflict us. Humanism had, until now, emphasized the vulnerable but irreplaceable structure of natural human corporeality. Of course, the priority would be to use technological means and resources to ensure an adequate quality of life for all people, rather than undertaking risky transformations of the brain, body, and mind. But even if radical transformations of the human body were not achieved or attempted, it would be enough for the cognitive agency of AI to expand in such a way that it would end up diminishing and damaging the cognitive, creative, and deliberative capacities of millions of people; for example, their ability to read and write (which is not a natural trait, but a cultural and learned one, based on the natural ability to acquire and speak one or more languages[4]). Today, AI translation systems replicate the skills of human translators and support, for instance, academic research across numerous languages, helping scholars from various cultures and countries communicate more effectively. However, relying on AI for communication and comprehension may lead to a decline in individual linguistic abilities.

4 For example, we know that the capacity for articulate language is innate, according to Chomsky, and universal; it unfolds in a multiplicity of idioms with their natural generative grammars. But in the history of humanity, writing systems (predominantly ideographic and syllabic) emerged as a mental innovation that allowed linguistic messages to be preserved and transferred. This technical innovation, which dates back more than 3,000 years, enriched our cognitive and symbolic capacities and enhanced our social skills of communication and organization. Today, AI can generate, through its mastery of writing and grammar, linguistic messages in text and speech in a wide variety of languages, emulating the cognitive abilities of majority of human beings. As we know, given the conventional nature of writing systems, each person must learn to read and write in one of these systems. In contrast, the AI of chatbots emerges with this ability and powers it to levels that few people can achieve; for example, "speaking" and translating between so many different languages.

Let us therefore briefly analyze the evolutionary features of human existence that humanism should preserve in the face of the rapid advance of AI's agency power (Linares, 2019):

1. *Birth-mortality*: as Hannah Arendt (2018) pointed out, awareness of finitude and contingency largely determines the symbolic image of the human (Jonas, 1995) and the universal recognition of our common condition: we are all biologically born from the union of gametes of individuals of the human species, and we will all inevitably die. Transhumanist expectations of indefinitely extending life, even of transcending biological death in a digital form, are contrary to any humanist conception.
2. A *common natural corporeality*: our phenotypic and genotypic endowment is not determined or decided by anyone, at least not yet. Of course, the subjective experience of the body is culturally constructed, but it is a biological constant in history that sets the limits of what can be experienced and establishes common vital needs, as well as the capacities and potentialities that *humanize* the existence of any person. Human nature is flexible and dynamic; it does not need to be biotechnologically transformed to improve it, much less to achieve a superior quality of life or mental well-being.
3. *Mutability* and *vulnerability*: the human body is vulnerable and exposed to genetic mutations and metabolic, neurological, and physiological disorders. But so far, the most significant mutations that humanity has undergone have been more "symbolic" or cultural (that is, mental); humanity has shaped its own being because of a process of consciousness modification rather than technical-material intervention in its biological structure. Nevertheless, *vulnerability* and *mutability* are organic because the body exists in a fragile balance with the environment, which has endured serious and irreversible damage in our time[5]; different environmental agents –natural or human-made– could alter its genetic composition and balance with the environment. Humanism must propose that improving human living conditions necessarily involves remedying the environmental crisis we have caused through excessive pollution and exploitation of natural resources, as well as the acceleration of global climate change, which is now unstoppable.

5 All of us who live in large cities (metropolis) have already been affected by the high toxicity to which we are exposed, as well as by water, soil, and air pollution; for example, pollution from lead and other metals that damage our nervous system. Our biological nature is seriously affected by the technological world, and this explains our physical and mental illnesses and ailments, such as the epidemic of several types of cancer that are appearing in increasingly younger people.

4. *ndividuality*: each person *is* a unit with their own body, feels, suffers, and loves in their "own flesh", which is why bodily intelligence is, until now, irreplaceable, and non-transferable to another material, biological, or technological medium. Although the social value of individuality is relative because of its cultural codification, it remains the basis for recognizing an ontological community among human beings: we are all individuals with finite and vulnerable bodies and brains. It can be very risky to attempt to merge the brain/mind with AI or to duplicate consciousness and memory by transferring them to computer and AI devices. Humanism must advocate for the preservation of the natural conditions of neurobiological experience, consciousness, and unique and unrepeatable individual bodily identity (Habermas, 2002).
5. *Social interdependence*: through sensitive corporeality, we are open and connected to the world; each human body/brain is able of interacting and empathize with other human beings, other animals, and the entire natural world. It is essential to understand that human experience and intelligence are social and historical processes and products. AI cannot replace these human attributes yet.
6. *Bodily sensitivity*: interdependence is based on bodily sensitivity, through which human beings are able of developing varying degrees of sensations and emotions, to have feelings of sociability and concern for the lives of other living beings. Human life loses its most fundamental properties without the deployment of sensory and moral capacities. AI could help to repair and prosthetically replace the sensory and motor capacities of people who have suffered bodily damage due to accidents or illness, but we should not use, precautionary, to "augment" these capacities in healthy people. The integrity of human genome must be protected by international laws (González, 2005).
7. *Cognitive faculties*: perception, linguistic expression, imagination, and thought. Human beings are capable of being aware of their interdependence in the world through their mental representations and linguistic expressions; they are capable of imagining the situation of others and anticipating events to plan and guide their actions. Consciousness, which emerges from neuronal activity and the nervous system, remains a mystery that we cannot replicate in AI systems. As a precaution, we must preserve the natural and societal conditions in which human intelligence arises and develops.
8. *Ethical sense*: human beings are able of valuing things to guide their lives and are also capable of evaluating and choosing between options to achieve a *good* life. Values are historically and culturally conditioned, but the *need* to

value and choose between the possibilities that unfold in existence is a constant in humanity. The ability to evaluate is based on a free nature, that is, one that is possible and not predetermined. But the fact that human life is *possibility* implies that it is open to diverse options that are ambivalent, can improve or decline, can be humanized or dehumanized, can be confined to selfish illusions (such as transhumanist ones) or can expand toward a cooperative morality and justice. So far, even though AI devices can simulate moral deliberation processes and make decisions, only humans are capable of consciously and voluntarily assuming responsibility for their decisions and actions.

The various AI systems, and not only those of dialogical and communicative AI (which use artificial language), would not necessarily have a massive impact on the conditions we have just listed, but they could jeopardize the system of social interactions and the development of individual cognitive abilities, which together form part of our *biocultural nature*.

Consequently, the ethical meaning of human existence is based on an awareness of finitude, contingency and indeterminacy, freedom and vulnerability, and the essential interdependence of our being, which emanate from a natural corporeality that we have received from evolution and that we have shaped symbolically rather than technologically, based on the limits of a given bodily configuration, with all its restrictions and potentialities. Transhumanist projects assume that rational characteristics and the capacity for human self-transformation (which had only been symbolic but not technical) are an immutable constant; that human rationality is something like a subtle property that would not be affected by any biotechnological intervention. In this idea, if expressed correctly, there appears a remnant of spiritualist ontological dualism: it seems that transhumanism trusts that we can modify ("improve") *the* biological *body* and *brain* at will without altering the other non-material substance: the human mind or spirit. But humanity does not belong to two distinct substances, it only has one duality: body-mind, bodily experience or self-consciousness.

For its part, excessive confidence that AI will remain merely a subordinate auxiliary to humans will be a disastrous illusion. Our intelligence is complex and contextual and has developed in tandem with changes in the physical and environmental settings we inhabit, as well as the societal structure on which we depend. The rise of AI points to greater virtualization of lived space and time, AI-mediated communication (physically isolating and separating people), and weaker social interaction in favor of AI control and automation systems in all types of tasks and processes.

If in the past, the first industrial revolution generated the promise of liberation from physical human labor (Marx himself believed that the socialist revolution was possible through the mechanization of all production), nowadays the accelerated development of AI has raised the promise of liberation from routine or mechanized intellectual work that would free us from many hours to devote to leisure and pleasure. However, what is more likely to happen is the loss of cognitive abilities essential for social life as we become more dependent on AI for all kinds of tasks, however simple they may seem, such as locating ourselves geographically and traveling from one place to another in a city.

Thanks to mechanization and industrialization, many of us no longer must do hard physical work, but we are also no longer know how to produce our own food, how to repair or maintain our machines, or even how to fix a bicycle. Similarly, in the near future, we may not be able (due to our subordination to AI devices) to perform simple mathematical calculations, remember practical information, read in depth, write properly and with our own style, communicate our most intimate thoughts and feelings, or engage in intelligent conversation with our peers. The new industrial revolution of AI may bring enormous productive advantages, but it may also lead to serious cognitive losses for most people. These essential attributes of humanity would be put at risk, causing the greatest crisis of *Humanitas*. This cognitive destruction of the most essential qualities of human beings, even if silent and gradual, would be irreversible: humanity would lose its historical memory and its sociopolitical ability to organize, communicate, and project itself into the future.

Therefore, we must carefully analyze the social and individual effects, as well as the subtle and progressive changes in people's brain and mental configurations, as a result of the global development and expansion of AI in all its forms. We must assess these technological transformations, perhaps the most radical that our species will undergo (as Kurzweil himself argues), considering their effects on people's self-awareness, brain-mind *integrity*, sensitivity, and moral consciousness, as well as on the cognitive abilities that make interaction and rapport with others and with the entire world possible.

In principle, the employment of AI could be useful for remedying disabilities or recovering organic functions damaged by severe neurological diseases or accidents or even preventing them. For example, AI could be useful as a prosthesis for verbal communication in people who suffer aphasia, as well as for people who have lost their memory. For people with motor, visual, or hearing disabilities, robotic AI prostheses connected directly to their brains could also be employed for therapeutic purposes.

However, morally complicated cases will arise –in contrast– in those AI interventions and prostheses in which the enhancement of certain bodily or neural properties (mobility, strength, living memory, calculation and reasoning, sensitivity, environmental and contextual information processing) could alter equality among human beings, the self-understanding of the subject and their social interactions, in such a way that the individual could not recognize themselves as a free and autonomous moral agent on equal terms (Habermas, 2002). This is the paradox of possessing cognitive "superpowers": we undermine equal rights and the freedom to act, to make mistakes and to learn, or to try different alternatives. The human condition intervened by AI would inevitably become *robotized*, rigid, and determined, just as the condition of the industrial worker became *mechanized,* restricted to the regular and necessary rhythms of productive work and its purposes, to the detriment of creativity, flexibility, and the very freedom of individuals[6]. People who are dependent on or irreversibly live with or connected to AI systems would lose their natural faculties of cognition, reasoning, linguistic expression, memory, emotional reaction, and perhaps even independent thought. These are not minor risks that governments and civil societies should carefully anticipate and examine. Therefore, the era of AI seems to be an inherent risk to humanism as we have defined it.

How *can we remain human* and cultivate *humanitas* in the age of AI? In these perspectives of risks from technological transformations, the development of AI could jeopardize the cognitive abilities of most people, due to their weakening and replacement by AI systems. Therefore, we will have to redefine the goals and principles of *humanity* in our era.

To this end, we will briefly analyze the *Amsterdam Declaration* of the International Humanist Association, which the 50th World Humanist Congress promulgated in 2002 and updated in 2022. According to the IHEU website[7], the humanist principles are:

- Humanism is ethical
- Humanism is rational
- Humanism supports democracy and human rights

6 Just remember Chaplin's brilliant film *Modern Times* (1936) that criticized the meccanization of depauperating workers.

7 International Humanist and Ethical Union or Humanists International. https://humanists.international (accessed October 2025). The Amsterdam Declaration: https://humanists.international/policy/amsterdam-declaration-2002/

- Humanism insists that individual freedom must be combined with social responsibility
- Humanism is a response to the existing demand for dogmatic religions.
- Humanism values artistic creativity and imagination and recognizes the transformative power of art.
- Humanism is a philosophy of life that promotes greater fulfillment through the cultivation of ethics and creative living (HI, 2002).

Accordingly with the organization Humanists International states, contemporary humanism rejects on principle any autocratic and dogmatic politics that does not respect the balance of political powers and does not defend individual freedoms and human rights, and that is not based on a secular conception. But it is not only authoritarian and technocratic states (such as China and Russia) or theocratic states (such as Iran) that threaten *Humanitas*, both the ontological essence and moral ideal of humanity, but also the technological power of AI (held by a few global companies, mainly the GAMAX complex[8]), which can become an irrational global power because it is undemocratic, opaque (not transparent in terms of accountability), racist or chauvinistic, belligerent, and threatening to human rights.

If a few global companies (almost all of them American) continue to monopolize the development of this disruptive technology, AI systems could become an *anti-humanist* force, not only in politics but in all areas of social life.

Based on the Amsterdam Declaration (HI, 2002), we can summarize the characteristics that humanism must defend in the 21st century:

- Humanism must affirm the irreplaceable value, dignity, and autonomy of individuals, as well as the right of every person to the greatest possible freedom compatible with the rights of others.
- Humanism advocates the responsible, equitable, and sustainable use of science and technology; it promotes free scientific research into issues of human and ecological well-being, including the risks and dangers of AI.
- Humanism defends democracy and human rights; it aims for the equitable development of every human being and maintains info-democracy (Han, 2022) and human development (including technological and scientific development) are fundamental rights.

8 Google, Amazon, Meta, Apple, and X, along with OpenAI, companies that hold a monopoly on AI.

- Humanism strives to strengthen social responsibility alongside individual freedom. Humanism is not dogmatic, does not impose any creed on its followers, and is free from indoctrination. Likewise, humanism must question any ideology or proposal that advocates radically transforming human *nature*.
- Humanism is a way of life that is not dogmatic but pluralistic, not religious but secular, aiming at the highest possible realization of human virtues through the cultivation of an ethical and creative life. Implicit in the ideas of humanism is the notion that every person can develop all the virtues and natural and common potentialities of our species, if they have adequate living conditions, goods, and services, such as access to education, scientific knowledge, and healthcare.

In our time, it is necessary to add these two principles to the theses of 21st-century humanism:

- Humanism defends the free will, autonomy, and cognitive abilities inherent to human beings and prevents their weakening or replacement by AI.
- Humanism protects and preserves the irreplaceable value of natural human intelligence and the intelligence of other animal species. It reaffirms that only human beings are competent to deliberate and make crucial decisions about their individual and collective lives, and that they should not delegate these decisions to AI.

VII. THE FUTURE OF HUMANISM IN THE AGE OF AI

In the era of the emergence of AI as a global technological power that emulates and can replace human cognitive faculties, humanism must reemerge as a moral response to protect and preserve the natural human condition, and to protect our neurocognitive characteristics in a new modality of "neuro-rights" (Yuste, 2024).

Only in this century have we been able to begin to scientifically investigate and understand how the mind and brain work (Yuste, 2024), as well as the languages and cognition of other animals; but a new form of *non-natural*, non-animal intelligence has emerged, a new form of expressive otherness that simulates another form of subjectivity and thought: AI. This new form of technological agency can supplement and replace humans in increasingly complex and crucial cognitive tasks.

For that reason, in the age of artificial intelligence, we must delve deeper into the study of the neurobiological and cognitive traits inherent to human nature, because they are what separate us from AI. We must better understand our cognitive abilities and potential to protect and defend them from the onslaught of new neuro-technologies and AI systems. Thus, neuroscientists who are aware of the new risks, such as Rafael Yuste[9], have already proposed a legal system of "neuro-rights" that would provide international protection for all people who could be affected by these neurotechnological and AI devices. These neuro-rights highlight personal identity, free will, mental privacy, equity and social justice, and the right to protection against discriminatory biases arising from the use of neuro-technologies (for example, those that could reproduce or externalize people's internal thoughts). We must protect and preserve what makes us human in the face of the prodigious advance of AI. This does not mean that all our cognitive abilities are positive and praiseworthy, for the ambiguity of human nature implies that intelligence can also degrade toward negative and destructive acts. However, humanism implies a permanent and alert awareness of this essential ambivalence:

> Humanism also entails the wisdom of *human limits*. […] It implies knowing about the greatness of man (and his horrors), as well as his contingency, his weakness, and his smallness. It is, at the same time, knowing about his power and his impotence […]

Humanism flourishes where there is no despair about human being, knowing its precariousness. It flourishes as a tragic, properly ethical knowledge: when the glory of man does not make us forget pity for man, for his descent, his inertia, and his alienation. And conversely: when human misery does not make us forget the depth and greatness of man and faith in human excellence and the struggle for humanization remains alive (González, 1996, p. 40).

Humanism advocates the *humanization* of human beings through education, access to knowledge, the arts, sciences, and universal culture. Currently, contemplative, theoretical, playful, or artistic ways of life emphasize serenity, slowness (Han, 2015), or pleasant leisure seem to be antidotes to the acceleration, rigidity, and stressful excesses of technological productivism now embodied by AI. But humanism also affirms that we must reclaim science and technology to

9 Rafael Yuste is a Spanish neuroscientist, director of the Center for Neurotechnology at Columbia University, and president of the Neuro Rights Foundation. https://www.neurorightsfoundation.org/ (accessed October 2025).

humanize them; that is, we must reclaim physical labor, technical activity, and genuine scientific research as a daily quest for knowledge and the generation of causal explanations of reality for resolution of practical life.

That is why a truly humanistic person cultivates the arts, sciences, and techniques, as Leonardo da Vinci did. Thus, scientific and philosophical thought remain the primary bastion of humanism, the enclosure in which human beings *become humanized* through knowledge and the experience of understanding the world. We must not allow AI to dominate these sanctuaries of thought. However, we must also *humanize* science and technology to restore their ethical conscience, their responsibility to protect the limits of nature and to protect the essence of humanity. Science without *conscience* and technology *without limits or moral regulations* are the perfect recipe for disaster and the destruction of *Humanitas*.

> Genuine humanism […] *does not imply the "deification" of man,* nor a break with what is not human. One thing is dignity and the main place that can be assigned to man in the cosmos, and another is […] *hybris*: human arrogance, the breaking of balance; the loss of all extra-human horizons, of all limits or proportion (González, 1996, p. 39).

VIII. BIBLIOGRAPHY

Anders, G. (2025). *The Obsolescence of Human.* Vol. I. (Trad. by Christopher J. Müller). University of Minnesota Press.

Arendt, H. (2018). *The Human Condition.* Second ed. University of Chicago Press.

Braidotti, R. (2013). *The Posthuman.* Cambridge: Polity Press.

Díaz Gómez, J. L. (2021). *El enredo mente~cuerpo [The Mind-Body Entanglement].* Mexico: Herder.

Diéguez, A. (2017). *Transhumanismo: la búsqueda tecnológica del mejoramiento humano [Transhumanism. The technological quest for human enhancement].* Barcelona: Herder.

Echeverría, B. (1995). *Las ilusiones de la modernidad [The Illusions of Modernity].* Mexico: UNAM / El equilibrista.

Echeverría, J., & S. Almendros, D. (2020). *Tecnopersonas: cómo las tecnologías nos transforman [Technopersons. How technologies transform us].* Gijón: Trea.

González, J. (1996). *El* ethos, *destino del hombre [Ethos, destiny of man].* Mexico: FCE/UNAM.

González, J. (2005). *Genoma humano y dignidad humana [Human Genome and Human Dignity].* Barcelona: Anthropos.

Goñi, C. (2020). *Pico della Mirandola. El filósofo de la armonía.* Barcelona: Arpa.

Habermas, J. (2002). *El futuro de la naturaleza humana [The Future of Human Nature].* Barcelona: Herder.

Han, B.-C. (2015). *El aroma del tiempo. Ensayo filosófico sobre el arte de demorarse [The scent of time. A philosophical essay on the art of lingering].* Barcelona: Herder.

Han, B.-C. (2021). *No-cosas. Quiebras del mundo de hoy [Non-things. Failures of today's world].* Madrid: Taurus.

Han, B.-C. (2022). *Infocracia. La digitalización y la crisis de la democracia [Infocracy. Digitalization and the crisis of democracy].* Madrid: Taurus.

Hottois, G. (ed.). (2015). *Encyclopédie du trans/posthumanisme.* Paris: Vrin.

Humanists International. (2002). *Amsterdam Declaration.* Retrieved October 2025, from Humanists International, World Humanist Congress, Noordwijkerhout, Netherlands, 2002: https://humanists.international/what-is-humanism/the-amsterdam-declaration

Jonas, H. (1995). *El principio de responsabilidad: ensayo de una ética para la civilización tecnológica.[The Imperative of Responsibility: An Essay on an Ethics for Technological Civilization].* Barcelona: Herder.

Kurzweil, R. (2024). *The Singularity is nearer. When we merge with AI.* New York: Viking.

Lara, F., & Savulescu, J. (2022). *Más (que) humanos. Biotecnología, inteligencia artificial y ética de la mejora [More (than) Human. Biotechnology, Artificial Intelligence, and the Ethics of Enhancement].* Madrid: Tecnos.

Linares, J. E. (2019). *Adiós a la naturaleza. La revolución bioartefactual [Farewell to Nature. The Bioartifactual Revolution].* Madrid: Plaza y Valdés.

Nicol, E. (1977). *La idea del hombre. 2ª. versión [The Idea of Man, 2nd version].* Mexico: Fondo de Cultura Económica.

Nicol, E. (1980). *La reforma de la filosofía [The Reform of Philosophy].* Mexico: Fondo de Cultura Económica.

Pico della Mirandola, G. (2020). *Discurso sobre la dignidad del hombre [Discourse on the Dignity of Man].* Barcelona: Arpa.

Riechmann, J. (2004). *Gente que no queire viajar a Marte. Ensayos sobre ecología, ética y autocontención [People Who Don't Want to Travel to Mars. Essays on Ecology, Ethics, and Self-Limitation].* Madrid: Libros de la Catarata.

Yuste, R. (2024). *El cerebro: el teatro del mundo [The Brain, the Theater of the World].* Barcelona: Paidós.

CHAPTER TWO

THE NEW POSSIBILITIES OF DIGITAL ECONOMIC-INDUSTRIAL DEVELOPMENT

MEXICO'S SEMICONDUCTOR STRATEGY AND THE EMERGENCE OF A DIGITAL INDUSTRIAL POLICY FOR THE AI ERA: POSITIONING THE COUNTRY AS A STRATEGIC HUB IN THE GLOBAL AI ECONOMY

CONTENT:

I. Introduction **II.** The Evolution of Mexico's Semiconductor Strategy **III.** The New Geopolitics of Technology **IV.** Mexico within the Global Semiconductor Value Chain **V.** Digital Infrastructure as Critical Infrastructure **VI.** Mexico as a Strategic Node in Emerging Digital Chains **VII.** Policy Recommendations **VIII.** Conclusion **IX.** Bibliography

> DIEGO FLORES JIMÉNEZ

Head of the Electronics and Digital Industry Sector at the Secretariat of Economy. He holds a Master's degree in Government and Public Policy from Universidad Panamericana and a Bachelor's degree in International Relations from the Instituto Tecnológico Autónomo de México (ITAM). He specializes in national and regional security, having completed specialized studies at the School of Intelligence for National Security of the National Intelligence Center.

I. INTRODUCTION

Artificial intelligence has become the defining general-purpose technology of the twenty-first century. Unlike earlier waves of digitalization, contemporary AI systems depend on unprecedented computational capacity, massive volumes of data, advanced cloud infrastructures, hyperscale data centers, sophisticated communications networks, and, crucially, semiconductors. The strategic significance of semiconductors therefore extends far beyond their role as industrial inputs. They have become the foundational infrastructure of the digital economy and an essential determinant of national competitiveness, technological sovereignty, and economic security.

The emergence of generative AI and the accelerated development of applied quantum technology has profoundly altered perceptions regarding semiconductor production. Demand for advanced chips has risen dramatically as firms, governments, and research institutions invest in AI training and inference capacities. Simultaneously, geopolitical tensions, particularly between the United States and China, have transformed semiconductor supply chains into arenas of strategic competition; including the conflicts in the Middle East region, particularly the Hormuz Strait situation that affects the fragile global supply chains. Governments increasingly view chips not merely as commercial products but as assets critical to national resilience and geopolitical influence.

Within this changing landscape, Mexico occupies a unique position. The country combines deep industrial integration with the United States; a large manufacturing base; a growing ecosystem of semiconductor-related firms engineering capabilities; long-term specialized advanced manufacturing skills; and geographic proximity to major North American technology clusters. OECD analysis highlights Mexico's strengths in semiconductor design activities,

assembly, testing and packaging (ATP), trade openness, skilled human capital, and integration into electronics and automotive value chains (OECD, 2026). The FUMEC-USAID semiconductor roadmap similarly identifies Mexico as an emerging strategic partner in the effort to develop a more resilient North American semiconductor ecosystem (**FUMEC & USAID**, 2024).

However, attracting semiconductor investments alone will not guarantee meaningful economic transformation. The experience of many export-oriented economies demonstrates that participation in global value chains can generate manufacturing growth without necessarily creating domestic technological capabilities. Consequently, the key policy question for Mexico is no longer whether the country can enter semiconductor value chains, but whether it can leverage existing opportunities to build a comprehensive Digital Industrial Policy capable of supporting innovation, advanced manufacturing, AI infrastructure, and domestic technological development.

This paper argues that Mexico's semiconductor strategy should be understood as the foundation of a broader Digital Industrial Policy. Such a policy must integrate industrial development, digital infrastructure, talent formation, innovation systems, energy policy, data governance, and regional economic integration. Only through this broader approach can Mexico position itself as a strategic hub in the emerging AI-driven economy.

II. THE EVOLUTION OF MEXICO'S SEMICONDUCTOR STRATEGY

Mexico's contemporary semiconductor strategy did not emerge spontaneously. Rather, it evolved through a gradual process involving industry associations, academic institutions, research centers, federal agencies, state governments, international organizations, and North American stakeholders.

The FUMEC-USAID roadmap represented one of the first systematic efforts to identify Mexico's opportunities within the North American semiconductor supply chain (**FUMEC & USAID, 2024**). The study emerged from extensive consultations, interviews, workshops, and regional analyses designed to identify existing capabilities and future opportunities. Rather than focusing solely on attracting fabrication facilities, the roadmap examined technology capacities, workforce requirements, regional strengths, innovation ecosystems, and infrastructure needs.

A significant contribution of the roadmap was its recognition that Mexico already participates in multiple semiconductor-related activities. Existing

strengths include electronics manufacturing, semiconductor design services, engineering activities, automotive electronics, aerospace applications, and ATP-related capabilities. Consequently, the roadmap shifted the policy discussion from creating an industry from scratch toward upgrading and expanding capabilities that already exist.

Subsequently, Mexico's semiconductor agenda gained momentum through numerous national and international forums, including collaborative mechanisms involving government agencies, CANIETI, academic institutions, and North American partners. OECD analysis observed the emergence of multiple semiconductor-related initiatives at federal and state levels and identified increasing policy coordination efforts through institutional agreements established in 2024 (OECD, 2026).

The iteration of these initiatives achieved a major milestone by being incorporated into Plan México, and becoming public policy, which gave the semiconductors sector to the status of a strategic national priority. This transition is important because it reflects a conceptual shift. Semiconductors are no longer viewed exclusively as an industrial sector. Instead, they increasingly function as an enabling platform for broader economic transformation.

This evolution also reflects an important lesson derived from international experience. Semiconductor policy cannot be reduced to investment attraction. Successful ecosystems require coordinated interventions across education, research, infrastructure, financing, intellectual property, energy systems, regulatory frameworks, and digital governance. Mexico's policy evolution therefore represents a movement from sector-specific promotion toward a more integrated development paradigm.

III. THE NEW GEOPOLITICS OF TECHNOLOGY

The contemporary semiconductor industry cannot be understood outside the context of geopolitical transformation. Over the past decade, competition between the United States and China has increasingly centered on technological capabilities, particularly in advanced semiconductors, artificial intelligence, quantum computing, and strategic digital infrastructure.

The COVID-19 pandemic exposed vulnerabilities associated with highly concentrated semiconductor production networks. Supply disruptions affected industries ranging from automobiles to consumer electronics, demonstrating the risks associated with excessive geographic concentration.

Policy responses have been substantial. The United States enacted the CHIPS and Science Act, mobilizing significant public resources to strengthen domestic semiconductor production, research, and workforce development. Similar initiatives have emerged in Europe, Japan, South Korea, and other economies.

Three concepts increasingly define this new industrial geography: friendshoring, nearshoring, and economic security: **Friendshoring** seeks to relocate strategic production to trusted partners; **Nearshoring** prioritizes geographic proximity to reduce logistical vulnerability; and, **Economic security** links industrial capabilities to broader national security objectives.

Mexico benefits directly from these dynamics. Its participation in USMCA, geographic proximity to the United States, established and specialized advance manufacturing ecosystem, and extensive cross-border industrial integration position the country as a natural partner in North American supply-chain diversification.

Nevertheless, geopolitical opportunities should not be mistaken for automatic development outcomes. Countries that successfully leverage geopolitical shifts typically combine openness with strategic capability-building. The critical challenge for Mexico is ensuring that geopolitical advantage translates into technological upgrading and adding value, rather than simply expanded assembly operations.

IV. MEXICO WITHIN THE GLOBAL SEMICONDUCTOR VALUE CHAIN

Mexico already participates in important segments of the semiconductor ecosystem. OECD analysis identifies semiconductor design capabilities alongside assembly, testing, and packaging activities as key components of the domestic industry (OECD, 2026).

The country's strongest competitive advantage lies at the intersection of manufacturing competence, engineering talent, and regional integration. Industrial clusters in Baja California, Chihuahua, Jalisco, Nuevo León, and other regions possess extensive experience in electronics manufacturing and advanced production systems.

Yet, Mexico's participation remains concentrated in activities characterized by relatively lower value capture compared with cutting-edge design, advanced process technologies, semiconductor equipment production, and frontier research.

Moving upward within semiconductor value chains requires strategic prioritization. Rather than competing directly in leading-edge wafer fabrication,

where capital requirements frequently exceed tens of billions of dollars. Mexico could focus on higher-value niches aligned with existing strengths. These include supply chain, RAM and DRAM markets, legacy chips, chip design, embedded systems, heterogeneous integration, advanced packaging, testing services, automotive semiconductors, power electronics, and specialized engineering services.

The more fundamental challenge concerns domestic value creation. OECD evidence highlights the historically limited domestic contribution within electronics exports (**OECD, 2026**). Expanding local supplier networks, increasing domestic research intensity, strengthening technology transfer, and developing national firms are therefore essential components of industrial upgrading.

V. DIGITAL INFRASTRUCTURE AS CRITICAL INFRASTRUCTURE

A defining characteristic of the AI era is the convergence of physical and digital infrastructures. In previous industrial revolutions, competitiveness depended on ports, railways, highways, and electricity. In the twenty-first century, competitiveness increasingly depends on fiber-optic networks, cloud computing, data centers, AI compute capacity, digital platforms, and secure data ecosystems.

Semiconductors provide the hardware foundation of this infrastructure. However, their economic value materializes only when integrated with broader digital systems.

Data centers are a particularly important dimension of this transformation. AI development requires massive computational resources. Training and deploying advanced models increasingly depend on hyperscale computing infrastructure connected through high-capacity telecommunications networks.

Mexico possesses distinctive advantages for attracting data-center investment. Geographic proximity to the United States, growing demand for digital services, industrial integration, and expanding renewable energy potential create favorable conditions. Yet significant constraints remain.

Electricity availability emerges repeatedly across OECD and industry analyses as a critical challenge. Semiconductor production, data centers, and AI infrastructure all require reliable, high-quality power supplies. **Monitor** Deloitte's recommendations specifically emphasize the need for reliable energy infrastructure, faster permitting, renewable energy integration, and semiconductor-focused industrial clusters supported by guaranteed energy capacity (Monitor Deloitte, 2026).

Digital infrastructure must therefore be treated as strategic infrastructure. The future competitiveness of Mexico's industrial economy will depend increasingly on its capacity to store, process, transmit, and secure data.

CHALLENGES FOR A DIGITAL INDUSTRIAL POLICY

The transition from a semiconductor strategy toward a Digital Industrial Policy requires addressing multiple structural constraints

Talent and Skills

Human capital remains perhaps the most important long-term determinant of competitiveness. OECD (**2026**) recommendations consistently emphasize STEM education, English-language proficiency, vocational training reform, industry-academia collaboration, reduced dropout rates, and semiconductor-specific educational programs.

The challenge extends beyond workforce quantity. AI, semiconductors, and advanced digital industries require specialized competencies that evolve rapidly. Consequently, educational institutions must become more adaptive and closely integrated with industrial ecosystems.

Innovation and Research

Mexico possesses important research institutions working in semiconductor-related fields. Nevertheless, research commercialization remains limited.

Future policy should focus on converting scientific capabilities into commercial innovation through stronger industry-academia collaboration, technology transfer mechanisms, intellectual property frameworks, startup ecosystems, and venture financing.

DATA GOVERNANCE

Industrial competitiveness increasingly depends on data governance. AI systems require large-scale access to data while simultaneously demanding robust privacy, cybersecurity, and governance mechanisms.

Mexico lacks a comprehensive framework linking industrial policy, artificial intelligence, data infrastructure, and digital sovereignty. Developing such a framework is essential to supporting domestic AI ecosystems.

FINANCING

One of the most persistent obstacles to technology-intensive development involves access to long-term capital.

Monitor Deloitte **(2026)** recommends public guarantee mechanisms, public-private financing vehicles, and greater participation from institutional investors, including pension funds. Such instruments could reduce investment risks while supporting ecosystem development.

INSTITUTIONAL COORDINATION

OECD **(2026)** findings repeatedly stress the need for coordination mechanisms spanning government agencies, regional authorities, industry associations, universities, and research centers.

Digital Industrial Policy inherently crosses institutional boundaries. Without coordination, fragmented initiatives risk generating duplication, inefficiency, and policy uncertainty.

VI. MEXICO AS A STRATEGIC NODE IN EMERGING DIGITAL CHAINS

Mexico's strategic potential derives from geography, but its future depends on capabilities.

North America is gradually evolving toward a more integrated technology region. The United States provides frontier innovation, advanced research, venture capital, hyperscale computing infrastructure, and semiconductor fabrication investments. Canada contributes research strengths and specialized technological capabilities. Mexico provides manufacturing scale, engineering talent, industrial integration, and geographic connectivity.

This complementariness suggests that Mexico's long-term role should not be defined primarily as a manufacturing platform. Rather, the country has the potential to become a strategic node connecting digital innovation, advanced manufacturing, semiconductor services, AI infrastructure, and regional value chains.

Such a position would generate benefits that extend beyond economic growth. Increased domestic technological capabilities would strengthen resilience, reduce vulnerability to supply disruptions, expand opportunities for local firms, and improve productivity across multiple sectors.

Importantly, this vision is not based on technological nationalism. Semiconductor ecosystems are inherently global. The objective is not self-sufficiency

but strategic participation in high-value segments of interconnected global networks.

VII. POLICY RECOMMENDATIONS

RECOMMENDATION 1: Establish a National Digital Industrial Policy Council

Problem: Fragmented governance.

- **Objective**: Integrate industrial, digital, innovation, energy, education, and AI policies.
- **Mechanism**: Permanent coordination body involving federal agencies, states, industry, academia, and civil society.
- **Expected Impact**: Reduced policy fragmentation and stronger strategic coherence.

RECOMMENDATION 2: Create a National Semiconductor and AI Talent System

Problem: Skills mismatch.

- **Objective**: Align education with industry demand.
- **Mechanism**: National standards, competency frameworks, scholarships, industry partnerships, and international exchanges.
- **Expected Impact**: Expansion of specialized talent.

RECOMMENDATION 3: Develop Strategic AI and Semiconductor Infrastructure Zones

Problem: Infrastructure bottlenecks.

- **Objective**: Concentrate investment in globally competitive technology clusters.
- **Mechanism**: Integrated zones combining renewable energy, fiber networks, data centers, research facilities, and industrial parks.
- **Expected Impact**: Attraction of higher-value investments.

RECOMMENDATION 4: Launch a Semiconductor Innovation Fund

Problem: Weak commercialization of research.

- **Objective**: Accelerate technology transfer.
- **Mechanism**: Competitive grants, venture co-financing, and university-industry collaboration programs.
- **Expected Impact**: Increased domestic innovation output.

RECOMMENDATION 5: Establish a National Data Infrastructure Strategy
Problem: Lack of integrated data policy.

- **Objective**: Support AI development and digital transformation.
- **Mechanism**: Data governance standards, cloud infrastructure development, cybersecurity frameworks, and public data platforms.
- **Expected Impact**: Enhanced digital competitiveness.

RECOMMENDATION 6: Strengthen Domestic Supplier Development
Problem: Limited national content.

- **Objective**: Increase domestic value added.
- **Mechanism**: Supplier mapping, technical assistance, certification programs, and procurement linkages.
- **Expected Impact**: Greater local participation in semiconductor value chains.

VIII. CONCLUSION

Mexico's semiconductor opportunity should not be understood merely as an industrial opportunity. It is a development opportunity of historic proportions. The convergence of AI, geopolitical realignment, North American integration, and supply-chain restructuring has created conditions rarely observed in modern economic history.

Yet the success of Mexico's strategy will not be determined by the number of factories installed or by investment announcements alone. The decisive question concerns whether the country can transform external opportunities into domestic capability of upgrading its national industry into the high value chain segments.

The evidence reviewed throughout this paper suggests that Mexico already possesses many of the foundations required for success: manufacturing capabilities, engineering talent, established industrial ecosystems, geographic proximity to major markets, and growing policy attention. However, converting these advantages into sustainable technological leadership requires moving beyond a narrowly defined semiconductor agenda.

The future lies in constructing a comprehensive Digital Industrial Policy that integrates semiconductors, artificial intelligence, digital infrastructure, innovation, talent development, energy systems, and regional economic strategy. Under such a framework, semiconductors become not the final objective but the

enabling platform through which Mexico can participate more effectively in the emerging architecture of the AI-driven global economy, and change its national development matrix, securing a place in the digital economy for decades to come, by reassuring its key role in the hemisphere and in the top 20 economies.

This conclusion would be incomplete without engaging directly with the most comprehensive proposal advanced for Mexico's digital-industrial future: the Sistema Mexicano de Política Industrial Digital (Mexican Digital Industrial Policy System) put forward by Oropeza García and Berasaluce Iza (2024). Their proposal argues that Mexico's technological transformation cannot rest on isolated sectoral initiatives, however promising, but requires a permanent institutional architecture for the period 2024-2040, built around a National Council for Digital Industrial Policy, a long-term National Digital Industry Plan, a dedicated multiannual budget, and a Permanent Evaluation Committee tasked with tracking measurable national targets, including raising domestic content in exports to 40 percent, ICT exports to 20 percent of total exports, and total investment to 35 percent of GDP (Oropeza García & Berasaluce Iza, 2024). This proposal reinforces, almost point for point, the argument developed throughout this paper: that semiconductor attraction is a necessary but insufficient condition for upgrading, and that only a coordinated, whole-of-government framework can convert Mexico's manufacturing and geographic advantages into durable technological capability.

Building on this foundation, the public policy this paper proposes for positioning Mexico as a strategic node in the emerging global digital value chains is the institutionalization of a single, statutorily grounded National Digital Industrial Policy System that fuses the architecture proposed by Oropeza García and Berasaluce Iza (2024) with the sector-specific recommendations developed in this paper. This system should rest on four pillars. First, a permanent National Digital Industrial Policy Council –merging this paper's Recommendation 1 with Oropeza García and Berasaluce Iza's (2024) proposed Consejo Nacional– should bring together federal and state government, industry associations such as CANIETI, the innovation ecosystem, and academia under a single coordinating body empowered to align semiconductor, artificial intelligence, energy, and education policy. Second, a legally binding National Digital Industry Plan spanning 2026-2040, backed by a ring-fenced multiannual budget directed toward infrastructure and progressively rising development-bank financing, should fund the Strategic AI and Semiconductor Infrastructure Zones proposed earlier in this paper (Recommendation 3), directly operationalizing Oropeza García and Berasaluce Iza's (2024) call for sustained public investment. Third, a

Permanent Evaluation Committee should track a unified scorecard combining this paper's innovation and talent-formation indicators with the national targets proposed by Oropeza García and Berasaluce Iza (2024), keeping the policy adaptive to the pace of technological change described throughout this analysis. Fourth, sectoral prioritization should combine the strategic sectors identified by Oropeza García and Berasaluce Iza (2024) –including semiconductors, aerospace, automotive, and electromobility– with the higher-value semiconductor niches identified in this paper, such as advanced packaging, chip design, and testing services, concentrating public support where Mexico's existing manufacturing and engineering base can most credibly move up the value chain.

Implemented together, these four pillars would transform Mexico's role from that of an assembly platform into the connective hub described earlier in this paper: a strategic node linking United States innovation and hyperscale computing capacity, Canadian research strength, and Mexico's own manufacturing scale, engineering talent, and geographic proximity within an increasingly integrated North American technology region. This is precisely the outcome that Oropeza García and Berasaluce Iza (2024) envision for Mexico's digital industrial policy, and it is the policy pathway this paper ultimately proposes: not the pursuit of technological self-sufficiency, but strategic, well-governed participation in the highest-value segments of the emerging global digital value chains.

Future research should explore quantitative scenarios regarding semiconductor upgrading, quantum, AI infrastructure deployment, national value-added creation, and regional innovation spillovers, as well as comparative analysis of industrial-policy models implemented in East Asia, North America, and Europe.

IX. BIBLIOGRAPHY

Fundación México-Estados Unidos para la Ciencia (FUMEC), & U.S. Agency for International Development (USAID). (2024). *Mapa de ruta: Oportunidades para el nearshoring de semiconductores en México* [Roadmap: Opportunities for semiconductor nearshoring in Mexico]. https://fumec.org/semiconductors

Monitor Deloitte. (2026). *Recommendations to develop Mexico's semiconductor industry: U.S.-Mexico Semiconductor Collaboration Forum* (Cap. 7, San Diego).

Organisation for Economic Co-operation and Development. (2026). *Promoting the development of the semiconductor ecosystem in Mexico*. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/promoting-the-development-of-the-semiconductor-ecosystem-in-mexico_ab1c92e8/02c81dec-en.pdf

Oropeza García, A., & Berasaluce Iza, J. (2024). La transformación digital y la política industrial: Retos y oportunidades [Digital transformation and industrial policy: Challenges and opportunities]. Instituto para el Desarrollo Industrial y la Transformación Digital (INADI).

IMAGINED FUTURES, FICTIONAL EXPECTATIONS AND DYNAMICS OF CAPITALISM; THE ROLE OF ARTIFICIAL INTELLIGENCE

CONTENT:

I. Imagined futures and fictional expectations **II.** Social and Ethical Impacts **III.** Properties of a different capitalism **IV.** The Dynamics of Capitalism in the Age of AI **V.** Conclusion and Final Reflections **VI.** Bibliography

> CLEMENTE RUIZ DURÁN

Professor and Researcher at the Graduate Program in Economics, Faculty of Economics, UNAM. The preparation of this document was supported by Miguel Ángel Hernández-Ruiz, Rafael Ulises Ramírez-López, and Jean Pablo Vite-Sixto.

The interaction between imagined futures, fictional expectations, and the dynamics of capitalism takes on new relevance in the era of artificial intelligence (AI). This paper explores the topic through some key reflections that allow us to construct scenarios that structure the reality we are building with the emergence of artificial intelligence.

I. IMAGINED FUTURES AND FICTIONAL EXPECTATIONS

The concept of imagined futures, as conceptualized by Jens Beckert in his book "Fictional Expectations and Capitalist Dynamics: Imagined Futures", relates to how societies anticipate what comes through narratives, myths, and cultural representations. In the context of capitalism, these fictional expectations function as an engine that drives economic, political, and technological decisions. Currently, the rise of "smart cities" and the expectation of total automation are based more on fictional visions than on concrete realities. Narratives such as the "technological singularity" or the colonization of Marte reflect capitalist aspirations that shape investment and scientific-technological development. Capitalism feeds on these fictions to perpetuate the cycle of production and consumption. The promises of innovation and progress become commodities that sell specific futures, often favoring technology corporations.

Capitalism, characterized by its ability to absorb and transform any innovation into economic value, finds a strategic ally in AI. This is manifested in the acceleration of the market by optimizing production, distribution, and consumption processes, which further strengthens the dominance of digital giants like Amazon, Google, and Meta. This transformation is based on the massive exploitation of personal data, considered the new "oil", with AI as a

key tool for extracting value, which amplifies existing inequalities. Furthermore, many AI companies promote the idea that their technology is destined to "disrupt" entire industries, fueling an endless cycle of innovation and planned obsolescence that calls into question the long-term stability of economic and social structures.

AI acts as both a creator and destroyer of fictions. We must be aware that artificial intelligence not only executes tasks but also actively participates in the creation of fictional futures. Generative algorithms, such as language models, have the ability to create futuristic stories, advertising campaigns, and even economic predictions that can influence the collective imagination. However, this technology also poses significant challenges, such as the erosion of authenticity. The proliferation of deepfakes and narratives created by AI calls into question the trust in traditional sources of information, making it difficult to verify the truthfulness of what we consume. This is compounded by the growing debate about the energy that will be needed to meet the increasing demand for this technology. Additionally, automated decisions based on algorithmic predictions are having a tangible impact in areas such as financial markets, public policies, and social behavior, which highlights the growing power that AI is having in our daily lives.

II. SOCIAL AND ETHICAL IMPACTS

The intersection of fictional expectations and the dynamics of capitalism, exacerbated by AI, poses ethical and social challenges:

Digital exclusion is a significant consequence of the advancement of artificial intelligence and other advanced technologies. As these technologies become increasingly integrated into various aspects of daily life, people who do not have access to them or lack the necessary skills to use them can be left behind. This can result in a growing gap between those who can benefit from the opportunities offered by technology and those who cannot, exacerbating social and economic inequalities. It is crucial to address this digital divide to ensure that everyone has the opportunity to participate in the technological future.

Artificial intelligence plays a crucial role in surveillance capitalism, where advanced technologies are used to monitor public spaces. The promises of AI are intertwined with systems of social control, such as facial recognition and social credit platforms, which allow for detailed tracking and evaluation of individuals. Facial recognition can identify and track people in real time in public spaces, while social credit platforms can classify citizens according to

their behavior and compliance with norms, affecting their access to services and opportunities. Although these AI applications can offer benefits in terms of security and efficiency, they also raise serious concerns about privacy, individual freedom, and the potential for abuse of power. It is essential to develop ethical and legal frameworks that regulate the use of these technologies to protect people's rights and freedoms.

What futures are possible? Imagined futures are often monopolized by the media, sidelining alternative visions of more inclusive and sustainable development.

III. PROPERTIES OF A DIFFERENT CAPITALISM

A key question that arises from this rethinking of reality is the following: Is it possible to design AI systems that promote equity and sustainability? In this context of reimagining reality, movements like solarpunk or degrowth offer alternative visions for the future, proposing models that transcend traditional capitalist dynamics. These approaches suggest that it is possible to build a more just and sustainable society, where technologies, including artificial intelligence, are used for collective benefit and not just for a few. In this sense, if we manage to ensure that the benefits of AI are within everyone's reach, we could mitigate the inequalities that would otherwise be intensified by the implementation of these technologies. In this way, AI could not only contribute to efficiency and innovation, but also become a driver of social transformation that promotes a more equitable and sustainable future. This interaction between AI, narratives of the future, and capitalism suggests that "imagined futures" are not inevitable, but rather depend on the political, ethical, and social decisions we make today. The emergence of this new vision intertwines profound elements of sociology, economic theory, and the philosophy of technology. The interaction between imagined futures, fictional expectations, and the dynamics of capitalism acquires a new relevance in the age of artificial intelligence. Below, we offer some structured reflections to explore this topic.

A. IMAGINED FUTURES AND FICTIONAL EXPECTATIONS

Imagined futures are projections of how the world could be, built from dreams, longings, and assumptions. Throughout history, these concepts have been an essential driver for the progress of humanity, as they motivate policies,

technology projects, and business strategies. Ideas like smart cities or the colonization of Mars demonstrate how these future perspectives can foster progress, although they are often based more on assumptions than on situations that are truly tangible.

In capitalism, these narratives not only inspire, but they also become tools to mobilize resources, attract investment, and justify large projects. Promises of innovation and progress are presented as benefits for all, but in practice, they often concentrate the results in the hands of a few companies or regions, deepening existing inequalities. After all, every great breakthrough begins as a simple idea.

Therefore, although these visions of the future are necessary to move forward, we must analyze them critically. Only in this way can we ensure that the decisions we make today build a more equitable, sustainable, and inclusive future, not only for our generation, but for those to come. Solving the problems of the present with an eye toward the well-being of those who will follow us is, ultimately, what truly matters.

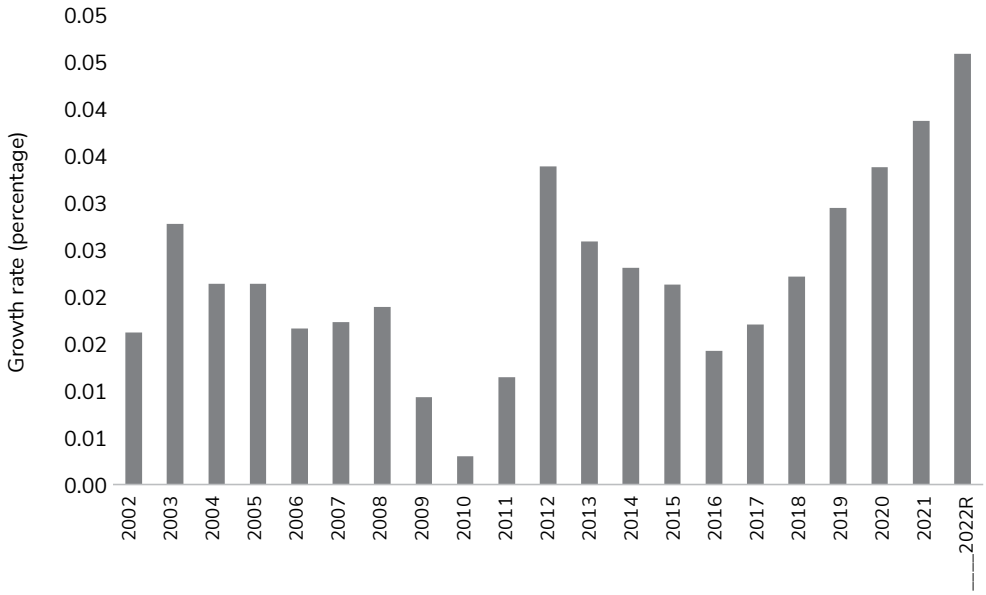
IV. THE DYNAMICS OF CAPITALISM IN THE AGE OF AI

Artificial intelligence has transformed the way markets, industries, value chains, and production operate. One of its most visible impacts is the acceleration of the market, as it optimizes production, distribution, and consumption processes, consolidating the power of technology giants.

Narratives of disruption have become a central element in the discourse of many artificial intelligence companies, which promote their technologies as tools capable of radically transforming entire industries. This approach not only reinforces the idea that technological progress must be constant and accelerated, but it also perpetuates dynamics of continuous innovation that often lead to planned obsolescence. By emphasizing the need to adopt the latest tools and solutions so as not to "be left behind", these narratives generate constant pressure on companies and consumers, increasing resource consumption and exacerbating inequalities by privileging those who have access to the most advanced technologies. Furthermore, this vision of innovation as an end in itself can hinder deeper reflections on how technology could be developed and applied to promote equity, sustainability, and collective well-being.

In this scenario, the total factor productivity model, expanded by the KLEMS[1] model, is presented as a valuable tool for understanding how sectors like software in Mexico can drive more inclusive and sustainable economic growth. This approach analyzes in detail how factors such as capital, labor, and materials contribute to the country's development. By connecting this data with ideas of alternative futures, we can imagine how strategic decisions could transform the software industry into a force for reducing inequalities and promoting a balance with the environment. The graph below helps us visualize these contributions and reflect on the role that each choice has in building the future. The KLEMS methodology helps us understand more clearly how economic growth is composed, by breaking it down into key factors such as capital, labor, energy, materials, and intermediate services. In this case, graph 1 shows how the software sector has contributed to Mexico's economic growth between 2002 and 2022, highlighting the fundamental role of intangible capital, such as software, in improving total factor productivity (TFP).

FIGURE 1 Mexico (2002-2022): Percentage rate of software contribution to growth



SOURCE: Author's own elaboration, based on data from INEGI (KLEMS Factor Productivity).

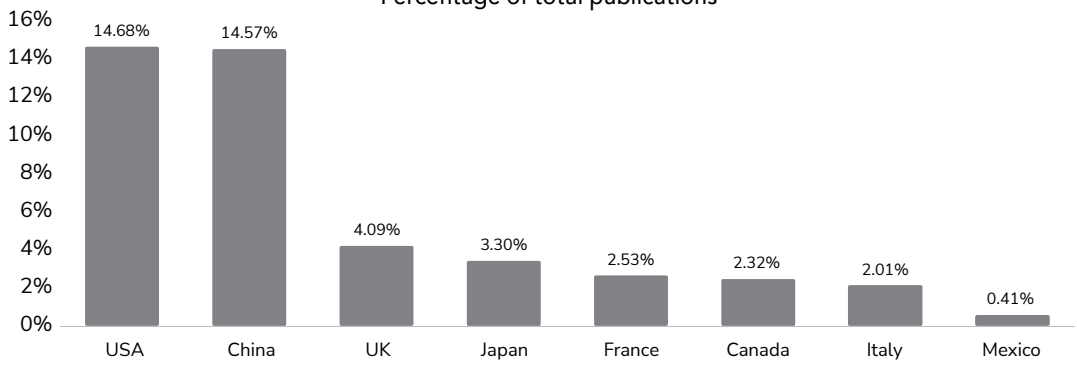
1 The KLEMS model developed by Dale Jorgenson of Harvard University in 2010

Throughout the analyzed period, the contribution of software has shown ups and downs. In the years 2009 and 2011-2013, significant peaks are observed, likely driven by strategic investments in technology and digital transformation processes. On the other hand, 2017 stands out for a notable drop, possibly related to a reduction in investments or changes in technology policies. However, starting in 2020, a sustained recovery is perceived, fueled by the acceleration of digitalization in response to the pandemic. This rebound suggests that the software sector is playing an increasingly relevant role in improving productivity and in strengthening the Mexican economy.

B. AI AS A CREATOR AND DESTROYER OF FICTIONS

Artificial intelligence, especially through generative algorithms like language models, has a significant impact on narrative generation. These algorithms can create futuristic stories, advertising campaigns, and economic predictions that influence the collective imagination in various ways. However, for Mexico, AI represents both an opportunity and a challenge: the country is at a crossroads where it can become a regional technological leader or fall behind in the global race. For the country to fully take advantage of these opportunities and remain globally competitive, it is crucial to increase the production of publications and studies on AI. A greater number of research and publications will not only foster innovation and technological development in the country, but will also position Mexico as a leader in the field of AI. The following graph illustrates the urgent need to increase publications on AI in Mexico, by comparing it with other countries.

FIGURE 2 Selected countries (2000-2023): Number of AI publications
Percentage of total publications



SOURCE: Author's own elaboration, based on data from the AI Index Report (2023)

TABLE 1 Mexico: AI publications by topic

Topic	Publications
Swarm intelligence optimization algorithms	108
Type 2 fuzzy logic systems and applications	104
Analysis of the economic impact and urban development	93
Fundamentals and applications of neural networks	87
Industry 4.0 and digital transformation in the manufacturing industry	85
Artificial intelligence in medicine	76
Sampling-Based Motion Planning Algorithms	69
Detection and diagnosis of process failures in industries	66
Educational data mining and learning analytics	61

SOURCE: Author's own elaboration, based on data from OpenAlex (2024).

To address the need to increase publications on artificial intelligence in Mexico, an effective solution would be to foster collaboration and support among universities, research centers, and the private sector. Universities can play a crucial role by establishing specialized research programs in AI, offering incentives for the publication of studies, and facilitating access to resources and funding. Furthermore, promoting the creation of research networks and consortia that include national and international experts can enrich the quality and scope of the publications. Implementing these strategies will not only increase academic production in AI but will also strengthen Mexico's position in the global landscape of technology and innovation.

TABLE 2 Mexico: Publications on artificial intelligence by institution 2000-2024

Institution	Publications
National Polytechnic Institute	966
Monterrey Institute of Technology and Higher Education	859
National Autonomous University of Mexico (UNAM)	842
Ministry of Public Education	372
National Technological Institute of Mexico	369
University of Guadalajara	359

SOURCE: Author's own elaboration, based on data from OpenAlex (2024).

However, the use of artificial intelligence can have significant destructive consequences, such as the erosion of authenticity and performative predictions. The proliferation of deepfakes and AI-generated narratives calls into question the trust in traditional sources of information, as it becomes increasingly difficult

to distinguish between what is real and what is fabricated. This can undermine the credibility of the media and institutions, generating widespread mistrust. Furthermore, automated decisions based on algorithmic predictions have tangible effects on financial markets, public policies, and social behavior. These performative predictions can influence decision-making in a way that reinforces certain patterns and trends, sometimes detrimentally, affecting stability andv equity in various areas.

C. STRUCTURAL RESTRUCTURING FROM AI

Artificial intelligence is presented as a technology with infinite possibilities, capable of revolutionizing multiple industries by optimizing processes, creating new business models, and redefining the value of data. AI is not only transforming the world, but it places us on the brink of a true digital revolution. The rapid advancement of artificial intelligence has captured global attention, generating both enthusiasm and concern, and sparking a crucial debate about its potential impact on the world economy.

Global investment in artificial intelligence has shown significant growth in several key areas, reflecting the transformative potential of this technology in strategic sectors. The customer service and natural language processing sector has led the investment, highlighting the importance of AI in improving communication and efficiency in services. For its part, medicine and healthcare continue to be an area of strong investment, underscoring the impact of AI in optimizing diagnoses, treatments, and the management of health resources. Data management and processing has also consolidated itself as a crucial area, showing how AI's ability to analyze large volumes of data can drive more informed and precise decision-making.

Furthermore, AI infrastructure and governance are gaining ground as essential areas to ensure the ethical and sustainable development of the technology, which is indispensable for addressing the challenges of transparency, regulation, and security. Other sectors, such as educational technology and energy, are also emerging as key investment areas, with enormous potential to transform the way we teach, learn, and manage energy resources, driving a transition toward more sustainable and efficient economies. With this data in mind, Mexico is in a privileged position to capitalize on these global trends, effectively integrating AI into sectors such as health, education, and energy, which would not only foster its economic development but also contribute to the creation of a more just and competitive environment globally.

TABLE 3 Global investment on AI 2013-2023

Billions of dollars at constant 2021 prices

Total	**568.65**
Natural language processing and customer service	48.05
Medicine and healthcare	33.21
Data management and processing	22.27
AI Infrastructure and Governance	17.82
Educational technology	12.47
Energy, oil and gas	12.25
Autonomous vehicles	11.69
Fintech	9.68
Creative, musical and video content	9.25
Semiconductors	9.18
Factial recognition	9.04
Drones	7.62
Insurance technology	5.09
Venture capital	4.71
Cibersecurity and data protection	4.62

SOURCE: Author's own elaboration based on data from Quid (2024).

Mexico has consolidated its leadership in the growth of artificial intelligence companies in Latin America, with an increase of 965% in the last six years, reaching 362 companies in 2024. This growth has generated more than 11,000 jobs and attracted investments exceeding 500 million dollars. Furthermore, Mexico City has become the main hub for AI development, housing almost half of the country's startups. Companies in this sector, which include key areas such as machine learning and robotics, report average revenues of 1.3 million dollars in 2024, reflecting an expanding market with great potential.

TABLE 4 AI companies in Latin America 2024

Country	**2018**	**2024**	**Growth rate**
Brazil	124	728	487%
México	34	362	965%
Chile	28	160	471%
Colombia	13	100	669%
Argentina	32	83	159%

SOURCE: https://www.santander.com.mx/ceb/2024/mexico-lidera-crecimiento-de-empresas-de-inteligencia-artificial-en-latam.html

D. SECTORS WITH HIGH POTENTIAL FOR AI

I. Healthcare

Artificial intelligence is no longer just a futuristic promise; it is already a reality. AI is transforming the healthcare sector, opening up new opportunities and challenges. An example illustrating this impact is the work of Demis Hassabis and John Jumper of Google DeepMind, along with David Baker, who won the 2024 Nobel Prize in Chemistry. Their creation, AlphaFold, succeeded in predicting the three-dimensional structures of proteins from amino acid sequences, solving a scientific challenge of over 50 years. This breakthrough allowed for the identification of the structure of almost 200 million proteins, providing tools to address problems such as antibiotic resistance and the design of enzymes capable of breaking down plastics. Furthermore, Baker advanced even further, developing completely new proteins with applications ranging from health to sustainability (Expansión Digital, 2023).

AI is already demonstrating its ability to transform diagnostics and treatments. For example, advanced algorithms trained with more than 60,000 mammograms can now identify subtle patterns in breast tissues, detecting possible malignant tumors at extremely early stages. This development promises to eradicate late diagnoses and revolutionize early cancer detection (National Geographic, 2023).

Additionally, AI has a key role in remote regions where access to doctors is limited. AI-based diagnostic tools can offer precise consultations and treatment recommendations without the need for immediate medical presence. However, this advancement raises questions about the future of certain medical roles, such as general practitioners, who could be partially replaced. There is a great opportunity to improve disease detection, as Kreutzer and Sirrenberg point out, the Camelyon Grand Challenge 2016 demonstrated that collaboration between humans and algorithms can generate better results. In this challenge, an algorithm developed by Harvard and MIT researchers achieved 92.5% accuracy in detecting metastatic breast cancer, while pathologists reached 96.6%. The combination of both approaches resulted in 99.5% accuracy, reducing the margin of error by 85%.

In the future, portable devices equipped with AI could radically change how people manage their health. These devices would be capable of collecting real-time data, recommending diets, suggesting personalized medications, and offering advice tailored to each individual's specific needs, improving overall well-being.

The implementation of AI will not be limited solely to diagnosis and treatment. It also promises to optimize hospital management, allocating resources more efficiently, improving patient flow, and reducing waiting times. Likewise, in drug distribution, AI could allocate resources more precisely, personalizing treatments to maximize results.

II. Finance

The financial sector is one of the industries with the greatest potential to benefit from the implementation of technologies based on artificial intelligence. Historically, this sector has led in technological adoption, pioneering the use of computers for innovations like high-frequency trading.

According to IBM (2023), artificial intelligence in finance is based on the application of advanced technologies, especially machine learning algorithms. These tools allow financial institutions to optimize efficiency, accuracy, and speed in tasks such as data analysis, projections, investment and risk management, fraud detection, and customer service. AI is transforming the sector by automating traditional banking processes and generating customer interactions that mimic human intelligence and behavior.

Plummer (2024) notes that AI is also revolutionizing fraud detection and prevention. Machine learning models can analyze millions of transactions to identify patterns that indicate fraudulent activities. Banks leverage these capabilities to identify suspicious transactions in real-time, significantly reducing losses from fraud, as well as personalizing services. Banks use advanced data analytics to offer product recommendations tailored to customers' spending habits and transaction histories.

Furthermore, Plummer also mentions how AI enables more sophisticated financial analysis and risk modeling. By identifying correlations in vast datasets that would not be apparent to humans, AI improves predictive analytics, scenario planning, and risk assessments. AI has the potential to offer many benefits to the financial sector, such as optimizing portfolio management, efficiently analyzing large volumes of data, making more accurate profitability predictions, optimally assigning credit ratings, and providing more exact risk assessments, positioning AI as the engine of the financial sector's future.

III. Manufacturing

Artificial intelligence has driven the so-called Industry 4.0, where production processes are increasingly automated and connected, which has an impact on operational improvement and cost reduction.

AI has revolutionized industry by promoting productive maintenance, advanced robotics, and supply chain optimization. There are sensors that monitor equipment in real time to prevent failures, while robots with AI work alongside humans on assembly lines, increasing productivity. Furthermore, tools like those used by Amazon predict demand, adjust inventories, and optimize delivery routes. These advances have increased productive efficiency, reducing costs and allowing for the improvement of production processes.

The automotive industry is one of the main sectors that has adopted artificial intelligence and robotics in Mexico, and it is also one of the most important industries in the country. Automotive companies use it to optimize supply chains, improve production planning, and detect vehicle failures before they occur. Artificial intelligence allows for predictive analysis, which reduces inefficient time and improves product quality. In addition, collaborative robots work alongside employees to increase efficiency in the production process.

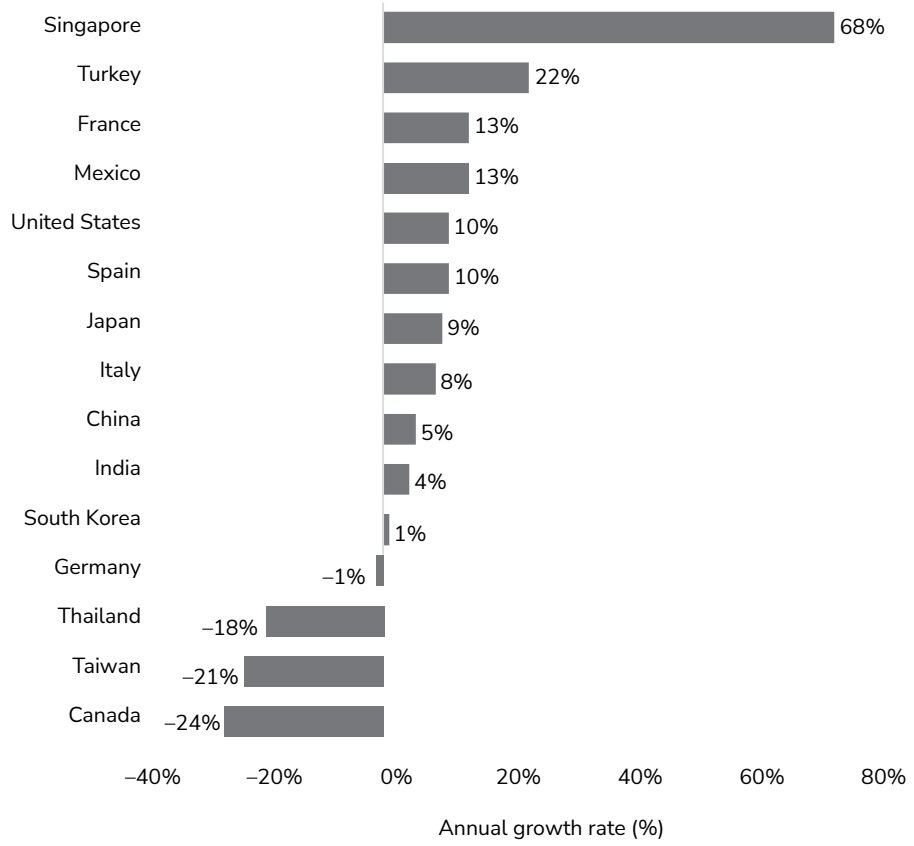
The electronics sector has also integrated AI solutions to optimize its production processes. Machine learning systems identify bottlenecks in production and adjust workflows, maximizing productivity. Computer vision is used for quality control, allowing for more precise and rapid detection of defects in electronic products.

In the food sector, companies are implementing AI to improve operational efficiency and ensure quality standards. Automation mechanisms and machine learning algorithms help manage inventories and forecast demand, optimizing production and reducing costs. In addition, AI technologies allow for automated inspection during the manufacturing process, improving the quality of the final production.

The pharmaceutical industry has also adopted AI technologies to improve its production processes. They use it primarily for the analysis of data related to the development of new drugs, optimizing the research and development process.

The textile sector has implemented AI-based solutions to customize products on a large scale. This not only improves consumer satisfaction by adapting designs to specific trends, but also optimizes the use of resources, reducing costs and increasing efficiency.

FIGURE 3 Annual growth rate of installed industrial robots by country, 2021 vs. 2022



SOURCE: Author's own elaboration based on data from IFR, AI Index Report 2023

The manufacturing industry is greatly benefiting from generative artificial intelligence, especially in areas like product design, process optimization, and automation. According to a report from Stack Overflow[2] (2023), 76% of professionals in the software industry already use or plan to use AI tools in their work. This trend is also reflected in the manufacturing sector, where a growing number of companies are adopting AI technologies to improve productivity and reduce operational costs.

82% of the professionals surveyed highlight that the main benefits of using AI tools include a significant increase in productivity, by allowing for greater

2 Stack Overflow 2023 Developer Survey

speed in design and production. Around 60% of those surveyed also report greater efficiency and accelerated learning, especially in the application of AI for the automation of repetitive tasks and in improving product quality. These tools are especially useful in the simulation of manufacturing processes and the optimization of the supply chain.

Despite this progress, critical challenges persist, such as the validity of data in real-world applications, the transparency and interpretability of AI models, and the need to integrate physical approaches with data-based methods. Opportunities are also identified for the development of generative AI with granular control and generalist systems that interact more naturally with users. The future of manufacturing will depend on the adoption of new tools and methodologies that allow for the leveraging of data-rich environments and a deep understanding of industrial processes. This approach will not only improve operational efficiency but will also consolidate the scientific foundations for advancing toward truly intelligent manufacturing (Gao *et al.,* 2024).

IV. Agriculture

Food waste is a persistent problem that affects both the economy and the environment. It is not just about what is left on the plate, but a massive loss that occurs throughout the entire supply chain, from production to distribution. In the United States, the USDA estimates that between 30% and 40% of the food supply is wasted, an alarming percentage that reflects inefficiencies at different stages of the process. These losses can be due to spoilage during transport, inadequate storage conditions, or even overproduction.

However, data science and artificial intelligence are beginning to change this reality. In agricultural production, technologies such as drones, sensors, and satellites collect data on soil productivity, crop health, and weather conditions. Processed in real time, this data allows farmers to make smarter decisions about crop rotation, planting, and harvesting. Furthermore, predictive algorithms can anticipate pests or diseases, facilitating early and effective interventions. Autonomous agricultural robots complement these tasks by planting, harvesting, and applying pesticides with precision, reducing waste and optimizing resources.

In the distribution stage, AI also plays a key role. Sensors connected to advanced analytical systems monitor critical parameters such as temperature, triggering alerts in case of variations that could compromise the quality of the products. In retail, automated systems for tracking expiration dates help plan promotions and prevent food from losing freshness. On the other hand, prediction algorithms adjust product orders according to demand, aligning supply

and consumption to minimize waste. These technologies not only improve efficiency but also contribute to building a more sustainable and conscious supply chain. (Innovación Digital 360, 2024)

V. Retail trade

AI has transformed shopping experience and logistics operations. Additionally, predictive inventory management systems optimize stock levels, reducing waste and ensuring that products are available when needed. An innovative example of the application of AI is Amazon Go, where the technology allows for a completely automated shopping experience without the need for cashiers.

AI algorithms analyze shopping behavior and suggest personalized products, as Amazon does. This increases the likelihood of cross-selling and repeat sales. Most retailers and large companies have already begun to implement AI-based chatbots; these can answer frequently asked questions from customers and improve customer service 24 hours a day.

These innovations have made it possible to create more personalized shopping experiences, improve efficiency, and develop new frictionless commerce models, benefiting both consumers and retailers.

VI. Education

Artificial intelligence is revolutionizing the way we learn and teach. In the last year, many academic platforms like Khan Academy and Coursera, known for offering videos, courses, and readings, have begun to incorporate chatbots into their pages designed to facilitate learning. These tools can generate personalized quizzes, summarize content, and provide syllabi adapted to the needs of each user.

The potential impact of AI in education goes much further. With the ability to create images and videos, the future could offer us complete classes taught by virtual teachers generated by AI, capable of writing on whiteboards as if they were real teachers. These classes would be completely personalized, adjusting to the pace of each student.

Artificial intelligence systems can identify the strengths and weaknesses of each student and adapt the content to their needs, allowing each one to learn at their own pace. This does not stop here; there is also a way to support teachers and thus improve their performance. For example, AI can take care of grading multiple-choice exams, organize student data, or scheduling activities, freeing up time for teachers to focus on teaching.

Beyond facilitating access to knowledge, AI is opening the doors to a more inclusive, personalized, and efficient educational experience. This has the potential to reduce educational gaps in regions of the world with limited access to quality education. The implementation of these technologies could be key to closing inequalities, promoting development in countries with educational lag, and generating new opportunities for millions of people.

VII. Entertainment and media

Artificial intelligence has transformed both the creation of content and the way users consume it. For example, ChatGPT allows for the automated creation of scripts, images, and audiovisual content, facilitating creative production on an unprecedented scale. Additionally, platforms like Spotify and YouTube use AI to personalize recommendations for music, videos, and podcasts, offering users content adapted to their tastes and preferences. In the film industry, AI optimizes production by predicting the success of movies and managing budgets more efficiently. These innovations have not only democratized content creation but have also improved the user experience by making it more personalized and accessible.

This has affected and benefited large companies, for example, The Walt Disney Company, as it is transforming various areas of entertainment: in content production, as in The Mandalorian, it improves visual effects and narrative by de-aging characters; in theme parks, like Pirates of the Caribbean, it facilitates interactive animatronics that offer personalized experiences; in animated films, such as Elemental, it brings realism and efficiency by simulating elements like fire and water. Furthermore, in content strategies, it optimizes processes and reduces costs through post-production tools, and in immersive experiences, it enriches interaction in theme parks with augmented reality applications.

The use of AI in the film and streaming industry has also had negative consequences, due to the recent SAG-AFTRA union strike, which began on July 14, 2023, and concluded on November 8 after 118 days of stoppage. It was one of the longest in the industry's history. During this period, the Californian economy suffered estimated losses of 6.5 billion dollars, in addition to the layoff of some 45,000 workers. One of the reasons was the growing concern about the use of AI to replicate voices or likenesses without adequate compensation for the actors.

VIII. Energy

Artificial intelligence and Industry 4.0 demand a change in humanity's energy pattern and production processes. With the adoption of artificial intelligence,

technology companies have seen an increase in emissions due to the growing energy demand of AI, particularly from the data centers needed to train and run AI models. As generative AI systems gain popularity, their energy consumption increases, with some models requiring much more energy than traditional software. As AI systems become more advanced, their energy consumption is expected to double approximately every 100 days, putting pressure on already overburdened electrical grids. Companies like Microsoft and Google have reported significant increases in CO2 emissions, partly due to the expansion of their data centers (Kemene et al, 2024).

Despite the high energy demand of AI, it has the potential to drive the energy transition by optimizing the integration of renewable energies, improving energy efficiency in carbon-intensive industries, and contributing to grid stability. AI can help predict energy production, shift energy loads, and optimize the use of renewable energy sources, ensuring efficient grid operation. However, the energy required to support AI's growth poses the challenge of balancing its benefits with environmental costs. As the European Parliament and other regulatory bodies begin to establish requirements for AI systems to be energy-efficient, hardware advancements, such as Nvidia's new "superchip", offer hope for reducing energy consumption (Kemene *et al.,* 2024).

AI has had a significant impact on the energy sector, promoting sustainability and efficiency. Intelligent electrical grid management systems balance supply and demand in real time, ensuring a stable and efficient supply. Climate prediction algorithms analyze meteorological data to optimize the generation of renewable energy, maximizing its use and minimizing waste. Furthermore, smart buildings use AI to automatically adjust lighting and temperature, improving energy efficiency. These innovations have contributed to the reduction of carbon emissions and greater integration of clean energies, advancing toward a more sustainable future.

Energy systems, once centralized, now require multidirectional electricity flows between distributed generators, the grid, and users, which increases management complexity. This phenomenon is accentuated by the growing interconnection between energy, transport, construction, and industry. For this reason, AI is positioned as a key tool to improve supply and demand forecasting, optimize the integration of renewable energies, and increase operational efficiency. For example, the use of machine learning models to predict wind production has allowed companies like Google to increase the accuracy of their forecasts, thereby improving the profitability of their wind farms and enabling the advance sale of energy at higher prices (EIA, 2024).

Furthermore, AI also plays a crucial role in preventing grid failures through predictive maintenance, which significantly reduces service interruptions. Companies like E.ON and Enel have implemented machine learning algorithms to detect problems in their infrastructures, anticipating failures before they occur. As digitalization and automation expand, AI promises to revolutionize grid management, facilitating demand response, improvingv electricity flow control, and offering more efficient consumer services (IEA, 2024).

IX. Transport and logistics

Autonomous transportation and route optimization are led by AI. Artificial intelligence has revolutionized the transportation sector, improving safety, reducing operational costs, and increasing efficiency in urban transport. Companies like Tesla and Waymo are at the forefront of developing autonomous vehicles, which allow for driving without human intervention, promising a future with fewer accidents and greater comfort. In the logistics field, companies like UPS use AI to optimize routes, which reduces delivery time and costs, improving operational efficiency. Additionally, smart cities implement AI-based traffic management systems to regulate vehicle flows in real-time, alleviating congestion and improving urban mobility. These innovations are transforming transportation, making it safer, more economical, and more efficient.

The International Energy Agency (2024) estimates that almost one in five vehicles sold in 2023 was electric, representing double the proportion recorded just two years ago. Regarding the impact of artificial intelligence on transportation, the advancement towards electric vehicles (EVs) is a key component of the transition to a more sustainable transportation system, with AI playing a crucial role in optimizing charging infrastructure, managing EV fleets, and improving operational efficiency in this sector.

In Mexico, the manufacturing of transportation equipment is the economic activity with the highest added value, and the integration of AI in the automotive industry could further transform this key sector for the country's economy. As emerging technologies, such as AI, are incorporated into production processes, Mexico has the opportunity to consolidate itself as a leader in the manufacturing of next-generation vehicles, including electric (EV) and autonomous vehicles.

D. SOME REFLECTIONS ON THE FUTURE

Although artificial intelligence has a transformative impact, it poses ethical and social challenges, such as the potential loss of jobs, inequality in access to these

technologies, and dependence on centralized systems. We can address both topics in a complementary way. First, we will explore alternative currents to capitalism that seek to incorporate or resist technology, including AI, and then we will examine specific cases of industrial transformation with AI.

Alternative currents to capitalism and their relationship with AI. Some contemporary currents are rethinking the dominant economic models and, in particular, the relationship with disruptive technologies like AI. These perspectives seek to leverage these tools in an inclusive, ethical, and sustainable manner:

A. **Economy for the Common Good.** Proposed by Christian Felber, this current promotes an economic model based on values such as cooperation, sustainability, and equity. In this context, AI could be used to: optimize the distribution of essential resources, such as water, food, and energy; analyze social and environmental data to make more responsible political decisions; create transparent platforms for collective governance. Example of AI use: Algorithms for crop optimization in agricultural cooperatives that work under sustainability principles.

B. **Degrowth.** Degrowth questions the obsession with infinite economic growth, proposing to reduce consumption to preserve the planet's resources. AI can play a role in: designing more efficient energy systems that prioritize renewable energies; modeling and predicting the ecological impacts of political or industrial decisions; boosting local economies through decentralized platforms, reducing corporate intermediaries. Example of AI use: Simulations of natural resource consumption for communities seeking to minimize their ecological footprint.

C. **Circular Economy.** This model is based on reusing materials and products to minimize waste. AI helps in: designing more durable and recyclable products; creating reverse logistics systems that recover used materials; monitoring industrial processes to avoid resource waste. Example of AI use: Platforms like Winnow use AI to reduce food waste in restaurants, maximizing reuse.

D. **Post-capitalist proposals.** Movements like "solarpunk" or the "digital commons" imagine futures that combine advanced technology (like AI) with cooperative social structures:
   - **Solarpunk:** Integrates technologies like AI into self-sufficient communities, with an emphasis on sustainability and ecological design.
   - **Digital commons:** Open and collectively managed resources, such as AI software for research or education.

Example of AI use: Open-source projects like Hugging Face democratize access to AI tools, moving away from corporate control.

E. SPECIFIC CASES OF INDUSTRIAL TRANSFORMATION WITH AI

AI is also reshaping key industries under capitalist logic, but with the potential to change the way they operate.

Artificial intelligence is reshaping key industries under capitalist logic, with significant potential to transform their traditional ways of operation. In the field of autonomous transport and logistics, companies like Tesla and Waymo are leading the development of driverless vehicles, revolutionizing the sector by reducing operational costs and eliminating the need for human drivers, which also generates a positive environmental impact through route optimization and the consequent reduction of carbon emissions.

In sustainable agriculture, companies like Blue River Technology, a subsidiary of John Deere, use artificial intelligence to identify and apply herbicides selectively, which reduces the use of chemicals and costs. In Africa, projects like Hello Tractor use AI to manage shared tractors, allowing small farmers to significantly improve their productivity.

The creation of cultural content is experiencing a democratization thanks to platforms like Runway ML, which allow artists to generate images and videos using AI, and Deepdub, which uses smart technology to translate and dub films into multiple languages, expanding cultural dissemination.

In the medical field and global accessibility, organizations like Butterfly Network are developing portable AI-powered ultrasound devices to perform diagnostics in remote areas. In India, smart tools translate medical content into local languages, facilitating access for rural communities.

Finally, in the energy sector, companies like Google DeepMind have achieved the prediction of wind generation, optimizing the use of renewable energy in electrical grids.

V. CONCLUSION AND FINAL REFLECTIONS

Artificial intelligence is at a decisive moment that could define the future of our societies. Today, its development is primarily driven by economic interests that seek to maximize profits and consolidate technological monopolies. However, not everything is written. AI also has the power to become a tool for building a more just and inclusive world. Of course, this change will not happen by magic. We need an ethical, political, and social framework that guides its implementation toward objectives that benefit the collective well-being, not just the interests of a few.

Solid ethical rules are needed that limit the abuses of large corporations and guarantee a responsible use of AI. These norms must address issues like privacy, transparency, and equity, preventing technological power from aggravating existing inequalities. Second, democratizing AI is essential. It cannot be an exclusive tool for those who have the resources; it must be within everyone's reach, regardless of their economic situation or where they live. This means reducing digital divides and promoting technological education in communities that are usually marginalized.

Finally, it is time to rethink the role of technology in our lives. Instead of perpetuating a model focused solely on economic growth, we need to build a future based on sustainability, social equity, and respect for the environment. AI can be a driver to achieve all this, but only if we make conscious decisions, with a long-term vision. So combining technological innovation with humanistic values is not just important; it is essential. In this way, we can ensure that technology is truly at the service of humanity, and not the other way around.

VI. BIBLIOGRAPHY

Argandoña, A. "Desafíos y oportunidades de la IA en el sistema financiero mexicano". En Transformación digital en América Latina, 73–98. Barcelona: IESE Business School, 2019.

Beckert, Jens. Imagined Futures: Functional Expectations and Capitalist Dynamics. Cambridge, Massachusetts: Harvard University Press, 2016.

Crespo Garay, Cristina. "Un método basado en inteligencia artificial detecta el cáncer del futuro". National Geographic Latinoamérica, 10 de abril de 2023. https://www.nationalgeographicla.com/ciencia/2023/04/un-metodo-basado-en-inteligencia-artificial-detecta-el-cancer-del-futuro.

Endeavor México. La Era de la IA en México. 2024. https://mexico.endeavor.org/2024-ia/.

Expansión Digital. "Científicos de Google usan la IA y ganan el Premio Nobel de Química 2024". Expansión, 10 de octubre de 2024. https://expansion.mx/mundo/2024/10/09/premio-nobel-de-quimica-2024.

Forbes Staff. "Inteligencia artificial, ¿la vía para transformar la economía mexicana?" Forbes México, 2024.

Franco, G. "Disney pone en marcha una iniciativa con Inteligencia Artificial que transformaría Hollywood para siempre". Hipertextual, 25 de octubre de 2024. https://hipertextual.com/2024/10/disney-pone-en-marcha-iniciativa-inteligencia-artificial.

Gao, Robert X., Qing Chang, Jorge F. Arinez, Chengying Xu, y Jianjing Zhang. "Artificial Intelligence in Advanced Manufacturing: Current Status and Future Outlook". Journal of Manufacturing Science and Engineering 142, no. 11 (2024): 110804.

Gao, Robert X., Jörg Krüger, Marion Merklein, Hans-Christian Möhring, y József Váncza. "Artificial Intelligence in Manufacturing: State of the Art, Perspectives, and Future Directions". CIRP Annals 73, no. 2 (2024): 723–49. https://doi.org/10.1016/j.cirp.2024.04.101.

Gómez Mendoza, C., y Ezar, R. "Manufactura avanzada en la era de la inteligencia artificial: ¿Qué relevancia tienen los datos para esta industria?" EY México, 2024.

IBM. "¿Qué es la inteligencia artificial en las finanzas?" Por Matthew Finio y Amanda Downie. Última modificación: 8 de diciembre de 2023. Consultado el 19 de noviembre de 2024. https://www.ibm.com/es-es/topics/artificial-intelligence-finance.

IEA. Global EV Outlook 2024. París: IEA, 2024. https://www.iea.org/reports/global-ev-outlook-2024.

IEA. "Why AI and energy are the new power couple". Última modificación: 2023. Consultado el 19 de noviembre de 2024. https://www.iea.org/commentaries/why-ai-and-energy-are-the-new-power-couple.

Innguma. "Disney y la IA: Reinventando la innovación y vigilancia tecnológica". Última modificación: 21 de febrero de 2024. https://www.innguma.com/disney-ia-innovacion-vigilancia-tecnologica/.

Innovación Digital 360. "Desperdicio de alimentos: cómo reducirlo con IA y ciencia de datos". Última modificación: 14 de julio de 2022. https://www.innovaciondigital360.com/agrotech/desperdicio-de-alimentos-como-reducirlo-con-ia-y-ciencia-de-datos/.

Kemene, Eleni, Bart Valkhorf, y Thapelo Tladi. "AI and energy: Will AI help reduce emissions or increase demand? Here's what to know". World Economic Forum, 2024.

Keywords/-/Meta/revistaeyn-Com. "Disney forma una unidad para coordinar el uso de IA y realidad aumentada". Revista E&N, 2 de noviembre de 2024. https://www.revistaeyn.com/empresasymanagement/disney-forma-una-unidad-para-coordinar-el-uso-de-ia-y-realidad-aumentada-FE22405167.

Kreutzer, Ralf, y Marie Sirrenberg. Understanding Artificial Intelligence: Fundamentals, Use Cases and Methods for a Corporate AI Journey. Cham: Springer, 2019.

Palma, J., y Alvarado, R. "Impacto de la inteligencia artificial en México: Perspectivas éticas y sociales". Revista Misión Jurídica, 2015.

Plummer, S. "La IA en las finanzas: Revolucionar el futuro de la gestión financiera". El Experto en Anualidades, 11 de septiembre de 2024. Consultado el 19 de noviembre de 2024. https://www.datacamp.com/es/blog/ai-in-finance.

UNESCO. Propuesta de Agenda Nacional de la Inteligencia Artificial para México 2024-2030. 2024.

USDA. "Food Waste FAQs". Consultado el 1 de noviembre de 2024. https://www.usda.gov/foodwaste/faqs.

STRATEGY FOR THE ADOPTION OF AI IN MEXICAN INDUSTRY

CONTENT:

I. Introduction: The Evolution of Artificial Intelligence **II.** AI in Industry **III.** Mexico and Industrial AI. **IV.** Mexico in the Latin American Artificial Intelligence Index (ILIA) **V.** AI Opportunities in Industry. **VI.** Governance. **VII.** Conclusions. **VIII.** Bibliography

> JOSÉ LUIS SOLLEIRO
Senior Researcher at the Institute of Applied Sciences and Technology (ICAT) at UNAM and Advisor to the Institute for Industrial Development and Digital Transformation (INADI)

> ROSARIO CASTAÑÓN
Academic at the Institute of Applied Sciences and Technology (ICAT) at UNAM

> DAVID GUILLÉN
Consulting and Project Management Manager at Cambiotec, A.C.

I. INTRODUCTION: THE EVOLUTION OF ARTIFICIAL INTELLIGENCE.

Artificial intelligence (AI) has evolved rapidly since its inception in the 1950s, transforming into a key technology that impacts industrial development and various aspects of daily life.

In 1936, Alan Turing, a British mathematician and cryptographer, introduced the concept of the Turing machine, a theoretical idea that formalizes the concepts of algorithm and computation, which was fundamental to understanding how modern computers work[1]. In 1950, Turing proposed the famous "Turing Test" in his article "Computing Machinvery and Intelligence". This test evaluates whether a machine can exhibit intelligent behavior indistinguishable from that of a human being. Turing's proposal laid the groundwork for the development of AI by raising questions about machines' ability to think and reason.

The term "artificial intelligence" was coined by John McCarthy in 1956 during the Dartmouth Conference, where the foundations for the development of this discipline were laid. It was during this era that the first algorithms and programming languages, such as LISP[2], were developed, allowing machines to perform simple rule-based tasks.

In the 1960s, AI began to generate some applications with the development of programs such as ELIZA, a chatbot that simulated human conversations, as well as expert systems like MYCIN, designed to diagnose diseases, which demonstrated AI's potential in practical applications[3,4].

1 During World War II, Turing played a crucial role in breaking Nazi codes, especially those generated by the Enigma machine.

2 LISP, an acronym for list processing, is a programming language designed for easy manipulation of strings of data. It is one of the oldest programming languages still in use.

3 MYCIN was able to identify the bacteria causing the infection in patients and suggested antibiotics and appropriate doses based on each patient's weight. It was also capable of detecting infectious blood diseases such as meningitis and bacteremia.

4 https://telefonicatech.com/blog/mycin-el-comienzo-de-la-inteligencia#:~:text=MYCIN%20era%20capaz%20de%20identificar,la%20meningitis%20y%20la%20bacteriemia accessed on November 20, 2024

Expert systems (software that emulates the reasoning and decision-making of a human expert in a specific area) boosted AI. One of the first was Dendral, created in 1965 by Edward Feigenbaum and other researchers at Stanford University, designed for the analysis of chemical compounds[5]. These systems use a specialized knowledge base and reasoning techniques to analyze information and provide solutions to specific problems, imitating the decision-making of a human expert. In the 1980s, expert systems evolved and were applied in various fields of economic activity.

Subsequently, in the 1990s, with the growth of computing power, it became possible to process more complex algorithms and simulate decisions. In a historic milestone, Deep Blue, a supercomputer developed by IBM, defeated world chess champion Garry Kasparov in an official match. This proved the advanced calculation and strategy capabilities that could be conferred upon machines, paving the way for identifying AI's potential in decision-making[6].

At the beginning of this century, the arrival of big data (the ability to process large volumes of data) revolutionized the field of AI. Machine learning algorithms began to evolve and consolidate, allowing machines to learn patterns and make predictions based on historical data. This advance has been fundamental for applications such as facial recognition and search engines.

Building on this foundation, the development of deep learning led AI to new, more complex applications. Using artificial neural networks with multiple layers, this technique has enabled significant advances in areas such as natural language processing (NLP), image recognition, and autonomous systems. Applications like Apple's Siri and the Alexa virtual assistant are clear examples of the impact of deep learning.

In 2012, the introduction of AlexNet, a deep neural network that won the ImageNet competition, marked a significant advance in visual recognition. This model demonstrated that neural networks[7] could outperform traditional methods in complex tasks such as image classification.

In 2020, DeepMind presented AlphaFold, a system capable of predicting the three-dimensional structure of proteins with high accuracy from their amino acid sequence. This advance has significant implications for molecular biology and drug development.

5 https://www.lifeder.com/sistemas-expertos/ accessed on November 18, 2024

6 https://academiaready.es/los-hitos-mas-importantes-en-la-historia-de-la-inteligencia-artificial/ accessed on November 18, 2024

7 A neural network is a computational model inspired by the functioning of the human brain. It is composed of interconnected units called artificial neurons that process information in a similar way to biological neurons.

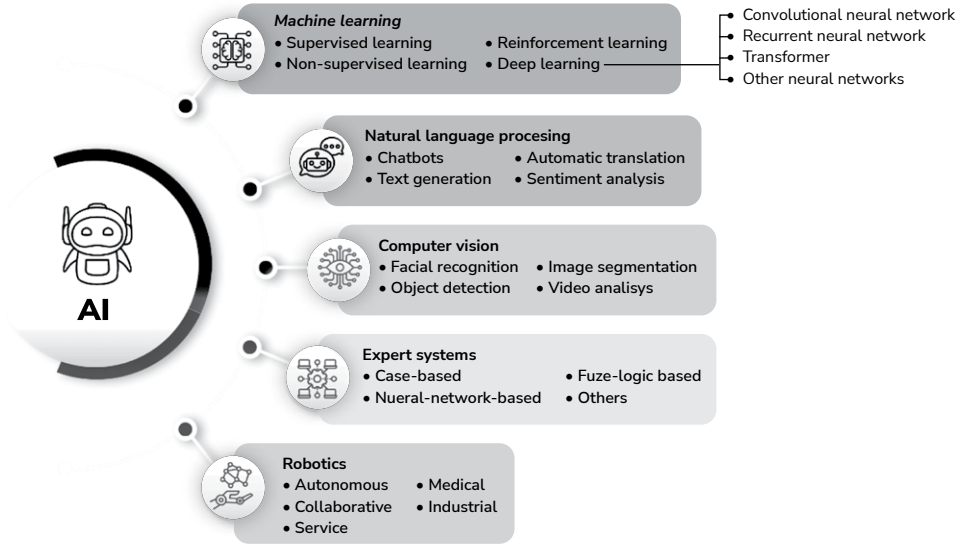
On the other hand, generative artificial intelligence has recently gained popularity, with tools like ChatGPT that can create coherent and relevant text from simple prompts. This type of AI is being used in various industries, from marketing to creative development. The exponential growth of generative AI has raised concerns about its ethical use and various risks, revealing the urgent need to establish regulations that guarantee transparency and equity in AI systems to ensure their responsible development, without limiting innovation.

AI encompasses a wide variety of technologies, making its taxonomy complex and varied. Historically, it has been classified according to various criteria, including:

- **Level of complexity:** reactive and limited AI.
- **Problem-solving capacity or intelligence level:** narrow, general, and superintelligence AI.
- **Rationality and mode of operation:** systems that act like humans, think like humans, or act rationally.

In addition to these classifications, Figure 1 shows the main disciplines related to AI.

FIGURE 1 Main disciplines related to Artificial Intelligence



SOURCE: Own elaboration

In 2024, artificial intelligence continued its expansion, with multinational companies being the ones that have most leveraged its potential (Oldemeyer, 2024), (Sections, 2021). According to the survey by Szedlak (2021), nearly 90% of SMEs still do not have AI applications in their operations. Due to the recent nature of different applications, there is still a significant gap between the application of technology for the development of concrete solutions and its implementation in companies.

II. AI IN INDUSTRY

Currently, terms such as lean manufacturing, advanced manufacturing, and Industry 4.0 are well known in the industrial environment. According to Lee (2020), all these techniques have been designed to efficiently solve known problems, and AI has been added to them, which uses various machine learning methods to integrate different types of data, allowing the extraction of invisible data, identification of relationships, and optimization of systems, thus avoiding (hidden or unknown) problems that have not yet occurred.

Industrial AI is a new emerging field that is expanding at a very rapid pace, thanks to the confluence of multiple sensors integrated into various products, the increasing availability of data (particularly digitized data), smarter algorithms, and the availability of processors and storage devices for large volumes of data.

According to the above, it is important to keep in mind that AI is a cognitive science that mainly includes six fields: natural language processing; computer vision; cognition and reasoning; game theory and ethics; and machine learning and robotics.

Traditional AI experts have deployed their great algorithmic capacity, but this is not enough for industrial applications, as it is difficult to solve problems of industrial systems in the absence of experts and knowledge of the business environment in various sectors. Therefore, the biggest challenge of industrial AI is to transform human-centered algorithmic thinking into the understanding of engineering systems.

The development of intelligent manufacturing systems and the rapid evolution of the Internet have shown a trend where the concepts of "localized intelligence" and "interconnected intelligence" go hand in hand in the context of industrial systems, which are evolving towards intelligent applications. In this sense, one of the biggest challenges for industrial AI does not lie in the technology itself, but in how to apply it to generate value. It is about using AI to solve problems such as the following:

- **Manufacturing quality,** including equipment, systems, and processes, as well as industrial organization and personnel capacity building. With AI models, early and automated detection of possible defects during production can be achieved using real-time data, for example in assembly and machining. Emerging applications include automated visual inspection using deep learning methods (Ojer, 2020), defect prediction (in line with the zero-defect manufacturing paradigm) (Peres, 2019), and online quality prediction (Schmitt, 2018).
- **Standardization of workflows,** which seeks to optimize and standardize tasks and processes within an organization, leveraging AI capabilities to automate and improve efficiency.
- **Predictive maintenance** using algorithms to analyze real-time data from sensors installed in machinery and equipment, which allows predicting failures before they occur and scheduling timely maintenance.
- **Supply chain** optimization through advanced data analysis, to adjust inventory levels and improve logistics.

In the words of Lee (2020, p. 9), the objective of the future Industrial Intelligence system is "to create an industrial environment that achieves the goals of zero accidents, zero pollution, zero waste, zero defects, and zero downtime".

CAPABILITIES REQUIRED FOR INDUSTRIAL AI

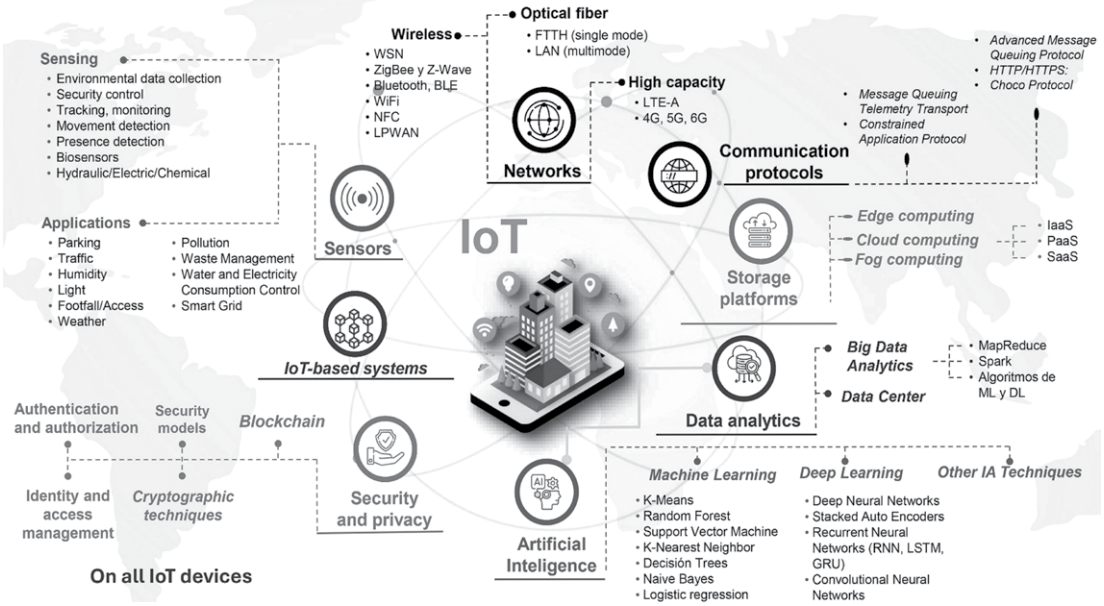
The development of industrial AI requires a series of technological capabilities that make its effective implementation and integration into existing processes viable. The main ones are detailed below:

- **Data infrastructure,** which implies systems that allow continuous data collection from sensors, machines, and processes. This requires robust, accurate, and updated databases, as well as the capacity to store large volumes of information.
- **Advanced data analysis tools.** Statistical analysis software and big data platforms are required to process and analyze large datasets in real-time. Similarly, the ability to develop and apply machine learning algorithms that allow machines to learn and adapt from data to make autonomous decisions is needed.
- **Automation capabilities** through systems to control and monitor processes, such as SCADA (Supervisory Control and Data Acquisition), LiDAR (Light Detection and Ranging), and PLC (Programmable Logic Controllers) systems.

- **Advanced robotics** through collaborative robots capable of working with humans and autonomously. Currently, the integration of artificial intelligence (AI) and Internet of Things (IoT) technologies is taking industrial robots to new levels of autonomy and adaptability. These advances allow robots to learn from their environment and optimize their operations, resulting in greater efficiency in manufacturing processes (Schwab, 2017).
- **Internet of Things (IoT).** The Internet of Things (IoT) describes the network of physical objects ("things") that are embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the Internet. These devices range from common household objects to sophisticated industrial tools[8]. Connectivity between devices is fundamental for data collection and exchange to generate. Figure 2 shows the infrastructure components for IoT operation.
- **5G networks** to improve the ability to transmit large volumes of data quickly and with low latency, which is crucial for industrial applications that require immediate response, such as automated systems and industrial robots.
- **Cybersecurity and data protection** to implement robust regulations and measures to protect industrial data and critical infrastructures. Effective systems to identify threats and react to cyberattacks are essential to achieve reliable operation in a highly connected context where every contact point in a network is a risk element (Solleiro *et al.,* 2022).
- **Research and development capabilities** in areas such as neural networks, deep learning, reinforcement learning, statistical methods, intelligent control, cyber-physical systems, sensors, simulation, and cybersecurity. Figure 3 shows the main technologies protected by patents in industrial AI in the period 2014-2024. The major fields where inventions are concentrated are machine learning, data analysis, cloud computing, deep learning, control systems, and industrial robots.

8 https://www.oracle.com/mx/internet-of-things/what-is-iot/ accessed on August 15, 2023

FIGURE 2 The infrastructure elements for IoT operation



SOURCE: Own elaboration.

FIGURE 3 Areas of concentration of industrial AI inventions (2014-2024).



SOURCE: Own elaboration using Patbase.

As can be observed, the effective application of industrial AI depends on the development and integration of technological capabilities that include everything from a solid infrastructure for data collection, analysis, and use to advanced automation, cybersecurity, and research. Only based on these capabilities, it will be possible to successfully implement industrial AI in Mexico and achieve the expected benefits. The following sections will delve into each of these capabilities and analyze Mexico's current situation in relation to them, as well as proposed strategies for their strengthening and development.

III. MEXICO AND INDUSTRIAL AI

In the Mexican context, the adoption of AI not only presents opportunities to improve the productivity and competitiveness of its companies, but also a series of structural obstacles derived from a poorly articulated national innovation system and inadequate public policies that have not placed the development of the digital economy as one of its pillars.

According to the International Monetary Fund (IMF) AI Readiness Index, Mexico ranks 56th out of 174 countries, with a score of 0.53. This index evaluates several critical components, including digital infrastructure, human capital, innovation and economic integration, and regulation and ethics. Compared to other Latin American countries, Mexico ranks behind Chile (0.58) and Uruguay (0.55) in terms of readiness to adopt this technology[9].

INFRASTRUCTURE

In the digital infrastructure dimension, Mexico obtained a score of 0.13, reflecting significant deficiencies in internet access, broadband speed, and mobile coverage. Inadequate infrastructure is a serious limitation to the development of industrial AI, as applications in this field require high speed and low latency, which depend on a widespread 5G network.

According to Piedras (2024), in the first half of 2024, "the balance of the public budget allocated and executed in the ICT sector in Mexico reveals a contrasting picture: some institutions or programs received a significant boost, while others face cuts that could limit their reach". The Federal Electricity Commission (Comisión Federal de Electricidad) and PEMEX saw notable increases: CFE

9 https://www.eleconomista.com.mx/opinion/El-futuro-digital-de-Mexico-Oportunidades-y-retos-en-IA-20240718-0003.html

339.4% in telecommunications infrastructure services and 203.5% for PEMEX telecommunications services. In stark contrast, the Ministry of Infrastructure, Communications and Transportation saw a 19.8% reduction in that period, and the General Directorate @prende.mx (formerly the General Directorate of Television Education), a program of the Ministry of Public Education (SEP) focused on educational digitalization and connectivity in schools, saw a 15.1% reduction, which shows that the government's discourse does not coincide with its commitment to digital education. What's remarkable about these budget data is that they reveal what priorities have been established.

In the case of 5G deployment, Mexico has increased its coverage in urban areas, but not outside of them[10]. But the 5G infrastructure deficit is not its only challenge, as the high cost of spectrum has hampered the allocation process, further delaying the development of this network.

HUMAN CAPITAL FOR AI

Investment in education, especially in STEM (science, technology, engineering, and mathematics) fields, is insufficient. Greater training and retraining are needed for the current workforce. A study by the Mexican Institute for Competitiveness (IMCO, 2023) indicates that the lack of trained personnel is one of the main obstacles to fully adopting AI.

Despite growing demand, the supply of specialized professionals in areas such as machine learning, data processing, and algorithm development remains limited. This shortage jeopardizes the country's ability to take advantage of the benefits of artificial intelligence and exposes it to greater technological dependence on foreign markets.

CISCO (2024), based on a global survey, concludes that by 2024, only 27% of companies in Mexico expect to have trained personnel to face the challenges of artificial intelligence, a figure that represents a decrease compared to the previous year. Furthermore, a growing cultural resistance is detected at the organizational level, especially at the management and operational levels.

10 https://www.5gamericas.org/resources/deployments/ accessed on November 29, 2024

AI IN BUSINESS

Cisco's recent report on the 2024 Artificial Intelligence Readiness Index reveals that Mexican companies' ability to implement AI has stagnated[11]. The study, which analyzes 8,000 companies in 30 countries, classifies organizations into four levels of readiness: leaders, chasers, followers, and laggards. In Mexico, only 13% of companies are in the leader category, while 38% are at the chaser level, slightly above the global average.

Compared to 2023, there is a decrease in the percentage of companies that consider themselves leaders in the use of AI. While the number of chasers remains stable, the number of organizations classified at the lowest levels –followers and laggards– has increased, highlighting a decline in national readiness to face the challenges of artificial intelligence.

"According to the data, one of the biggest obstacles in Mexico is inadequate infrastructure, including deficiencies in computing, networks, and cybersecurity. Only 21% of companies have the necessary technological resources, a considerable drop from the 28% recorded in 2023. Furthermore, the problem of data management remains significant, with 82% of companies working with fragmented information in silos, which limits the quality and reliability of their AI models"[12].

INNOVATION

In this area, Mexico has regressed. Spending on research is low compared to international standards, reaching only 0.27% of GDP in 2023, down from around 0.5% six years ago[13]. The National Strategic Programs defined by CONAHCYT (National Council of Agricultural Research and Technology) during the administration of Andrés Manuel López Obrador did not prioritize research for the manufacturing industry, and sectoral and mixed funds were eliminated. Regarding the promotion of technological development in companies, support from the Innovation Incentive Program and those managed by the National Entrepreneurship Institute were also eliminated.

11 https://comentariodeldia.com/2024/11/25/mexico-retrocede-en-preparacion-para-la-ia-segun-informe-de-cisco/

12 https://comentariodeldia.com/2024/11/25/mexico-retrocede-en-preparacion-para-la-ia-segun-informe-de-cisco/

13 https://www.elfinanciero.com.mx/opinion/eugenio-gomez/2024/08/27/que-tan-bien-preparado-esta-mexico-para-adoptar-la-inteligencia-artificial/

In the AI Meetings Roadmap for Mexico (May and Rojas, 2024, p. 28), it is mentioned, in the area of research and development, that "the group of participants highlighted that Mexico's investment in R&D is lagging compared to other Latin American countries, which is reflected in a diminished research ecosystem, with fewer researchers, lower productivity, and a smaller number of registered articles and patents". May and Rojas (2024, p. 29) also emphasize that "the specific challenges faced by SMEs around R&D are enormous given that they often lack the necessary financial resources and internal research capacity. The group of experts recognized the absence, to date, of any initiative, fund, or program aimed at technological research and development specifically aimed at these companies or multisector collaboration. This aspect was underscored as fundamental, given that SMEs constitute more than 95% of the country's economy".

Under these circumstances, the country's and its companies' capacity for innovation is not only low but has also been diminished.

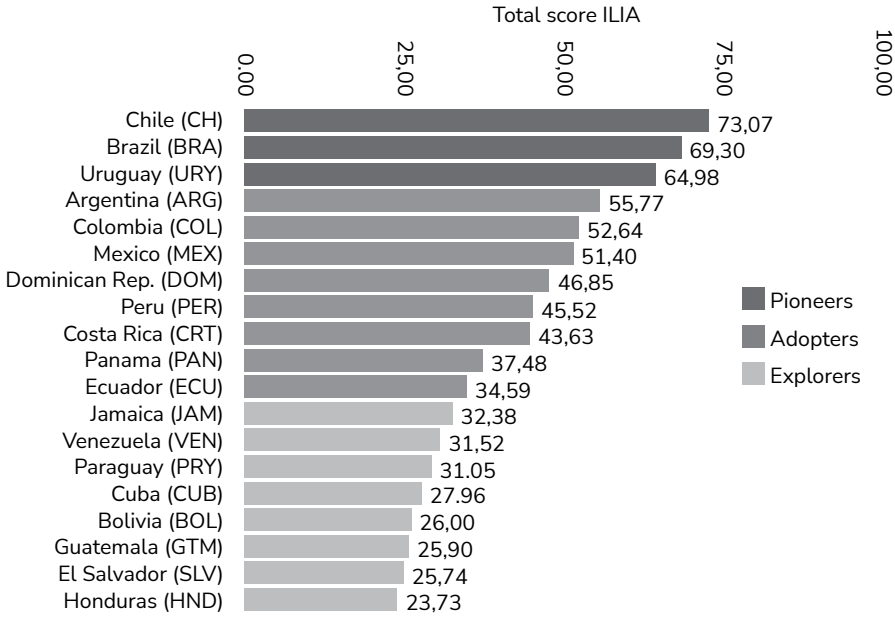
REGULATION AND ETHICS

The lack of a regulatory framework adapted to the needs of the digital environment presents an additional obstacle to the effective adoption of AI. Mexico needs to develop a flexible regulatory framework that protects citizens' rights while fostering innovation. This can facilitate faster and more effective adoption of AI-based technologies. May and Rojas (2024) recognize that Mexico needs to build a National AI Strategy based on ethical principles for technology, as well as designate an entity responsible for implementing the National AI Strategy.

IV. MEXICO IN THE LATIN AMERICAN ARTIFICIAL INTELLIGENCE INDEX (ILIA)

In 2024, the second edition of the ILIA was launched to offer a comprehensive and structured overview of the progress of AI in Latin America and the Caribbean, based on elements and variables that influence the development of AI ecosystems across three dimensions: enabling factors, research, development and adoption, and governance. Figure 4 illustrates that, even within the context of this region, Mexico has lagged, ranking only sixth in a ranking of AI "adopters" based on its level of maturity.

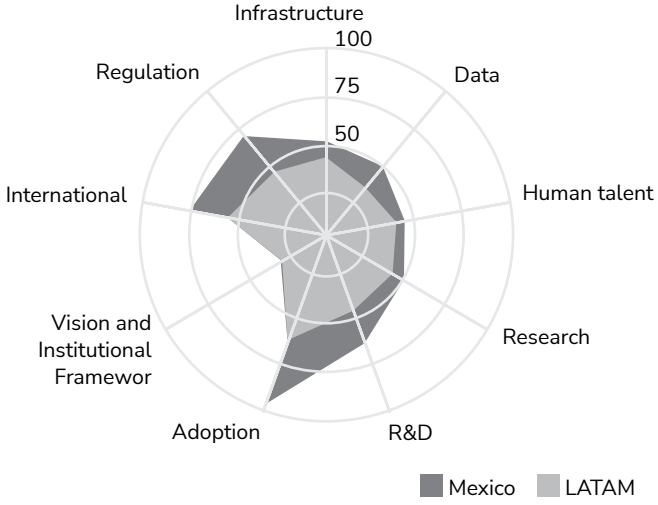
FIGURE 4 Latinoamerican AI Index 2024



SOURCE: Centro Nacional de Inteligencia Artificial (2024). https://indicelatam.cl/wp-content/uploads/2024/11/Ficha-Mexico.pdf

Figure 5 shows Mexico's position relative to the average for Latin American and Caribbean countries, revealing superior performance, albeit with a similar profile. However, Mexico's lack of vision and institutional framework in the field of AI is notable, placing it at a clear disadvantage.

FIGURE 5 Latin American AI Index 2024.Sub-dimensions of México and Latin America

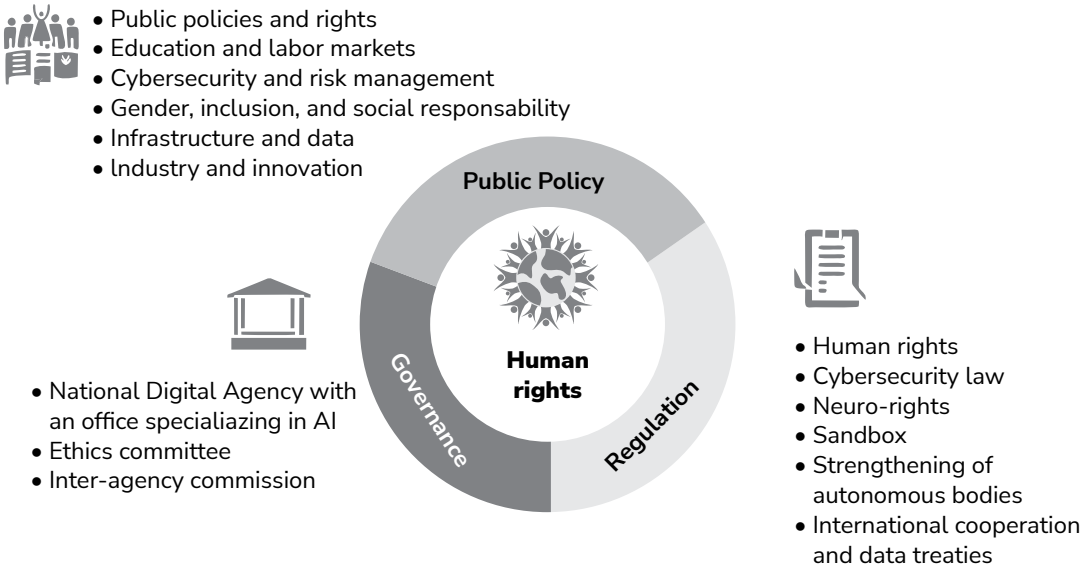


SOURCE: Centro Nacional de Inteligencia Artificial (2024). https://indicelatam.cl/wp-content/uploads/2024/11/Ficha-Mexico.pdf

In the document "Towards a national artificial intelligence strategy" (Ibarra *et al.,* 2024, p. 48), a general review of recent National Digital Strategies is presented, concluding that "the lack of a long-term vision of public policies and the lack of continuity of policies aligned with good practices or international standards in terms of digitalization are a serious problem and an obstacle for Mexico to achieve significant growth in AI. This situation is repeated every six years in the federal and local public administrations, and every 3 years in the municipal or mayoral administrations of Mexico City".

The National Alliance for Artificial Intelligence (Lagunes *et al.,* 2024, p. 12) mentions that, in Mexico, "the adoption of AI-based technologies shows an upward trend. The "Global AI Adoption Index 2022" indicates that 31% of companies in Mexico have actively implemented AI solutions (IBM Global AI Adoption Index 2022, 2022). Furthermore, a survey conducted by the Internet Association and Knowsy AI (2023) indicates that 55% of internet users in Mexico perceive AI as a useful tool, while 22% express concerns about its potential negative impacts. These figures reflect a duality present both in Mexico and globally: an enthusiasm for innovation and technological development, along with caution towards the ethical, economic, social, technological dependence, and cybersecurity risks that these advances could entail".

FIGURE 6 Axes of the ANIA proposal for building a national AI agenda 2024-2030.



SOURCE: Lagunes et alv. (2024)

Faced with this duality, various proposals have been launched in Mexico to define AI policies, all of which share a regulatory focus, relegating the promotion component. It is also notable that recent emphasis has been placed on generative AI and its ethical and responsible use. Such is the case of the National Alliance for Artificial Intelligence (ANIA), developed in the Senate of the Republic, whose proposal for a policy agenda is summarized in Figure 6. It is clear that the drive for the development and application of industrial AI has no specific focus.

V. AI OPPORTUNITIES IN INDUSTRY

A policy to promote industrial AI should focus on fostering its most promising applications, which include several innovative areas that seek to improve efficiency, decision-making, and operational capabilities. Among the most impactful applications in the short and medium term are expected to be the following:

PROCESS OPTIMIZATION IN MANUFACTURING

The application of industrial AI for process optimization seeks to increase the profitability of processes, while making them more sustainable and efficient.

These are real-time decision-support systems based on models based on historical data. Examples of tangible solutions being developed worldwide include:

- Energy consumption prediction and optimization (management) (Qin, 2018)
- Production efficiency (Liang, 2019)
- Demand forecasting (Chien, 2020)
- Optimization of the design and production of complex parts, rapid prototyping, and customization (mainly through 3D printing and manufacturing) (Liu, 2024)

QUALITY CONTROL AND ASSURANCE

With AI models, early and automated detection of potential defects during production using real-time data, for example in assembly and machining. Emerging applications include automated visual inspection using deep learning methods (Ojer, 2020), defect prediction (in line with the zero-defect manufacturing paradigm) (Peres, 2019), and online quality prediction (Schmitt, 2018).

HUMAN-ROBOT COLLABORATION

Currently, the integration of AI and IoT technologies is taking industrial robots to new levels of autonomy and adaptability. Cobots, or collaborative robots, are specifically designed to work safely and collaboratively with humans, representing a major shift from traditional industrial robots, which operated in isolation due to safety risks. Cobots are equipped with advanced sensors that allow them to detect the presence of humans and adjust their movements to avoid accidents. Support is provided for current human-centric tasks, either by improving operator well-being and safety or by making their tasks easier and more efficient. Applications of industrial AI include workforce training and task support (Lai, 2020), collaborative robotics (Bergamini, 2020), and ergonomics (Abobakr, 2019).

EXPERT SYSTEMS FOR DECISION-MAKING

The use of expert systems in industrial environments enables automated decision-making based on established rules and knowledge bases. These systems are applied in various areas such as risk assessment, fault diagnosis, and predictive maintenance (Yan, 2017).

LOGISTICS AND SUPPLY CHAIN OPTIMIZATION

Through automation and predictive analytics, the links in a supply chain can be seamlessly integrated, optimizing communication, standardization, and management of critical data (Oran, 2021). The use of AI helps overcome geographical and language barriers and generate highly efficient collaborative environments.

CHALLENGES FOR THE ADOPTION OF INDUSTRIAL AI IN MEXICO

Despite the potential benefits of industrial AI, its adoption in Mexico faces significant challenges that must be addressed to ensure a successful and competitive insertion into this technological phenomenon. These challenges include:

- **Lack of specialized human capital.** There is a shortage of professionals with skills in data science, machine learning, cybersecurity, and advanced robotics. This limits the ability of companies to develop and implement AI solutions.

- **Limited investment in R&D.** Investment in research and development in Mexico is still low compared to other countries, which hinders the generation of new knowledge and the development of proprietary AI technologies.
- **Insufficient infrastructure**. The lack of adequate digital infrastructure, especially in terms of 5G connectivity and cloud computing, limits the deployment of AI solutions that require high processing capacity and low latency.
- **Absence of clear public policies**. There is a need for a comprehensive national strategy that promotes the adoption of AI in industry, including incentives for innovation, training programs, and regulatory frameworks that facilitate its ethical and responsible use.
- **Resistance to change and lack of awareness**. Many companies, especially SMEs, are unaware of the benefits of AI or are reluctant to adopt new technologies due to fear of the unknown or lack of resources.
- **Data privacy and security concerns.** The implementation of AI solutions requires handling large volumes of data, which raises concerns about privacy and cybersecurity. It is essential to establish robust mechanisms to protect sensitive information.
- **High implementation costs.** The initial investment required to implement AI solutions can be high, which is a barrier for many companies, especially SMEs.

THE NEED FOR AN INDUSTRIAL AI POLICY

To address the opportunities outlined above, a clear policy with a long-term vision is required to help overcome the current shortcomings, summarized in Figure 7.

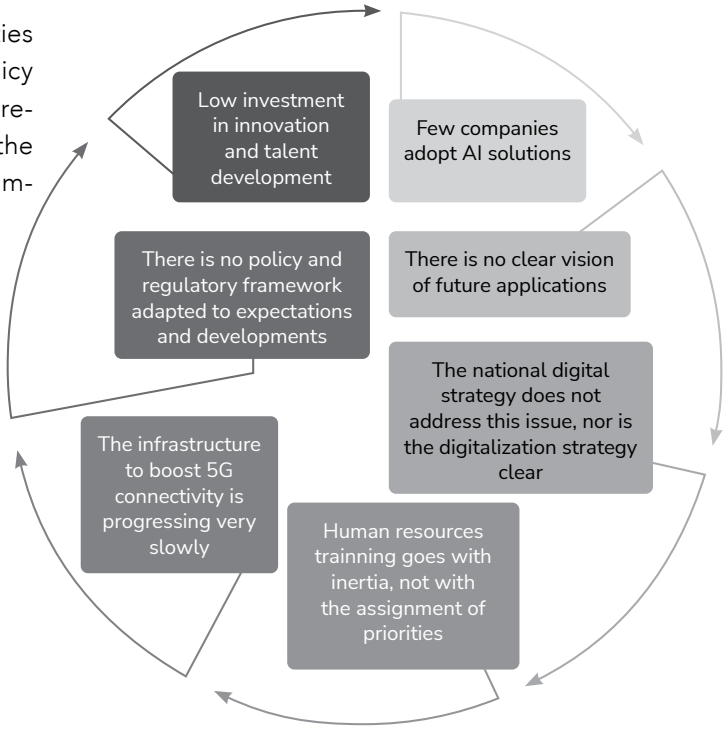


FIGURE 7 Obstacles to the development and adoption of industrial AI in Mexico.

SOURCE: Own elaboration

To overcome these obstacles, industrial AI policy must consider at least the following elements:

Financial support for companies that develop and adopt this technology, through direct economic incentives in the form of partial financing based on the merit of the business proposal, as well as tax incentives for investment in this field. SMEs should receive preferential treatment. It should be noted that the costs of technology and staff training are often high, which must be considered in the design of the support instrument.

Specific funds should be generated to promote joint research between research centers and companies[14]. These funds would be dedicated to financing innovative projects that offer practical solutions for the implementation of AI and Internet of Things solutions in companies. The integration of AI with other emerging technologies such as quantum computing and robotics promises to further expand its capabilities to solve complex problems and impact diverse economic sectors.

Build accessible high-speed, low-latency connectivity infrastructure in industrial development areas and facilitate access to internet services at affordable costs. To achieve these characteristics, it is essential to work within the framework of public-private partnerships with a long-term vision.

Likewise, there must be a reliable **supply of electrical power** to reliably operate the network of interconnected smart devices, as well as the data centers.

Regarding the latter, we agree with the proposal by Lagunes et al. (2024) regarding data infrastructure, as "it is necessary to promote investment in the development of Digital Public Infrastructure (DPI), including telecommunications networks, software, AI models, and open data structures designed under ethical principles". To achieve this, it is necessary to define a regulatory and promotional framework that facilitates the creation and operation of shared data centers, considering security, ethics, and energy efficiency standards. Operating a shared data center requires a significant investment in infrastructure, technology, personnel, and processes, as ensuring security, availability, and quality of service is essential to meet customer demands and maintain market competitiveness. This provides additional space for promoting public-private partnerships.

14 The proposed National AI Agenda for Mexico 2024-2030 suggests focusing investment funds on developing AI use cases in industries whose management teams are diverse in terms of professional experience and have equal numbers of women in leadership positions. Use cases will be able to utilize public digital infrastructure and controlled sandbox testing spaces, as well as access to national and international research and development networks.

It is necessary to invest in the creation and consolidation of data analysis platforms, with software for processing and analyzing them using advanced tools, so that companies can process large volumes of data generated by IoT devices, marketing operations, and logistics.

Cybersecurity and the protection of privacy and personal data are necessary elements for the responsible application of AI. Mexico faces very significant challenges in this area, as it is a frequent destination for cyberattacks. There is no national strategy or specific legislation, despite the proliferation of initiatives. In this regard, it is recommended to consider the proposals of the Guide for the Development of a National Cybersecurity Strategy promoted by the Federal Telecommunications Institute (Solleiro and Cuevas, 2024).

Increasing investments in technical and vocational education is essential to closing the skills gap needed to implement AI in industry. Promoting STEM-focused educational programs is crucial to preparing future generations. According to Lagunes et al. (2024), an ecosystem is required that includes key elements that enable the integration of AI-based technologies into educational environments (infrastructure, teacher professional development, provision of digital educational resources, as well as monitoring and evaluation mechanisms). This policy approach implies defining a long-term educational strategy that breaks with the inertia of the current educational system.

VI. GOVERNANCE

Governance refers to the set of processes, structures, and mechanisms used to make decisions in a participatory and inclusive manner to implement policies and manage resources within the policy framework. One of its objectives is to ensure that the AI strategy is integrated into the organizational design, considering the actors of the industrial ecosystem. This is a starting point for the design, development, and implementation of cyber governance. AI policy should be based on a participatory governance framework to foster multistakeholder dialogue. To this end, it is necessary to build analysis, consultation, and decision-making mechanisms specific to open and participatory governance systems[15]. Communication and continuous improvement mechanisms should include

15 Under the current administration, the Federal Telecommunications Institute ceased to be an autonomous body and was integrated into the Ministry of Economy. Although it has been assured that this action will not affect the activities currently underway, this claim cannot yet be verified.

ongoing dialogue with organizations from disciplines that are related to and impact the application of AI. It is important to adhere to the following principles:

- Participation of diverse stakeholders in the decision-making process to define, implement, and evaluate the national AI strategy, ensuring that diverse perspectives and needs are considered.
- Transparency so that decisions, actions, progress, and results are accessible to all stakeholders, promoting trust and accountability.
- Shared responsibility and distribution of responsibilities to ensure that individuals and entities carry out their assigned activities and are held accountable for their decisions and behaviors.
- Efficiency and effectiveness so that the resources allocated to implementing the Strategy are optimally used to achieve the established objectives, minimizing waste and maximizing results.
- Equity to avoid discrimination against groups, organizations, and technologies, promoting inclusion, technological neutrality, and equal opportunities.

In November 2024, Mexican President Claudia Sheinbaum Pardo presented the new Digital Transformation and Telecommunications Agency. Its main mission will be to digitize procedures currently carried out in person, allowing for the continuation of one of the main priorities of the Fourth Transformation: the fight against corruption. To simplify the procedures that citizens must complete, Llave MX, a new digital authentication mechanism that will streamline them, will be created. As can be seen, this Agency is oriented toward the government and, so far, does not consider taking on the task of generating an AI strategy that could address the issues associated with a digital industrial policy. In this context, the proposal by Lagunes et al. (2024, p. 16) in the sense that a main objective of the agenda for Mexico is "to incorporate public policy actions in Artificial Intelligence as part of the National Development Plan (PND), the National Digital Strategy (EDN), and the Plan for Compliance with the 2030 Sustainable Development Goals, defining a roadmap for action, responsible bodies, and performance indicators". It is not sustainable for Mexico to try to join this momentous technological wave without having a vision, strategy, and solid governance.

VII. CONCLUSIONS

Although Mexico shows high potential for adopting artificial intelligence in industry, primarily due to the operation of manufacturing sectors integrated into export markets and the possibilities of nearshoring, the country faces significant challenges that must be urgently addressed. Undoubtedly, the main one is defining a strategy for the development and broad adoption of AI in companies, particularly SMEs. This strategy must be supported by a public policy accompanied by effective instruments.

Among the instruments of AI policy, effective legislation and regulation stand out. The absence of a clear regulatory framework for the development and use of AI can generate legal and operational uncertainty for companies. This is especially relevant in Mexico, where there are currently no specific laws regulating the ethical and responsible use of artificial intelligence, as well as issues related to cybersecurity and privacy. The disappearance of autonomous organizations such as the National Institute of Transparency and Information Access (INAI) and the Federal Institute of Telecommunications (IFT) create uncertainty, so urgent action is needed to provide clarity on these matters.

Investment in digital infrastructure, strengthening education at all levels to ensure human talent in critical areas, and focusing on research, technological development, and innovation will be fundamental elements to improve Mexico's performance in this area.

It is necessary to be clear about the need for public and private investment so that the programs and projects to implement the strategy are viable. The government cannot face the challenges alone, so it must be open to collaboration with the private sector, with a long-term perspective. This requires an environment of legal certainty, which is currently not available.

With a strategic and collaborative approach, Mexico can not only reduce the existing gaps in international rankings but also ensure that the country's companies successfully incorporate artificial intelligence and achieve a level of mastery that allows them to extract maximum value from the associated technologies.

The effective adoption of AI in the industrial sector requires a comprehensive approach that combines advanced technology, efficient data management, staff training, and an organizational culture open to innovation. By developing these capabilities, companies can not only improve their operational efficiency but also position themselves competitively in an increasingly digitalized market.

Implementing AI projects can be costly, including expenses for hardware, software, and specialized technical personnel. This can be an impediment, especially for small and medium-sized businesses. Therefore, the catalytic role of public support for companies to adopt this technology is crucial.

It is good to keep in mind the five principles of industrial AI that Lee (2020) calls the five S's: Systematic, in the sense that the architecture should lead to full integration of AI with industrial systems; Speed, to enable the ability to design, model, validate, and deploy systems quickly to solve critical problems and respond to market needs; Streamline, to generate processes that ensure results and integrate into the industry in a simple and effective manner; Standards, to standardize data, analytics, formats, and industry operations; and Sustainable, to achieve consistent performance and repeated success.

VIII. BIBLIOGRAPHY

Abobakr. (2019). RGB-D ergonomic assessment system of adopted working postures. Appl. Ergonom., 75–88.

Asociación de Internet MX y Knowsy AI (2023). 19o Estudio sobre los hábitos de usuarios de internet en México 2023. Mayo. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://irp.cdn-website.com/81280eda/files/uploaded/19%20Estudio%20sobre%20los%20Hai-bitos%20de%20Usuarios%20de%20Internet%20en%20Mei-xico%202023%20.pptx.pdf

Bergamini. (2020). Deep learning-based method for vision-guided robotic grasping of unknown objects. Engonomic Informat, 13423.

Centro Nacional de Inteligencia Artificial (2024). Índice Latinoamericano de Inteligencia Artificial. https://indicelatam.cl/wp-content/uploads/2024/11/Ficha-Mexico.pdf

CISCO (2024). CISCO AI Readiness' Index. Hype Meets Reality. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.cisco.com/c/dam/m/en_us/solutions/ai/readiness-index/2024-m11/documents/cisco-ai-readiness-index.pdf

Chien. (2020). Deep reinforcement learning for selecting demand forecast models to empower industry 3.5 and an empirical study for a semiconductor component distributor. International Journal Production Research, 2784–2804.

Ibarra, Ernesto; De la Peña, Sissi; y Santoyo, Cristina; Panorama de la Inteligencia Artificial en México: hacia una Estrategia Nacional. 2024, disponible en: https://www.amcid.org/page/sandboxregulatoriomexico

IBM (2022). Global AI adoption Index 2022. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.eticanews.it/wp-content/uploads/2022/05/IBM-Global-AI-Adoption-Index-2022_FINAL.pdf

Instituto Mexicano para la Competitividad -IMCO- (2023). Inteligencia artificial: el futuro de la educación en México, Instituto Mexicano para la Competitividad, A.C., Ciudad de México

Lagunes A., Martínez Y., Cárdenas C., De la Peña, S., Mancilla, D., Xilotl R., Sánchez O., Moguel, A., Cárdenas J., (2024). "PROPUESTA DE AGENDA NACIONAL DE LA INTELIGENCIA ARTIFICIAL PARA MÉXICO (2024 - 2030)". Alianza Nacional de Inteligencia Artificial (ANIA). Disponible en: https://www.ania.org.mx/

Lai. (2020). Smart augmented reality instructional system for mechanical assembly towards worker-centered intelligent manufacturing. Journal Manufacture System, 69–81.

Lee, J. (2020). Industrial AI. Applications with Sustainable Performance, Springer, doi.org/10.1007/978-981-15-2144-7

Liang. (2019). Fog computing and convolutional neural network enabled prognosis for machining process optimization convolutional neural network enabled prognosis for machining process. Manuf. Syst, 32–42.

Liu, Q. (2024). Evaluating the Impact of 3D Printing Technology Innovations on Industrial Production Efficiency and Economic Growth in the Automotive Sector in China. Innovation Science and Technology.

May, C. y Rojas, D. (2024). Encuentros IA. Hoja de ruta para México. EON Institute y British Embassy in Mexico, Ciudad de México

Ojer. (2020). Real-time automatic optical system to assist operators in the assembling of electronic component. International Journal Manufacture Technology, 2261–2275. doi:10.1007/s00170-020-05125-z

Oldemeyer, L., Jede, A., Teuteberg, F. (2024). Investigation of artificial intelligence in SMEs: a systematic review of the state of the art and the main implementation challenges. Management Review Quarterly. https://doi.org/10.1007/s11301-024-00405-4

Oran. (2021). AI - Robotic Applications in Logistics Industry and Savings Calculation. Journal Of Organizational Behavior Research. doi:10.51847/juxqmvcvqf

Peres. (2019). Multistage quality control using machine learning in the automotive industry. IEEE Access, 79908–79916.

Piedras, E. (2024). Presupuesto TIC en la Primera Mitad de 2024: Asignaciones y Ejecución. The Competitive Intelligence Unit, Ciudad de México, 31 de octubre de 2024

Qin, J. (2018). Multi-so–850urce data analytics for AM energy consumption prediction. Adv. Eng. Informat, 840.

Schmitt. (2018). Implementation of cyber physical production systems for quality prediction and operation control in metal casting. Sensors, 1428.

Schwab, K. (2017). The Fourth Industrial Revolution. Crown Publishing Group, New York.

Sections. (2021). Striving for excellence in AI implementation: AI maturity model framework and preliminary research results. 363–376. https://doi.org/10.17270/J.LOG.2019.354

Solleiro, J. L., Castañón, R., Guillén, A. D., Hernández, T. Y., Solís, N. (2022). Boletín de vigilancia tecnológica en ciberseguridad 1. Instituto de Ciencias Aplicadas y Tecnología, UNAM. Descargable en https://www.icat.unam.mx/wp-content/uploads/2022/09/Vigilancia_Tecnologica_en_Ciberseguridad_Boletin.pdf

Solleiro, J.L. y Cuevas, J.L. (coord.) (2024). Guía para desarrollar una estrategia de ciberseguridad nacional, Instituto Federal de Telecomunicaciones, Ciudad de México

Szedlak, C., Leyendecker, B., Reinemann, H., Kschischo, M., Pötters, P. (2021). Risks and Benefits of Artificial Intelligence in Small and Medium Sized Enterprises. Proceedings of the International Conference on Industrial Engineering and Operations Management. Roma, Italua. Agosto 2-5. Disponible en chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://ieomsociety.org/proceedings/2021rome/175.pdf

Yan. (2017). Industrial big data in an industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance,'. IEEE Access, 23484–23491.

ARTIFICIAL INTELLIGENCE AND THE FOUNDATIONS OF A NEW DIGITAL ECONOMY

CONTENT:

I. Introduction: The Shift in the Productive Model in the Age of Artificial Intelligence **II.** The Effects of AI on the Future of Work **III.** Rising Market Concentration Through Digitalization and AI **IV.** Accumulation of Technological Capabilities and Their Impact on Development Paths **V.** The Advancement of Artificial Intelligence and Sustainability **VI.** Elements for an Economic Agenda on Artificial Intelligence **VII.** Concluding Remarks **VIII.** Bibliography

> JULEN BERASALUCE IZA

Professor-Researcher at the Center for Economic Studies, El Colegio de México (COLMEX).
Member of the Institute for Industrial Development and Digital Transformation (INADI)

"You insist that there are things machines cannot do.
If you tell me exactly what those things are,
I will always be able to build a machine that does exactly that".
JOHN VON NEUMANN, MANIAC

I. INTRODUCTION: THE SHIFT IN THE PRODUCTIVE MODEL IN THE AGE OF ARTIFICIAL INTELLIGENCE

Over the last several decades, we have witnessed a technological revolution fueled by the rapid expansion of digital technologies, whose effects extend across every aspect of socioeconomic life, including social interaction, information exchange, leisure, productivity, and many other areas. Labels such as *Industry 4.0* (Germany), *Industry 5.0* (European Commission), and *Society 5.0* (Japan) highlight different frameworks for envisioning new social paradigms in which digital technologies take central stage. By combining the analysis of emerging technologies with foresight exercises about their future trajectory, many countries –both developed and developing– are attempting to prepare for the transformations ahead. The scope of the socioeconomic changes tied to the Digital Revolution makes it essential to adopt a prospective approach that weighs not only the benefits but also the risks these technologies entail.

Among the most disruptive digital technologies, artificial intelligence (AI) clearly stands out. While its origin can be traced back to the earliest stages of digitalization, in recent years AI has exhibited exponential growth in its capabilities. As early as the mid-twentieth century, Alan Turing (1950), the father of digitalization, speculated about the possibilities of designing computer programs capable of learning through systems of variation and feedback that could guide their development in desired directions. Since then, a series of technological breakthroughs has expanded AI's functionality: long short-term memory networks (Schmidhuber & Hochreiter, 1997), convolutional neural networks (Krizhevsky *et al.,* 2012), generative adversarial networks (Goodfellow

et al., 2014), transformers (Vaswani *et al.*, 2017), among many others. As we will see, these advances were not achieved in isolation.

For the purposes of this discussion and given the absence of a universally accepted definition of AI, I will adopt the following working definition: AI refers to systems that, without the active intervention of any biological entity (in our case, humans), make decisions that adapt to a changing environment or accumulated experience. Two aspects of this definition are worth emphasizing. First, AI cannot be reduced to algorithms alone; its effective operation depends on an integrated system that also includes interaction environments, data, and computer power. Therefore, a proper evaluation of AI must consider this broader ecosystem. Second, AI systems are adaptive, imitating one of nature's most fundamental capacities. They evolve in response to environmental conditions and feedback from prior performance. This makes them inherently dynamic and difficult to analyze as fixed mechanisms, compelling us to consider their general features and capacities while also acknowledging that their continuous, accelerated transformation complicates forward-looking assessments.

Under this definition, AI encompasses a wide range of systems: artificial neural networks, machine learning, deep learning, expert systems, adversarial networks, large language models, and computer vision, among others. These technologies, with distinct applications across economic sectors, have demonstrated extraordinary disruptive potential, particularly in boosting productivity and automating tasks that until recently required medium- or even high-level cognitive skills from humans. Their effectiveness is closely tied to two factors: the availability of vast datasets to train the model, and the immense computer power needed to both train and operate. However, the most recent Chinese AI models, such as DeepSeek, have reduced the importance of this second factor.

The following sections examine several critical dimensions that should be central to any digital policy on AI's economic implications. As will be argued, the complexity of AI's socioeconomic effects requires active leadership by governments. This perspective stands in stark contrast to the current situation, in which governments have left significant gaps unaddressed, allowing Big Tech firms to shape AI's development largely according to their own interests. Entrusting the future of such disruptive technologies –and their profound socioeconomic consequences– to a handful of private actors is highly problematic, especially when some frame regulation as an obstacle to technological progress. History provides ample evidence of the risks associated with self-regulation and the dangers posed by the absence of active public involvement in the development

of new technologies. For this reason, the analysis that follows considers areas in which proactive public policy is urgently needed, including market concentration, labor impacts, inequalities in economic development, and the emergence of new strategic inputs.

II. THE EFFECTS OF AI ON THE FUTURE OF WORK

One of the most important implications of AI lies in its impact on labor demand. Following Korinek's (2024) classification of production schemes based on the scarcity and reproducibility of production factors, AI represents a rupture with the industrial production model, in which labor was naturally limited in its reproduction and complementary to capital. By contrast, AI, as a new form of capital, both tangible and intangible, has a much greater potential to substitute labor. While certain factors may slow or limit this trend, the possibility of a more generalized or sufficiently advanced AI replacing labor poses considerable challenges that warrant serious discussion of preventive and responsive economic policies. A declining demand for labor would exert downward pressure on wages, reducing workers' incomes, which remain the primary source of income for most of the population, and generate far-reaching socioeconomic consequences.

Focusing on substitution effects, Acemoglu and Restrepo (2018) argue that in activities where AI can replace human labor, productivity gains will drive adoption, resulting in output growth that outpaces wage growth. As capital per worker rises dvue to productivity improvements from automation, labor's share of total output will fall. The additional demand generated by higher productivity should, theoretically, expand the production of labor-intensive goods and services. However, this effect would depend on the absorptive capacity of existing and newly created sectors. Assessing that capacity requires considering whether shifts in consumer preferences might increase the value of human labor, the extent to which the state fosters demand for such goods and services, and the broader institutional framework that shapes wage pre- and post-distribution.

Current estimates indicate that a substantial number of occupations - especially those requiring medium to high-level cognitive skills- are vulnerable to AI, particularly large language models (LLMs). Eloundou et al. (2023) estimate that in the US, 19% of occupations could see more than half of their tasks affected by AI, while 80% could see at least 10% of their tasks impacted. It is worth mentioning that tasks directly related to digital technologies are not exempt from this menace. For instance, programming is increasingly shaped by tools

such as GitHub Copilot, which auto-complete and perform simplified coding tasks, notably reducing the demand for programmers. Furthermore, the tasks affected are becoming increasingly complex. Strikingly, the occupations most exposed to AI tend to be higher-wage jobs, meaning that well-paid activities are among those at risk.

The concept of robotization extends beyond simple substitution of labor and is closely tied to AI. The very word "robot" originates from the Czech term "robota", meaning forced labor, servitude, or drudgery, introduced by Karel Čapek in his science fiction play R.U.R. (Rossum's Universal Robots). This etymology adds an important dimension to our understanding of robotization, one that already has tangible applications today. Just as the mechanization of labor during industrialization manifested in the control of workers on the assembly line –a feature of Taylorism– the use of AI-driven hyper-surveillance of employees can be seen as a kind of "Taylorism on steroids".

In the same way that Taylorism was designed as a rationalization of production processes in the name of efficiency, AI monitoring tools subject workers to strict compliance with predetermined goals under continuous surveillance. These goals may include delivery deadlines, customer satisfaction ratings for every interaction, or detailed tracking of how workers engage with computer systems.

Workplace surveillance can reach dystopian extremes, a concern that already worries regulators. For instance, the *European AI Act* prohibits the use of emotion-recognition technologies in workplaces. Despite doubts about the universal scientific validity of linking facial expressions to emotions, tools for this purpose are already being developed at prestigious institutions like MIT and the University of California, and are being deployed commercially by firms such as Human and HireVue in job interviews (see Crawford, 2021). Moreover, biases in AI training data can amplify discrimination based on gender, ethnicity, or culture. Workers, in this context, are not only expected to perform adequately but also to show a facial expression that AI systems classify as pleasant or cooperative. A 21st-century remake of *Modern Times* could well feature an artificially animated Charles Chaplin, alienated not only by the optimized timing and minimization of tasks on an AI-driven production line but also compelled to smile convincingly so that the algorithm recorded his contribution as positive for the workplace environment. This raises ethical concerns about whether emotions should be treated as private and, therefore, be off-limits for performance evaluation.

Although AI can indeed provide valuable tools for measuring productivity, the use of such metrics can evolve in different directions. Unions, for example, could utilize more comprehensive productivity data to enhance their bargaining power

in collective bargaining. Yet, the asymmetry of power in the workplace suggests that employers are far more likely to harness these tools to their advantage. Comprehensive productivity tracking could individualize worker performance to such an extent that it undermined the pillars of collective bargaining.

Certain practices must be avoided, including collecting data outside working hours or off-site, using data in ways that may lead to discrimination, invading workers' privacy, or deploying AI to anticipate and suppress labor rights activities. For example, intensive data collection could identify the most active labor organizers, enabling employers to target them for dismissal. Given the inherent asymmetry in labor relations, worker consent cannot be treated as a sufficient safeguard for legitimizing surveillance practices in the workplace.

Finally, new digital platform-based business models raise pressing questions about the nature of labor relations and the definition of work itself. Traditionally, the distinction between ownership of capital and provision of labor has been a central factor in defining the employer-worker divide. Yet today, even when workers supply part of the capital (for instance, vehicles in ride-hailing or delivery platforms), clear asymmetries and dependencies remain between those who control the platforms and their users. Platforms often function as tools for reshaping employment relationships in ways that obscure –but do not eliminate– the fundamental employer-employee imbalance.

III. RISING MARKET CONCENTRATION THROUGH DIGITALIZATION AND AI

The shift toward a more digitalized economy, increasingly reliant on AI, asks for a reevaluation of industrial organization policy, with a particular emphasis on competition. Within this new landscape, the role of *Big Tech* –the large technology firms capable of leveraging their dominant positions for their own interest– cannot be ignored. As will be outlined briefly, there are structural features of the digital economy that naturally foster concentration, and these dynamics are likely to be intensified by the introduction of AI. The national origin of these corporations is not a trivial factor, as it is tied to the different regulatory models (or lack thereof) under which they operate. Within Big Tech, the so-called "Big Five" –Alphabet (Google), Amazon, Apple, Meta, and Microsoft– stand out, all of which are headquartered in the US. Still, one cannot disregard major Chinese players such as Huawei, Alibaba, Tencent, or ByteDance.

Many of these companies operate as platforms, controlling either social networks or operating systems, and thereby exercising direct control over how

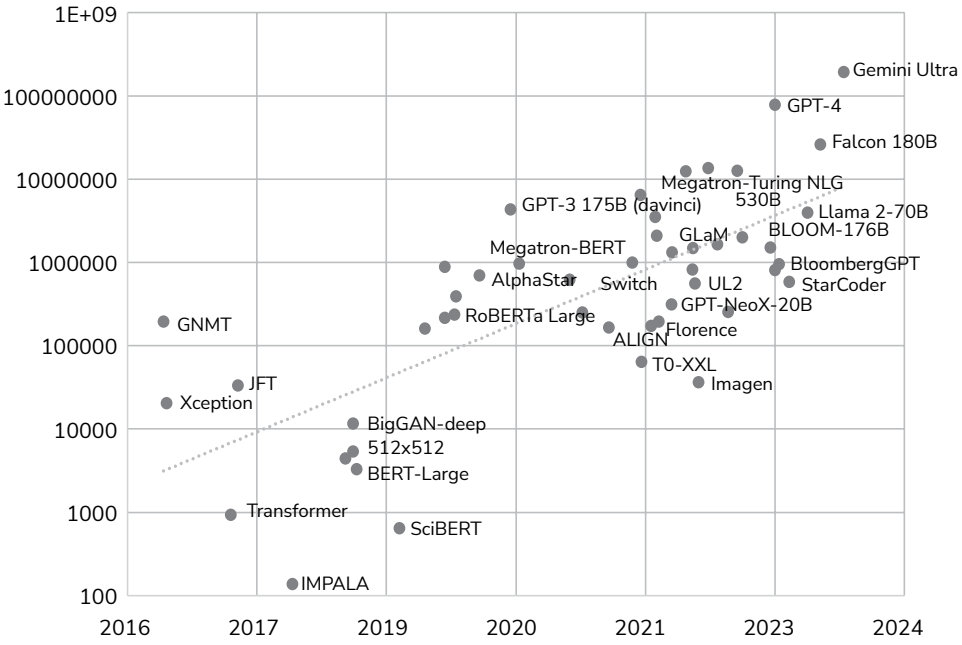
millions of users access other services, including those offered by competitors. In this sense, their role in the economy increasingly resembles that of a public utility, and their regulation may need to follow similar principles, particularly as they exhibit characteristics of natural monopolies (see Tirole, 2024). Unlike traditional natural monopolies, however, digital monopolies present a greater threat, as their reach is not confined to national or geographic boundaries. The ease with which digital technologies can operate globally, combined with liberalization in digital trade regulations and the new capacities of AI that reduce natural barriers such as language, enables Big Tech to extend its market dominance on a global scale.

AI may further reinforce this concentration of industrial power through at least two mechanisms: (1) the resource intensity of its three fundamental pillars and (2) its cumulative, path-dependent development process. Regarding the first point, AI systems rely on computational capacity, algorithms, and data. The complexity of recent LLMs demands ever greater computational power. While estimates vary widely due to firms' secrecy, the latest models are believed to contain over a trillion parameters, requiring tens of thousands of petaflop-days of computing power for training. The estimated training cost for Gemini Ultra alone approaches $200 million. These figures are the latest manifestations of an exponential trend in the costs of training advanced AI. Such computational requirements are far beyond the reach of smaller firms, constituting a formidable barrier to entry. Although the entry of the Chinese DeepSeek may seem an exception to this trend, since it was adapted to use modest-scale GPUs and the training cost of DeepSeek-V3 is estimated at around $5 million, this figure does not account for investments in research, infrastructure, data preparation, etc.

The availability of data is essential for building AI model capabilities, particularly during training. One of the key advantages Big Tech enjoys over potential competitors lies precisely in the vast amounts of user data it captures through interactions on its platforms. These interactions range from purchases and searches to seemingly trivial –and often uncontrolled– behaviors, such as how long a user lingers on a given post. Individually, such data points may have little value, but collectively they create value well above the marginal cost of extending their service to an additional user, which is close to zero.

Two main sources could ease these data constraints. First, expanded data capture through permanently connected devices, collectively referred to as the "Internet of Things". A growing ecosystem of wearables, cameras, microphones, and sensors –permanently online through 5G networks– could multiply the available data, even if such data are not publicly accessible. This dynamic only

FIGURE 1 Training Cost by Computational Usage (in US dollars, inflation-adjusted)



SOURCE: own elaboration with data from HAI (2024)

heightens the strategic value for firms to promote the widespread adoption of these devices. Second, synthetic data is generated by techniques such as generative adversarial networks, variational autoencoders, or advanced simulations. For example, Waymo's autonomous driving unit reported accumulating 20 million real-world driving miles, but multiplied that figure by a factor of 1,000 in simulated environments (Waymo, 2024). Simulations can be continuously enriched with real-world inputs while recombining them in novel ways to test algorithmic performance under varied conditions.

Cumulative advantages in AI development reinforce market concentration. Dominant firms can leverage existing market power to expand into adjacent or newly created sectors with high growth potential. For instance, synergies in data exploitation may enable digital giants to expand their reach into finance, where predictive capacity is a decisive competitive advantage. This dynamic raises competition concerns: if the continued use of a given AI system allows it to learn and adapt to the unique needs of a user, that user will face switching costs when adopting a rival product. One possible response would be to establish rules for the portability of individual training histories or user-specific data, giving end-users greater control.

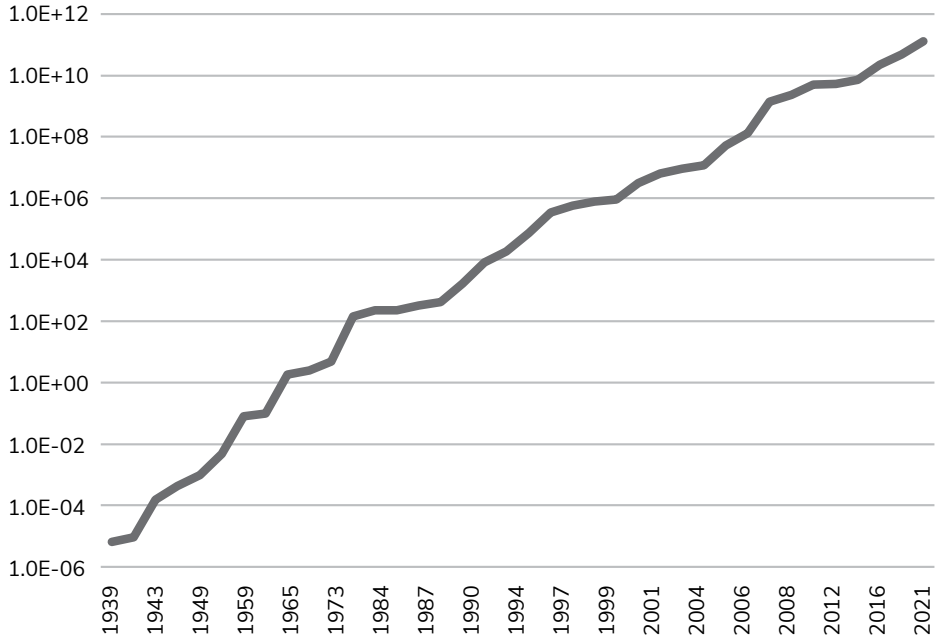
Competition oversight, therefore, must go beyond traditional measures such as consumer prices. Companies can build market dominance by accumulating data rather than lowering prices. Regulators should closely monitor data-sharing strategies between firms or between subsidiaries of the same conglomerate. User consent is often insufficient, as data collected for one app (e.g., location or health information) can be repurposed and monetized in ways that the user does not anticipate. Antitrust law and competition authorities should consider data-related effects when reviewing mergers and acquisitions. The recent antitrust measures issued against Google in the case brought by the U.S. Department of Justice show regulators' reluctance to curb concentration in the digital sector meaningfully.

Finally, AI's own development may accelerate technological change in ways that reinforce concentration. Experts increasingly discuss the medium-term prospects of artificial general intelligence (AGI), capable of autonomous learnvting across a wide variety of domains, or even the so-called technological singularity, where AI systems improve themselves in a recursive feedback loop. While such forecasts may underestimate obstacles or overstate feasibility, it is undeniable that AI represents a unique class of technology, one that can drive advances in other technologies, amplifying its disruptive potential.

The risks of market concentration in the digital economy and AI are profound. The stakes extend well beyond consumer welfare in terms of prices. Companies with such dominance can influence regulation itself -for example, whether social media firms should bear specific responsibilities for their potential role in shaping electoral outcomes. Similarly, their reluctance to disclose details about algorithmic architectures, training data, or model parameters hinders both scientific progress and the development of precautionary regulations. Due to the enormous scale of these players and their involvement in sensitive areas, ranging from surveillance to military applications, government actions are often deeply entangled with the strategies of leading AI firms.

Running counter to concentration trends is the exponential increase in computing power per constant dollar. This means that, for the same purchasing power, greater computational capacity can be achieved, rapidly lowering the cost of access to high-performance computing. If we combine this with algorithmic novelties, such as the refined Mixture-of-Experts architecture and the Multi-Head Latent Attention, which drastically reduce memory use, there could be opportunities for new players in these markets, whose dominant players are still being defined. Also, this could narrow the gap for developing countries, enabling them to gain access to advanced computational capacities at lower costs within a relatively short timeframe.

FIGURE 2 Computations per second per dollar at peak performance (2023 constant prices)



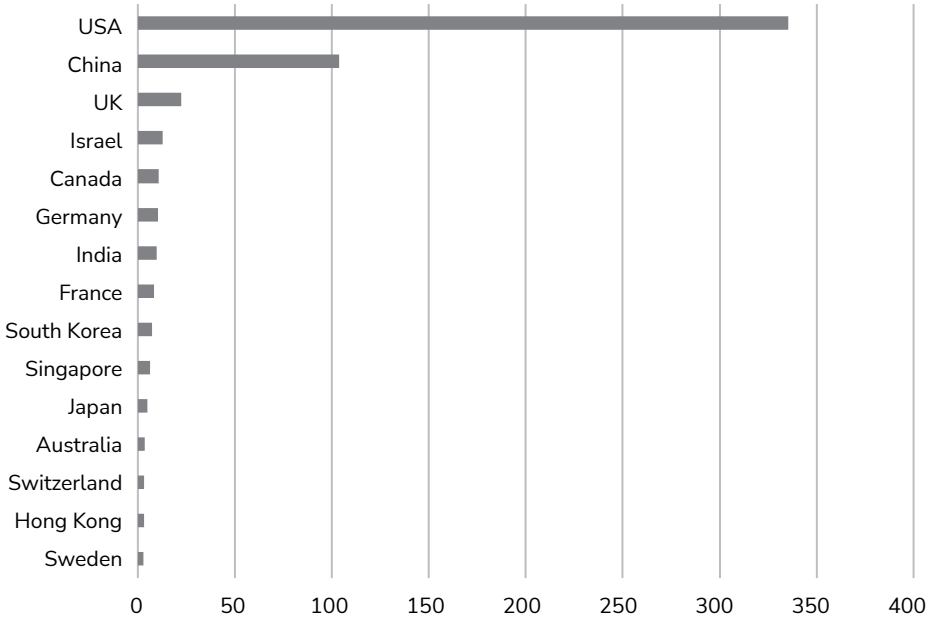
SOURCE: own elaboration with data from Kurzweil (2024, 165-166)

IV. ACCUMULATION OF TECHNOLOGICAL CAPABILITIES AND THEIR IMPACT ON DEVELOPMENT PATHS

Several of the elements discussed above, taken together, can significantly constrain the ability of developing countries to achieve technological convergence, thereby limiting their prospects for economic growth and broader socio-economic development. Technological development has a strongly cumulative dimension, and AI may reinforce this tendency. Because new knowledge and infrastructure are built upon prior capabilities, it is challenging to advance in one domain without already possessing the necessary foundations. These foundations extend beyond knowledge to include strategic inputs, know-how, and institutional experience. For instance, developing robust digital infrastructure is virtually impossible without parallel investments in energy infrastructure, given the enormous electricity demands of these facilities, as well as firms' growing requirements for green energy.

Unless new technologies emerge in entirely distinct, non-complementary fields –or the costs associated with earlier developments fall dramatically– the accelerated pace of technological change associated with AI may significantly widen the gap between countries, making convergence more difficult. In some cases, restricting access to powerful technologies can have positive effects, since it allows tighter control over tools with potentially harmful applications. For example, maintaining limits on access to advanced AI models capable of designing novel biological agents could reduce the risk of terrorist groups exploiting such capacities. At the same time, however, such restrictive access also curtails opportunities for beneficial civilian applications.

FIGURE 3 Cumulative Corporate Investment, 2013-2023 (billions of US$)

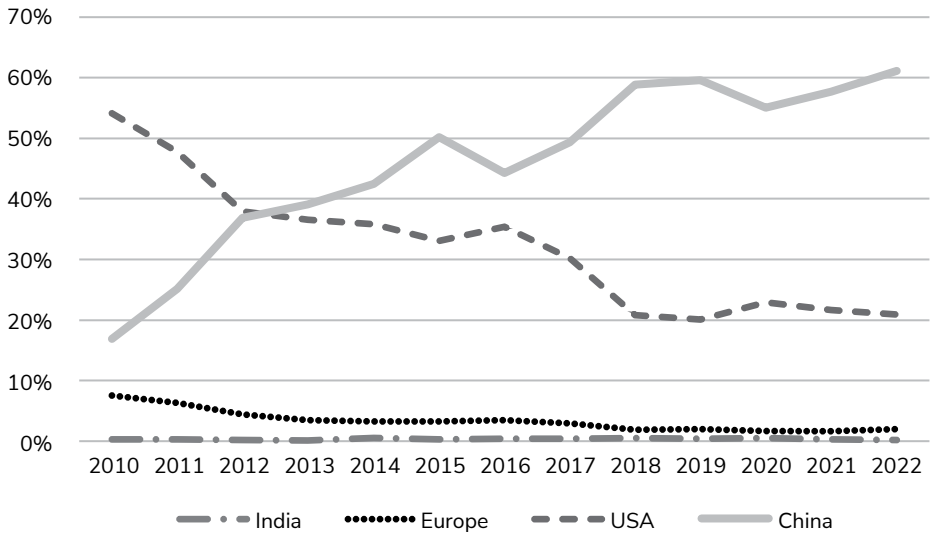


SOURCE: HAI (2024)

As previously noted, the latest AI models –characterized by greater capabilities– require increasingly large-scale investments, a dynamic that reinforces corporate concentration. In parallel, the clustering of major firms in the United States (predominantly) and China (to a lesser extent), as well as in other advanced economies, is mirrored by the concentration of corporate activity and private investment at the national level. Over the past decade, private investment has been overwhelmingly concentrated in the United States,

exceeding $300 billion in AI-related projects. China follows, with less than one-third of that amount, although it achieves notable outcomes relative to its investment levels. Beyond these two global hegemons, investment is more dispersed across other advanced economies. Europe, whether considered regionally or through the European Union as the most relevant political actor, emerges as the third actor, although its cumulative weight in private investment remains far behind that of the other two. By contrast, developing countries account for a marginal share, with India standing out as a notable exception. India's *Digital Public Infrastructure* (DPI) initiative is particularly relevant, as it offers an alternative model of digital development (including AI) that explicitly addresses the differentiated challenges faced by the Global South.

FIGURE 4 Granted AI patents as a share of the global total, by geographic area.



SOURCE: HAI (2024)

The trends on the geographic concentration of AI-specific patents provide useful insights into how technological concentration may evolve in the future. Two clear patterns emerge from the aggregated data. First, AI-related technological development is highly concentrated at the country level, and this concentration has, if anything, intensified in recent years. Whereas the U.S. and China together accounted for 71% of AI patents in 2010, their combined share rose to more than 82% by 2022. Europe is the region that has lost the most ground: in 2010, it held 7.54% of AI patents, but by 2022 that share had fallen to barely 2%. These relative

figures must be understood in the context of the exponential growth in the absolute number of AI patents; yet, the relative loss of dynamism is nonetheless significant. On the one hand, the EU has made considerable progress in regulation –through the Digital Services Act, the Digital Markets Act, and the AI Act– creating an asymmetric competition framework that constrains Big Tech and embeds human rights protections in AI development. On the other hand, we must not forget that large corporations are primarily driving technological breakthroughs in AI, and European regulation may reduce incentives to invest in the region.

Second, the shift in global leadership from the U.S. to China is striking. In 2010, the U.S. led with 54.13% of granted AI patents compared with China's 16.89%. By 2022, the positions had reversed: the U.S. had fallen to 20.90%, while China had surged to 61.13%. Needless to say, patent counts alone do not capture the full picture of technological development, as patents differ widely in their economic value and technological centrality. Nevertheless, this sharp reversal can also be found in several indicators that show a broader trend of shifting technological leadership across sectors. Adjusted for population, the U.S. still outpaced China in 2022, with 4.23 patents per 100,000 inhabitants, compared to China's 2.51. Yet even here, U.S. leadership is challenged: South Korea recorded 10.26 AI patents per 100,000 inhabitants, the highest figure worldwide. The trajectories suggest that while AI will likely remain a highly concentrated technological and economic field, its center of gravity is increasingly shifting toward East Asia.

Other regions of the Global South remain largely excluded from AI technological development. India, the most notable case, barely exceeded 0.5% of global AI patents in any given year, and in 2022 accounted for just 0.23%. Latin America, Southeast Asia, and Africa all register minimal representation, far below their economic weight, signaling an unfavorable outlook for regional equity in the digital economy over the coming years. A more optimistic countertrend lies in the recent expansion of open foundation (or base) models. Of the 149 foundation models released in 2023, nearly two-thirds were open-sourced, a reversal of the proportion just two years earlier. The growing prevalence of open-source foundation models may provide opportunities for countries in the Global South to collaborate and pursue a more inclusive path to technological development.

V. THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE AND SUSTAINABILITY

The implementation of AI has significant implications for one of the central objectives of global economic policy: mitigating climate change, particularly through the reduction of greenhouse gas emissions. AI brings both environmental costs and benefits, and ongoing research is needed to assess its net impact.

On the negative side, training and operating large AI models require vast amounts of energy, primarily to power the data centers and computational infrastructure on which they depend. While cloud computing offers efficiency and security advantages, it is highly energy-intensive, especially in terms of electricity consumption. Despite industry efforts to expand the use of renewable energy, the carbon footprint of the computing sector has been compared to that of the aviation industry. For example, Luccioni et al. (2022) estimate that training BLOOM, a 176-billion-parameter language model, emitted 24.7 metric tons of CO2-equivalent, based solely on its required computational power. Factoring in hardware production and other operational processes, the total footprint would roughly double. Even so, BLOOM's emissions remain far below those of GPT-3, a model of similar size, which is estimated to have generated around 502 metric tons of CO2-equivalent, nearly ten times more. Although training efficiency is continuously being improved, the continued growth in model size and complexity increases the computational power required, offsetting these gains. By 2020, Freitag et al. (2021) estimated that information and communications technologies accounted for between 2.1% and 3.9% of global greenhouse gas emissions. Given the limited transparency in corporate reporting, these figures are highly uncertain, and the share is likely to rise as AI tools –more energy-intensive than other digital applications– become more widely adopted. For instance, while a single web search produces roughly 0.2 grams of CO2-equivalent, a query to ChatGPT may generate between 50 and 90 times that amount (Vanderbauwhede, 2023). If AI-based search were to replace traditional methods on a large scale, the effect on total emissions would be substantial.

AI systems are also highly water-intensive. Recent estimates suggest that global AI demand could, in the near future, consume as much water as half of the United Kingdom's annual usage (Li *et al.,* 2023). Beyond operational needs, computing hardware requires extensive inputs of critical raw materials. Producing finished components may consume up to 400 times their weight in extracted resources. Many of these are rare earth minerals, often mined through environmentally destructive processes that involve large-scale land disruption and

notable contamination risks. At the end of their lifecycle, electronic components also generate hazardous waste, with elements such as mercury posing significant environmental challenges.

Mitigating AI's environmental impact will require targeted regulation and monitoring. For instance, the EU's *Artificial Intelligence Act* includes provisions to encourage the development of energy-efficient AI models (European Parliament, 2024). Among the required technical disclosures for providers of general-purpose AI systems are the energy used for model training and processing. Environmental safeguards and monitoring mechanisms are explicitly integrated into its regulatory framework. In the US, there have been attempts to regulate this dimension, such as the *Artificial Intelligence Environmental Impact Act*; however, since the recent arrival of Donald Trump in office, such attempts have been halted.

Balanced against these risks are potential environmental benefits. By substituting human labor in certain processes, AI may, in some cases, lower overall resource use for equivalent output. This raises a critical question: to what extent is AI more efficient per unit of economic activity? Intensifying AI deployment will inevitably carry environmental consequences, and we need an assessment that surpasses the analysis of relative efficiency to reconsider its sustainability.

VI. ELEMENTS FOR AN ECONOMIC AGENDA ON ARTIFICIAL INTELLIGENCE

The disruptive potential of AI –both as a distinct technology and as a driver of the broader digital economy– will inevitably reshape key aspects of economic systems and labor regimes worldwide. Given the scale of these transformations, a forward-looking policy agenda is essential. The following outlines several issues that such an agenda should address, with a particular emphasis on labor markets.

Technological progress associated with the Digital Revolution and AI offers enormous opportunities for productivity gains. Yet the profitability of certain AI-driven and platform-based business models often relies on problematic practices, including the circumvention of labor law through contracts of association, Taylorist forms of pervasive surveillance, or the extreme individualization of workers as a means of undermining collective bargaining. Formal and informal labor institutions remain critical for safeguarding workers' rights and ensuring fair wages. These, in turn, are foundational for sustaining stable democratic societies.

Regulatory policy –both in the form of legislation and through enforcement agencies– must remain vigilant to how AI and digital business models may create new forms of labor rights denial. Policymakers should resist uncritical adoption of Big Tech's narratives of "flexible work" and instead focus on applying long-standing international labor principles to new contexts. To cite one such principle, enshrined in the preamble of the International Labour Organization's Constitution: "Labor is not a commodity". Beyond recognizing that work does not belong in the category of commodities, this principle implies that work should not be treated as if it were one. On this basis, precautionary regulation is justified, and the evaluation tools often employed in regulatory sandboxes must be scrutinized, particularly given the political power and influence of Big Tech firms.

Perhaps the most pressing labor concern is the risk of technological unemployment, in J. M. Keynes's words. While some argue that previous technological shifts ultimately generated new sectors and jobs, such inductive reasoning may not be extended to the Digital Revolution's radically different conditions. Assuming that market forces will create new labor opportunities is dangerous if it justifies inaction. Even the most radical advocates of free markets should acknowledge that labor markets adjust far more slowly and with greater social costs than other markets. Halting AI development is neither practical nor desirable, given global competitive dynamics. Instead, the challenge is to manage the inevitable restructuring of work and the potential rise in inequality.

The first set of measures should be aimed at mitigating –not halting– labor substitution. Preserving human agency in key systems and technologies can be justified on ethical grounds, preventing full delegation of decision-making to machines. Aviation regulators worldwide, for example, mandate the presence of two pilots on commercial flights. Similar regulatory principles could be applied to AI-driven domains such as autonomous driving, healthcare, or diagnostics: AI can be integrated, but not wholly replace human oversight.

Still, even partial substitution can dramatically reduce labor demand. For instance, AI coding assistants have sharply lowered the hours of human labor required in software development. The result is fewer programming jobs, with remaining tasks concentrated among a smaller group of highly skilled workers. Such dynamics risk widening inequality, as Acemoglu and Restrepo (2022) argue in relation to automation. In the face of this, proposals such as Universal Basic Income (UBI) are gaining attention.

A UBI would guarantee all citizens a minimum income. Although designed to be universal and avoid disincentives for productive activity, true universalism

may be fiscally unfeasible, even in developed countries. Pilot programs, such as those in Canada and Finland, have typically targeted the unemployed or low-income households. Finland's trial, for instance, reduced stress associated with unemployment and income insecurity, without any negative effects on employability (Kangas *et al.,* 2021). Yet the fiscal burden remains a central barrier, one that is especially acute in developing countries with weaker tax capacity.

To finance UBI, some have proposed targeted taxes on AI technologies. The logic is straightforward: tax the technology that displaces jobs to internalize its social costs. But such a policy raises serious challenges. First, defining precisely which AI technologies fall under such a tax would be complex, because many productivity-enhancing technologies reduce labor demand. Second, the geographic mismatch between where AI is developed (primarily in advanced economies) and where jobs are displaced (globally) means that revenues would concentrate in the Global North. Finally, the unilateral adoption of AI taxes at the national level would create competitive disadvantages, discouraging countries from acting on their own. These obstacles suggest that any similar approach would require international coordination.

Beyond labor, AI amplifies long-standing concerns over market concentration. The inherent dynamics of the digital economy already favor dominant players, and AI intensifies this pressure. Developing countries, with limited regulatory capacity, may struggle even more when confronting global tech giants. As seen in cases ranging from EU antitrust actions to political disputes between platforms and national courts (e.g., Brazil vs. X/Twitter), concentrated corporate power can overwhelm even the strongest states. This highlights the importance of international cooperation in countering the influence of Big Tech. Moreover, competition policy must move beyond consumer-price frameworks to include dimensions such as political influence, institutional capture, and data control.

Environmental sustainability presents another critical agenda item. Precautionary regulation is warranted. Following the EU example, policymakers could require AI developers to disclose energy usage both during training and operation. Public procurement policies could then prioritize energy-efficient models. At the same time, national energy strategies should align the siting of new data centers with expansions in renewable energy capacity.

Currently, three regulatory models stand out globally. The European model, anchored in the *Digital Services Act*, the *Digital Markets Act*, and the *Artificial Intelligence Act*, emphasizes asymmetric regulation of dominant firms. It restricts certain applications outright –as we mentioned, banning emotion-recognition tools in the workplace and continuous facial recognition in public spaces– and

places human rights at the center of AI governance. While this approach may slow private-sector growth and Europe's digital competitiveness, it remains the most interventionist model.

The US model is more fragmented and market-driven, with regulation lagging behind technological deployment. Although bills such as the *AI in Government Act* and the *Algorithmic Accountability Act* have been proposed, none have passed at the federal level. The most important step was the issuance of Executive Order 14110 in 2023, which established broad principles for AI safety, innovation, and privacy. States have also begun acting independently; for example, California's *AI Transparency Act* requires generative AI systems to label machine-produced content.

The Chinese model features stronger state oversight, with the government explicitly prioritizing national security and public interest in its 2023 *Interim Measures for the Management of Generative AI Services*. These include privacy provisions, safety standards, and content controls to prevent politically or socially undesirable outputs.

Finally, while not an AI-specific model, the DPI offers an alternative framework for developing economies. DPI emphasizes open-source systems and interoperable standards that can be shared between the public and private sectors to build incremental digital capacity. It also includes mechanisms for managing personal data under transparent standards, potentially offering countries in the Global South a more inclusive and sustainable pathway for AI adoption.

VII. CONCLUDING REMARKS

The disruptive force of AI extends across a wide range of socioeconomic dimensions, making regulatory planning and state intervention essential. This chapter focused on a subset of economic dimensions, with particular attention to four areas that have been most visibly affected by AI and the digital economy in these early stages of their development. The most ardent critics of state intervention, including even precautionary regulation, argue that the intrinsic complexity of AI, coupled with its rapid acceleration, renders regulatory efforts clumsy at best and merely obstructive to technological progress at worst.

Yet, in the face of these structural transformations of the economy –both present and forthcoming– it is essential to clarify who holds the agency in directing technological change. This is especially critical for countries in the Global South, whose legitimate technological aspirations risk being sidelined by the agendas of AI's leading nations. Many of the economic consequences

of AI are not the result of deliberate policy choices, but rather of insufficient oversight. Whether these effects can be mitigated ultimately depends on the ability of states to exercise some control over the technology. Such control is not straightforward, nor should it necessarily be understood as a barrier to technological development.

Big Tech firms have consistently advanced the case against regulation, opposing calls for precautionary measures and anticipatory state action in response to the profound socioeconomic changes brought about by the Digital Revolution. Their arguments, centered on the extraordinary technical complexity of digital technologies and the inherent unpredictability of their future trajectories, suggest that governments may struggle to regulate effectively. While it is true that the range of AI's potential applications is difficult to grasp in its entirety, this should not justify inaction. Rather, it calls for more sustained and systematic research into the societal and economic implications of AI.

AI holds undeniable promise. It could exponentially accelerate scientific research, helping to eradicate diseases, drastically reduce traffic fatalities through autonomous driving, and enable countless other breakthroughs. But there is also a darker horizon. Between 37.8% and 54.1% of AI researchers estimate that AI systems could lead to catastrophic events, including human extinction, with a probability of at least 10% (Grace *et al.,* 2024). That is, more than a third of experts in the field assign a non-negligible probability to AI having terrible results, which should give us pause.

More immediate risks are less speculative and already unfolding, such as technological unemployment, deepening inequality, and corporate concentration. These are not distant possibilities but pressing realities. Confronting them requires that economic policy approaches AI not merely as a driver of growth but also considers its risks and harmful effects. Any other stance risks indulging in illusions we cannot afford.

VIII. BIBLIOGRAPHY

Acemoglu, Daron and Pascual Restrepo. 2022. Tasks, automation and the rise in US wage inequality. *Econometrica* 90(5): 1973-2016.

Acemoglu, Daron and Pascual Restrepo. 2018. Artificial intelligence, automation and work. Working Paper 24196. *NBER working paper series*.

Crawford, Kate. 2021. *Atlas of AI*. New Haven and Londres: Yale University Press.

Official Journal of the European Union. 2022a. Digital Services Act. https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32022R2065

Official Journal of the European Union. 2022b. Digital Markets Act. https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32022R1925

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv:2303.10130

Freitag, Charlotte, Mike Berners-Lee, Kelly Widdicks, Bran Knowles, Gordon S. Blair and Adrian Friday. 2021. The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns* 2(9) https://doi.org/10.1016/j.patter.2021.100340

Goodfellow, Ian J. et al. 2014. Generative Adversarial Nets. *Advances in Neural Information Processing Systems* 27: 2672-2680.

Grace, Katja et al. 2024. Thousands of AI Authors on the Future of AI. AI impacts. https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf

Human-Centered Artificial Intelligence. 2024. *Artificial Intelligence Index Report 2024*. Standford University.

Kangas, Olli, Signe Jauhiainen, Miska Simanainen, Minna Ylikännö. 2021. *Experimenting with Unconditional Basic Income. Lessons from the Finnish BI Experiment 2017-2018*. Cheltenham (UK) and Northampton (MA, EE. UU.): Edward Elgar Publishing.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25

Kurzweil, Ray. 2024. *The Singularity is Nearer. When We Merge with AI*. Nueva York: Viking.

Li, Pengfei, Jianyi Yang, Mohammad A. Islam and Shaolei Ren. 2023. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Fooprint of AI Models. https://arxiv.org/pdf/2304.03271

Luccioni, Alexandra Sasha, Sylvain Viguier and Anne-Laure Ligozat. 2022. Estimating the Carbon Footprint of BLOOM a 176B Parameter Language Model. https://arxiv.org/pdf/2211.02001

Parlamento Europeo. 2024. Artificial Intelligence Act. Consolidated text. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html#title2

Schmidhuber, Jürgen and Sepp Hochreiter. 1997. Long Short-Term Memory. *Neural Computation* 9(8): 1735-1780.

Tirole, Jean. 2024. Competition and industrial policy in the 21st century. *Oxford Open Economics* 3: i983-i1001.

Vanderbauwhede, Wim. 2023. Emissions from ChatGPT are much higher than from conventional search. https://limited.systems/articles/google-search-vs-chatgpt-emissions/

Vaswani, Ashish. et al. 2017. Attention Is All You Need. arXiv:1706.03762

Waymo. 2024. Waymo Driver. https://waymo.com/intl/es/waymo-driver/ Reviewed on August 25, 2024.

INDUSTRIAL ARTIFICIAL INTELLIGENCE IN CHINA

CONTENT:

I. Artificial intelligence for industrialization **II.** Concept of Industrial Artificial Intelligence **III.** Industrial Artificial Intelligence Policies in China **IV.** Current situation and outlook **V.** Conclusions **VI.** Bibliography

> BEATRIZ JUÁREZ AGUILAR
Professor at the Graduate Studies Division of the UNAM Faculty of Law and Advisor to the Institute for Industrial Development and Digital Transformation (INADI)

I. ARTIFICIAL INTELLIGENCE FOR INDUSTRIALIZATION

In July 2025, the White House released *Winning the AI Race: America's AI Action Plan*, stressing the importance of artificial intelligence (AI) for "human flourishing, economic competitiveness, and national security", stressing the need for the American people to win the AI race. The main three lines are accelerating innovation, building American artificial intelligence infrastructure, and leading in international diplomacy and security (07/2025a).

Among the policy actions of this plan, "Support Next-Generation Manufacturing" highlights AI's capacity –along with robotics and other technologies– for the creation of new paths in manufacturing and logistics, that could include applications for national defense and security. Therefore, America and their "trusted allies" should be "world-class manufacturers of these next-generation technologies" in order to lead this "new industrial renaissance" (The White House, 2025b: 7).

The new emphasis of developed country governments on reindustrialization, explains Philip Pilkington, is due to "the excesses" caused by the New Economy of the 1990s, the post-industrial economy, and the two financial crises of 2000 and 2008, which, together with current geopolitical tensions, have led to the exploration of reindustrialization under concepts such as "'reshoring'" and "'decoupling'" (2024). The use of artificial intelligence in these processes is feasible, because reindustrialization "can […] assume different forms", could it be "backshoring of activities that already existed" or "the creation of new industrial activities" (Capello & Cerisola, 2022: 2).

In his recommendations to the West, Pilkington suggests that "quick-fix solutions" –especially those leading "to dangerous levels of global conflict around

trade flows"– should be set aside and the focus should be on "developing a multi-decade plan to massive boost investment" (2024).

This is precisely what China is already working on: a comprehensive industrial modernization plan spanning more than ten years, in which artificial intelligence is an essential element.

This chapter introduces a preliminary analysis, demonstrating that China has been developing a framework of policies and legal instruments aimed at addressing the main technical tasks necessary for the successful and profitable implementation of industrial artificial intelligence: managing data collection and processing, integration of legacy systems, investing in workforce skills and adopt security and privacy safeguards.

DELIMITATION

This study will analyze China's national public policies, because those frame the government's main objectives and serve as a guide for national, regional, and local authorities, as well as for national and international companies. The focus will be industrial policies from 2022-2024, that specifically relate to both *New Industrialization* and industrial artificial intelligence. However, given the complexity and enormous amount of material, the analysis will center on manufacturing, which is considered the backbone of Chinese industry and economy.

It should be noted that the growing geopolitical confrontation between the US and China has made artificial intelligence a very sensitive issue, which is why China is very selective in the information it publishes.

For reasons of space, this analysis will also not cover the complex relationship between intra-industry actors (e.g., equipment manufacturers) or supply chain relationships.

After this introduction, part II. will delineate the concept of industrial artificial intelligence; part III. will dive into the principal industrial artificial intelligence in China; part IV. will focus on the implementation processes of IAI in China; and part V. will offer conclusions.

II. CONCEPT OF INDUSTRIAL ARTIFICIAL INTELLIGENCE

Industrial artificial intelligence (IAI) is a specialization of artificial intelligence that is applied in industrial settings (IFS, n. d.), central to our economies like industry, infrastructure, mobility and healthcare (Siemens 2025). This more

focused application does not seek to develop systems with human-like intelligence (Precognize, 2022). IAI as an interdisciplinary research field, it encompasses machine learning, natural language processing and robotics; it is a systematic discipline focused "on the development, validation, deployment and maintenance of AI solutions" for industrial applications that have a non-variable performance (Peres and others, 2020: 220122).

An important characteristic of IAI is that it has "zero-tolerance" for uncertainty, i.e. "false positives or negatives, delayed insights, or unreliable predictions" (Precognize, 2022). Therefore, the designation of industrial-grade artificial intelligence denotes a level of quality that is characterized by reliability, security, and the capacity to meet the stringent requirements and standards associated with the most demanding professional media environments (Siemens 2025).

In this way, IAI contributes to the construction of smart and resilient industrial systems (Lee, Singh & Azamfar, 2019: 5). In particular, it contributes to increase efficiency by reducing human error, streamlining production, and improving operational efficiency. It impacts predictive maintenance because, by analyzing equipment data, it predicts failures before they occur, reducing downtime and preventing costly repairs. It allows for greater customization options by making production more flexible and responding more quickly to market changes (IFS, n. d.).

IAI also facilitates workflow automation, increases safety, improves supply chain management and performance, and removes obstacles to operational efficiency. As a result, it drives profitability, maintains long-term competitive advantages (Precognize, 2022), and improves decision-making and productivity, while also helping to drive innovation. In this way, IAI becomes "a crucial tool for industrial transformation" (IFS, n. d.). A bonus is that while IAI applications optimize the performance of industrial systems, they also inspire theoretical advances in the field of artificial intelligence (Palhares, Yuan & Wang, 2019: 9636).

It is important to note that each industrial sector may have different uses for industrial artificial intelligence. In the case of manufacturing, the application of artificial intelligence can be divided into three sectors: the upper, middle, and lower basic layers. The first includes industrial software and hardware such as infrastructure and smart industrial equipment; the middle layer of the system includes industrial control systems and industrial internet platforms; and the application layer can be applied to electronic communications, energy, and electricity and is mainly used in design, on-site operations, sales forecasting, energy conservation, emissions reduction, etc. (36 Krypton Research Institute, 2024). The integration of manufacturing with the new generation of information technologies is called smart manufacturing (Dù 2024).

For example, automation is based on the premise that a machine, piece of equipment, or robot performs a certain predefined activity, following specific rules in a limited number of scenarios. Artificial intelligence enables pattern recognition within data, learning from experience, and improving future performance, allowing these machines, equipment, or robots to identify solutions to complex problems within a given space (Lee, Singh & Azamfar, 2019: 5).

GENERAL REQUIREMENTS

Industrial artificial intelligence requires a set of advanced technologies for effective performance, including:

A. Machine learning and deep learning: these facilitate systems' ability to learn from data, enhance their performance progressively, and generate informed predictions, which leads to an optimization of efficiency in industrial processes.

B. The Internet of Things (IoT): these devices aggregate substantial volumes of data from machines and systems that collect them, which is subsequently analyzed by artificial intelligence to optimize processes, minimize waste, and enhance productivity.

C. Robotics and automation: the core of Industrial AI is robotics, because it facilitates the automation of tasks spanning from assembly line operations to quality inspections.

D. Edge computing: given the IAI tasks are time-sensitive, edge computing allows AI systems to make real-time decisions through local data processing, instead of depending on cloud computing (IFS, n. d.).

Despite the great potential of industrial artificial intelligence, its implementation faces many challenges. A basic reason for this is the large investment required to make the necessary adjustments for the integration of AI applications into corporate structures and throughout the entire value chain. Companies are therefore advised not only to understand its potential, but also the main requirements and organizational changes necessary to harness its full potential (Peres and others, 2020: 220122-220124).

Industrial and Financial Systems (IFS), a Swedish enterprise software company, highlights the following among the core tasks necessary for the successful and profitable implementation of IAI:

A. Data collection and processing: IAI is highly dependent on large volumes of high-quality data; however, for many organizations it becomes a challenges to effective collect, organize, and analyze such data.

B. Integration of legacy systems: the existence of older industrial systems that are incompatible with IAI can become an obstacle.
C. Workforce skills: Workers need new skills, and their employers need to invest in retraining.
D. Security and privacy: In the face of cyber threats, protecting the data collected and to be analyzed becomes an essential priority (IFS, n.d.).

In this regard, research by Peres et al. highlights that a critical obstacle to digital transformation and technology adoption is that many companies do not know how to integrate AI into their business models and processes, which is why guidance on how to do so is needed, including common practices, evidence, and successful experiences. For example, companies often collect data in multiple formats, without standards and with varying degrees of quality, so systematic approaches are needed for the implementation, deployment, and management of IAI solutions. (2020: 220122-220124).

A useful tool for solving this problem is "lighthouse factories", whose main asset is intelligence, which is fully exploited through the adaptation and application of IAI, making it a key driver of development. The incorporation of advanced technologies in these companies has brought them significant economic benefits, making them a "true guiding light" (Lee, Singh & Azamfar, 2019: 8) that illuminates and shows the way for others.

III. INDUSTRIAL ARTIFICIAL INTELLIGENCE POLICIES IN CHINA

In the context of the digital economic sphere, particularly within the domain of artificial intelligence, the government has implemented a series of techno-economic public policies since 2015, with a notable emphasis from 2017 onward. These policies have played a pivotal role in shaping the primary direction of research, innovation, development, and AI's application scenarios, emphasizing its application for industry. Subsequently, a series of voluntary ethical frameworks and industrial self-regulation were released. (Juárez-Aguilar, in press, a).

Meanwhile, businesses, government, and society quickly adopted AI through e-commerce, the sharing economy, gaming, fintech, social media, and many other areas. However, the indiscriminate use of AI began to create multiple social problems that existing regulations and voluntary measures (such as ethical frameworks and self-regulation) were unable to solve. As part of a broader reform of the digital economy, the government has been gradually developing

and applying specific regulations since 2022, aimed at addressing the most pressing challenges faced by the stakeholders involved, such as consumers, minors, older adults, and workers in the sharing economy, among others (Juárez-Aguilar, in press, a).

At the same time, the US increased restrictions on the Chinese technology sector, such as the export of products, services, data, and even labor mobility and investment, focusing specifically on halting Chinese advances in artificial intelligence. In response, the Chinese government emphasized the relationship between security and technological self-sufficiency, while reorienting technology governance toward specific objectives such as industrial strength and national security, discouraging areas more related to leisure, such as gaming, e-commerce, etc. (Creemers and others, 2023).

It is at this point that the concept of *New Industrialization* emerged, with a renewed emphasis on artificial intelligence.

NEW INDUSTRIALIZATION + ARTIFICIAL INTELLIGENCE

For almost twenty years, the Chinese government has promoted multiple public policies that work to integrate new technologies with the real economy in order to drive development and innovation. However, in recent years it has emphasized the primacy of the real economy for the country, particularly manufacturing.

Since 2022, the Chinese government has been developing the concept of *New Industrialization* (新型工业化, *xīnxíng gōngyèhuà*)[1], which refers to the idea that new generations of technology should contribute to modernizing the entire industrial landscape in order to create new models of production, business, and new forms of economic growth, prioritizing safety over economic results. The target date for achieving these objectives has been set for 2035 (Juárez-Aguilar, 2024). This *New Industrialization* recognizes the centrality of manufacturing and the domestic market and lists a series of measures for industrial transformation, including updating traditional industries; consolidating and expanding industries where advantages already exist; promoting and expanding emerging industries; designing future industries in a forward-looking manner; and deepening international cooperation. (State Council Information Office Website, 2023).

1 To enable better visualization, Chinese characters will not be inserted in italics.

Several experts from Chinese academia have agreed on the importance of fully integrating artificial intelligence into industry in order to achieve the goals of *New Industrialization*. Wén Jūnléi (文钧雷), vice president and professor at the *Virtual Reality Research Institute of the Cloud Computing Center* of the Chinese Academy of Sciences, asserts that since AI and smart manufacturing are at the heart of the new industrial competition and manufacturing is the foundation of the country, their combination can enable China to become a manufacturing power capable of overcoming domestic and international challenges, as well as providing a major boost to the development of new economic fields (Chengdu Daily, 2022).

Dù Chuánzhōng (杜传忠), a special researcher at the *China New Generation Artificial Intelligence Development Strategy Research Institute* and professor and director of the *Industrial Economics Institute* at Nankai University, details how AI enables new industrialization. Part of his argument reveals that from the standpoint of permeability, AI "can be deeply integrated into all aspects of the manufacturing industry and the upstream and downstream industrial chains, and play an enabling, intellectual and assigning role to industrial development"; from the perspective of synergy, AI "organically connects the fields of consumption, production, circulation, and distribution, which helps improve the coordination, order, and efficiency of the entire process of social reproduction"; and from the perspective of innovation, AI "can achieve self-evolution and self-upgrading of technology through deep learning and rapid iteration, which in turn leads to the optimization and upgrading of the entire industry" (Dù, 2024).

Zhōu Jì (周济), professor and former president of the *Chinese Academy of Engineering*, emphasizes that smart manufacturing requires both 5G and the industrial internet to achieve digital transformation, development, and intelligent improvement of the manufacturing industry, as well as innovation in the industrial model. Therefore, "5G + industrial internet + artificial intelligence" will empower *New Industrialization*, enabling innovation in products, production technology, industrial models, and manufacturing system integration, thus reaching the climax of the *Fourth Industrial Revolution* (Chén, 2024).

INDUSTRIAL ARTIFICIAL INTELLIGENCE POLICIES

The plans presented below are organized under the current Five-Year Plan, which is the *14th Five-Year Plan for National Economic and Social Development of the People's Republic of China and Long-Term Objectives Through 2035*. They coincide in seeking to take advantage of the opportunities offered by the new

scientific and technological revolution to support industrial strengthening and achieve the objectives of *New Industrialization*, but each plan approaches this issue from different angles.

On December 30, 2023, the Ministry of Industry and Information Technology (MIIT) and eight other institutions presented ***Guiding Opinions on Accelerating the Transformation and Upgrading of Traditional Manufacturing Industries***[2]. This plan considers that active adoption and upgrading of traditional industries will improve the resilience and security of the industrial and supply chain, strengthening China's competitiveness and position in the global division of industrial labor, and in industrial transformation It is proposed that by 2027, there will be an improvement in China's traditional manufacturing industry, with high-level, intelligent, green, and integrated development.

Some of the proposed measures emphasize advancing the smart transformation of enterprises and developing digital networks, while accounting for the specific characteristics and requirements of each industry, accelerating the integration of the entire manufacturing process and elements with artificial intelligence, big data, cloud computing, 5G, the Internet of Things, among others (CAICT, 2023).

On January 31, 2024, MITT and seven other institutions presented ***Implementation Opinions on Promoting the Innovation and Development of Future Industries***[3], which seeks to bolster manufacturing through the development and application of cutting-edge and disruptive technologies, such as metaverse, humanoid robots, brain-computer interfaces, and quantum information. Specifically, "future manufacturing" seeks to develop smart manufacturing, biological manufacturing, nano manufacturing, laser manufacturing, and circular manufacturing, as well as to achieve disruptive advances in smart manufacturing, smart detection, analog simulation, and other key technologies. With targets for 2025 and 2027, it also aims to promote flexible manufacturing, shared models, and the development of the industrial internet and the industrial metaverse, among other things (MITT + 6, 2024).

2 《关于加快传统制造业转型升级的指导意见》 *guānyú jiākuài chuántông zhìzào yè zhuǎnxíng shēngjí de zhǐdǎo yìjiàn.*

3 《关于推动未来产业创新发展的实施意见》 *guānyú tuīdòng wèilái chǎnyè chuàngxīn fāzhǎn de shíshī yìjiàn*

Also in January 2024, news circulated that the State Council of the PRC met to work on the ***Action Plan to promote artificial intelligence to boost New Industrialization***[4]. The meeting highlighted the need to coordinate high-quality development with high-level security; the plan's central objective was smart manufacturing; the deep integration of AI with manufacturing as the main means, and application scenarios as the driving force. Other key objectives include the need for the smart upgrading of key industries, the development of smart products, and the strengthening of high-level manufacturing systems (Chinese Government Website, 2024).

On May 11, 2024, the State Council approved the ***Action Plan for the Digital transformation of the Manufacturing Industry***[5]. It emphasizes that this transformation not only involves industrial and development model optimization, but it is also directly related to national, economic, industrial, scientific, and technological security. It was agreed to explore typical scenarios in different fields, according to the specific needs of the industry. Similarly, it is necessary to accelerate research, promotion, and application of advances in key technologies; work on equipment connection, mutual recognition of protocols, formulation of standards, construction of platforms, among others. It also highlighted the importance of increasing long-term support mechanisms for the digital transformation of small and medium-sized enterprises and combining them with large-scale equipment renewal actions and the implementation of technological transformation and modernization projects, as well as improving public service platforms (Xinhua, 2024).

In November 2024, Hé Hǎilín (何海林), head of the MIIT's *Operations Monitoring and Coordination Bureau*, recognized the current work on strengthen and coordinating public policies, deepen international cooperation, and striving to contribute to China's efforts to build a global industrial system and supply chains that are "secure, stable, flexible, efficient, open, inclusive, and mutually beneficial" (State Council, 2024).

The steps to be taken are:

1. Actively guide and promote companies to integrate into international industrial and supply chains.

4 《推动人工智能赋能新型工业化行动方案》 *tuīdòng réngōng zhìnéng fù néng xīnxíng gōngyèhuà xíngdòng fāng'àn.* At the time of writing this text, it has not yet been published (September 2025).

5 《制造业数字化转型行动方案》 *zhìzào yè shùzìhuà zhuǎnxíng xíngdòng fāng'àn*

2. Strengthen the reputation of the *Made in China* brand through multiple measures, such as the boost that industrial design gives to quality improvement.
3. Vigorously develop digital commerce and green commerce. Regarding the first point, work must be done on action plans designed for the digital transformation of the manufacturing industry and for artificial intelligence to empower *New Industrialization*. Build smart factories and supply chains, develop smart logistics equipment, promote the construction of e-commerce standards that are compatible with international standards on data flow, traceability, and electronic signatures, and expand new spaces for digital commerce.
4. Improve the level of service for companies that "go abroad" (State Council, 2024).

IMPLEMENTATION PROCESSES OF INDUSTRIAL ARTIFICIAL INTELLIGENCE IN CHINA

As explained in the second section, industrial artificial intelligence requires data collection and processing, the integration of legacy systems, adequate worker skills, and the maintenance of security and privacy (IFS, n.d.). Since these areas are cross-cutting and interact with each other, within the Chinese governance framework, several policy documents and legal instruments address these topics also in an intertwined way. Below is a brief overview.

A. Data

   In 2020, the Chinese government designated data as a factor of production and began to create ways for it to underpin the economy, such as data markets. However, geopolitical confrontations have been an essential element in subordinating economic development to security. Since 2021, several legal instruments have been stipulated to regulate data processing activities in the field of industry and information technology. Particularly in the case of industrial data, it is first necessary to determine which data should be protected and then indicate which data can be used to promote economic development, while safeguarding the rights and interests of individuals, organizations, and the nation (Juárez-Aguilar, in press, b).

B. Integration with legacy systems

   As mentioned in *Guiding Opinions on Accelerating the Transformation and Upgrading of Traditional Manufacturing Industries*, authorities recognize the need for a comprehensive modernization that includes digital infrastructure and equipment throughout the industry, high-tech zones, industrial parks, etc.

For example, in terms of production equipment, it will include the application of new sensors, advanced controls, and other smart components, as well as accelerating the promotion of smart equipment and the updating and replacement of software. (CAICT, 2023). The idea to replace all equipment is quite massive, but if there is a country who may be able to do it, it is China.

C. Worker skills

Although China has been working hard for decades to adapt its education system to meet the scientific and technological challenges of modernization, the Ministry of Education is continuing this work today and is often the convening institution for many public policies related to digital transformation[6].

From the wide transformation of education, here is one example. In July 2024, the Ministry of Education published a document encouraging universities to provide support in key areas such as integrated circuit manufacturing, artificial intelligence, quantum technology, life sciences, and energy, and called for greater alignment of professional education with industry development (Bangkok Post 2024). Among the tasks outlined by the Vice Ministry of Education are dynamic adjustment of the profession, scientific design of course content, updating of teaching materials, formulation of standards for internships, and building a training base to integrate industry and education, among others (CAEA, 2024).

In August 2024, 19 universities were already in the process of revamping their engineering and technology programs in favor of high-tech areas such as AI and big data. Similarly, now there are also training programs in big data, artificial intelligence, smart manufacturing, integrated circuits, data security, among others, aimed at engineers, technicians, and workers (Bangkok Post, 2024).

D. Security

Faced with growing national and international challenges, the Chinese government has decided to strengthen its security in a multidimensional way. Among the included areas are political, military, territorial, economic, financial, cultural, social, food, and ecological security, as well as technological security, which includes artificial intelligence, data, digital infrastructure, among others (Juárez-Aguilar, in press, b).

---

6 For example, in *Guiding Opinions on Accelerating the Transformation and Upgrading of Traditional Manufacturing Industries* one of the nine convening institutions was the Ministry of Education (CAICT, 2023).

The highest relevance of security is reflected in numerous public policies and a growing legal framework. One example of a public policy instrument is the aforementioned *Action Plan for the Digital transformation of the Manufacturing Industry*, which links industrial optimization with national economic security. And one example of the legal framework is *Measures for the Administration of Data Security in the Field of Industry and Information Technology (Trial)*[7], which came into force on January 1, 2023.

IV. CURRENT SITUATION AND OUTLOOK

In November 2024, the Chinese government presented the *China Internet Development Report*[8], it specifies that this country possesses the world's largest 5G network, which integrates 74 categories of the national economy and accelerates key sectors such as manufacturing, mining, electric power, and healthcare, among others. More than 4,500 core AI companies operate nationwide, and enterprises have established nearly 10,000 digital workshops and smart factories; 421 of the latter serve as lighthouse factories, and more than 90% have AI and digital twins (CCTIME, 2024; Qinghai Daily, 2024).

According to the *International Federation of Robotics*, China is the largest market for industrial robots, with 276,288 units installed in 2023, representing 51% of the global total. The growth potential in the manufacturing sector is forecast to be between 5% and 10% annually until 2027 (IFRI, 2024b). However, in terms of robot density, China ranked third with 470 robots per 10,000 employees in 2023, only behind South Korea and Singapore. Nevertheless, this achievement is surprising considering that China entered the top 10 in 2019, implying that it has doubled its density every four years (IFRI, 2024a).

In May 2024, *36 Krypton Research Institute* (36氪研究院) presented the *Research Report on China AI+ Manufacturing Indus*try[9], which points out that artificial intelligence is transforming the country's manufacturing industry because it seeks to meet the personalized needs of users. By focusing on the user, production becomes more flexible, leaving behind the rigidity of focusing primarily on the product (2024).

---

7 工业和信息化领域数据安全管理办法 (试行) *Gōngyè hé xìnxī huà lǐngyù shùjù ānquán guǎnlǐ bànfǎ (shìxíng)*

8 《中国互联网发展报告2024》*zhōngguó hùliánwǎng fāzhǎn bàogào 2024*

9 《中国AI+制造产业研究报告》*zhōngguó AI+zhìzào chǎnyè yán jiù bàogào gǔn*

The transformations resulting from the interactions of artificial intelligence with its environment are giving rise to new challenges and opportunities. On the one hand, a whole range of application scenarios are emerging that pose security risks, as well as potential vulnerabilities in infrastructure, software, hardware, privacy, data transmission, among others. Therefore, the identification and management of emerging security risks have become increasingly significant from a strategic perspective (CCTIME, 2024).

Dù Chuánzhōng asserts, the digital transformation of many Chinese manufacturing companies has reached such a level that the mere digitization of traditional industries is not enough; only the deep integration of artificial intelligence into manufacturing can realize the goals of *New Industrialization* (2024).

And there is no time to waste, because Jīn Zhuànglóng of the MIIT has pointed out that this institution is already aiming in new directions: future manufacturing, future information, and future materials; humanoid robots, brain-computer interfaces, the metaverse, next-generation internet, 6G. In addition to seeking innovative advances in key core technologies, it aims to create a series of essential products, achieve significant breakthroughs, and build a series of business incubators, among other goals (Guāngmíng Daily, 2024).

V. CONCLUSIONS

With regard to the field of artificial intelligence, Chinese public policy instruments are founded upon an overarching concept, encompassing research and development, infrastructure, AI components, and application scenarios. Within this comprehensive approach this chapter has shown that artificial intelligence is now boosting Chinese industrial modernization. This new kind of industrial reinforcement is aimed at updating traditional industries and creating new industrial activities. Its scope is very broad, ranging from all kinds of current application scenarios as well as future industries; from world-leading companies to SMEs; from traditional industries to green development.

A more detailed revision shows that China appears to be strategically well-positioned for this transformation, as it has already incorporated the core technical tasks essential for the successful and profitable implementation of industrial artificial intelligence into its public policies and legal frameworks, i.e. data collection and processing, integration of legacy systems, reskilling of workers, security and privacy.

UNCTAD emphasizes that it is critical for developing countries "to consider the diffusion and direction of frontier technologies and their impact on the economy, and to adapt catch-up strategies accordingly". This UN agency stresses particularly the need to tailor timely strategies for "integrating artificial intelligence into broader industrial and innovation policies that support inclusive and sustainable development", or we may the risk of falling behind (2025:1).

In order to do this, UNCTAD remind us, we can learn "from global examples" (UNCTAD 2025:1). And given the complexity of each industry and the heterogeneity of the Chinese landscape at the local level, countless examples can be found. Therefore, within the industrial transformation with AI –just like a lighthouse factory– China shows us, developing countries, a possible way forward.

Regarding artificial intelligence, developing countries have limited resources for AI development, therefore we need to select carefully which areas should be supported first, particularly the essential industries' situation should be meticulously analyzed.

Yet one first step, as indicated by Prof. Zhōu Jì, is to construct basic infrastructure for IAI (like 5G and industrial internet), in order to take advantage of what big data, the new generation of artificial intelligence, and smart manufacturing have to offer (Chén, 2024). Keeping always in mind that each industrialization follows a different path (Pilkington 2024).

VI. BIBLIOGRAPHY

36 Krypton Research Institute 36氪研究院 (24/05/2024). AI+ China, **中国AI+制造产业研究报告**Research report of manufacturing industry. https://finance.sina.com.cn/tech/roll/2024-05-30/doc-inawzavy4636958.shtml

Bangkok post (23/08/2024). Universities in China change course to meet government's need for hi-tech workforce https://www.bangkokpost.com/world/2852553/universities-in-china-change-course-to-meet-governments-need-for-hi-tech-workforce

CAEA China Adult Education Association **中国成人教育协会** (25/11/2024). **教育部副部长吴岩:以产教融合推动职业教育高质量发展**, Wu Yan, Vice Minister of the Ministry of Education: Promote the high-quality development of vocational education through the integration of industry and education. https://www.caea.org.cn/newsinfo/7786781.html

CAICT (30/12/2023). **关于加快传统制造业转型升级的指导意见**, *Guiding Opinions on Accelerating the Transformation and Upgrading of Traditional Manufacturing*. https://mp.weixin.qq.com/s/CVjsdPHa5RkggX-nZRRAqQ

Capello, R., & Cerisola, S. (2022). Regional reindustrialization patterns and productivity growth in Europe. *Regional Studies*, 57(1), 1–12. https://doi.org/10.1080/00343404.2022.2050894

CCTIME (28/11/2024). **速看!互联网发展报告重点来了**, Quick look! Here comes the focus of the Internet Development Report. https://finance.sina.com.cn/tech/roll/2024-11-28/doc-incxqqsy9863893.shtml

Chén Yǒngquán **陈永权** (20/11/2024). **没有强大的**5G+**工业互联网, 就没有新一代人工智能**, Without a powerful 5G + industrial Internet, there will be no new generation of artificial intelligence. *Yangtze River Daily***长江日报**. http://m.ce.cn/cy/gd/202411/20/t20241120_39208886.shtml

Chengdu Daily**成都日报** (04/11/2022). **人工智能如何赋能制造业**?, How can AI empower manufacturing?. https://cdst.chengdu.gov.cn/cdkxjsj/c108732/2022-11/04/content_bcf3c911acb24e8ebd6c443e66870e64.shtml

Chinese Government Website **中国政府网** (27/01/2024). **国务院本周政策**, State Department Policy of the Week, https://www.gov.cn/yaowen/liebiao/202401/content_6928620.htm

State Council RPC (22/11/2024).**国务院政策例行吹风会回应关切**, State Council policy briefing to respond to concerns. https://www.gov.cn/xinwen/2024zccfh/27/index.htm

Creemers, Rogier, Johanna Costigan, Paul Triolo, Tom Nunlist, Lauren Dudley, Mei Danowski, Martin Chorzempa, Karman Lucero, Seaton Huang (2023). "Is China's Tech 'Crackdown' or 'Rectification' Over?". *DigiChina*. https://digichina.stanford.edu/work/is-chinas-tech-crackdown-or-rectification-over/

Dù Chuánzhōng**杜传忠**(03/2024). **人工智能全面赋能新型工业化**, Artificial intelligence fully empowers new industrialization. *China Social Sciences Journal*. https://www.xuexi.cn/lgpage/detail/index.html?id=1737982780920682307&item_id=1737982780920682307

IFR International Federation of Robotics (20/11/2024 B). Global Robot Density in Factories Doubled in Seven Years. https://ifr.org/ifr-press-releases/news/global-robot-density-in-factories-doubled-in-seven-years

IFR International Federation of Robotics (24/09/2024 A). Record 4 Million Robots on Factory Floors Worldwide. https://ifr.org/downloads/press2018/2024-SEP-24_IFR_press_release_World_Robotics_2024_-_global_market.pdf

Industrial and Financial Systems IFS (n. d.). What is Industrial AI™? A Comprehensive Overview. https://www.ifs.com/nl/what-is/what-is-industrial-ai (Accessed 12/19/2025)

Jīn Zhuànglóng 金壮龙, (16/11/2024). 金壮龙: 进一步全面深化工业和信息化领域改革 为推进新型工业化注入强大动力, Jīn Zhuànglóng: Further deepen the reform in the field of industry and information technology in an all-round way and inject strong impetus into the promotion of new industrialization. *Qiushi Magazine*《求是》杂志. https://www.gov.cn/zhengce/202411/content_6987599.htm

Juárez-Aguilar, Beatriz (2024). Después de Hecho en China 2025: la Nueva industrialización en China. En Arturo Oropeza García (Coord.), *Hacia una nueva visión del desarrollo industrial digital en México*. INADI-CCE. 344-358. https://www.inadi.mx/libros-inadi

Juárez-Aguilar, Beatriz (in press, a). La gobernanza de la inteligencia artificial en China.

Juárez-Aguilar, Beatriz (in press, b). Los datos industriales en China.

Lee, Jay, Jaskaran Singh & Moslem Azamfar (2019). Industrial Artificial Intelligence, *Computers and Society*. DOI: 10.48550/arXiv.1908.02150

MIIT (15/08/2024 a). 工业和信息化部办公厅关于组织开展人工智能赋能新型工业化典型应用案例征集工作的通知, Notice of the General Office of the Ministry of Industry and Information Technology on organizing and carrying out the collection of typical application cases of artificial intelligence enabling new industrialization. https://www.ncsti.gov.cn/kcfw/xmsb/202408/t20240815_175395.html

MIIT (31/01/2024 b). 关于推动未来产业创新发展的实施意见 解读. *Interpretation of the Implementation Opinions on Promoting the Innovation and Development of Future Industries*. https://zwgk.mct.gov.cn/zfxxgkml/zcfg/zcjd/202401/t20240131_951103.html

MIIT + 6 (31/01/2024). 关于推动未来产业创新发展的实施意见, *Implementation Opinions on Promoting the Innovation and Development of Future Industries*. https://zwgk.mct.gov.cn/zfxxgkml/kjjy/202401/t20240131_951102.html

State Council Information Office Website 国务院新闻办公室网站(03/03/2023). 工业和信息化部部长金壮龙: 加快推进新型工业化 做强做优做大实体经济, Jin Zhuanglong, Minister of Industry and Information Technology: Accelerate the promotion of new industrialization and make the real economy stronger, better and bigger. https://www.ncsti.gov.cn/kjdt/ztbd/2023qglh/bztd/202303/t20230303_109930.html

Palhares, R. M., Y. Yuan and Q. Wang, (2019). Artificial Intelligence in Industrial Systems. *IEEE Transactions on Industrial Electronics*, 66, 12. 9636-9640. DOI:10.1109/TIE.2019.2916709.

Peres, Ricardo Silva, Xiaodong Jia, Jay Lee, Keyi Sun, Armando Walter Colombo & José Barata. (2020). Industrial Artificial Intelligence in Industry 4.0–Systematic Review, Challenges and Outlook. *IEEE Access*, 8, 220121–220139. DOI: 10.1109/ACCESS.2020.3042874

Pilkington, Philip (15/04/2024). Reindustrialization in the Age of Fragmentation. https://hiia.hu/en/reindustrialization-in-the-age-of-fragmentation/

Precognize (2022) Industrial IA. https://precog.co/glossary/industrial-ai/

Qinghai Daily (22/11/2024). **我国人工智能与制造业深度融合**, Chinas artificial intelligence and manufacturing industry are deeply integrated. https://finance.sina.com.cn/jjxw/2024-11-22/doc-incwwmew4200591.shtml

Siemens (2025). Artificial Intelligence Glossary. https://www.siemens.com/global/en/company/stories/artificial-intelligence.html

The White House (07/2025a). White House Unveils America's AI Action Plan. *The White House*. https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/

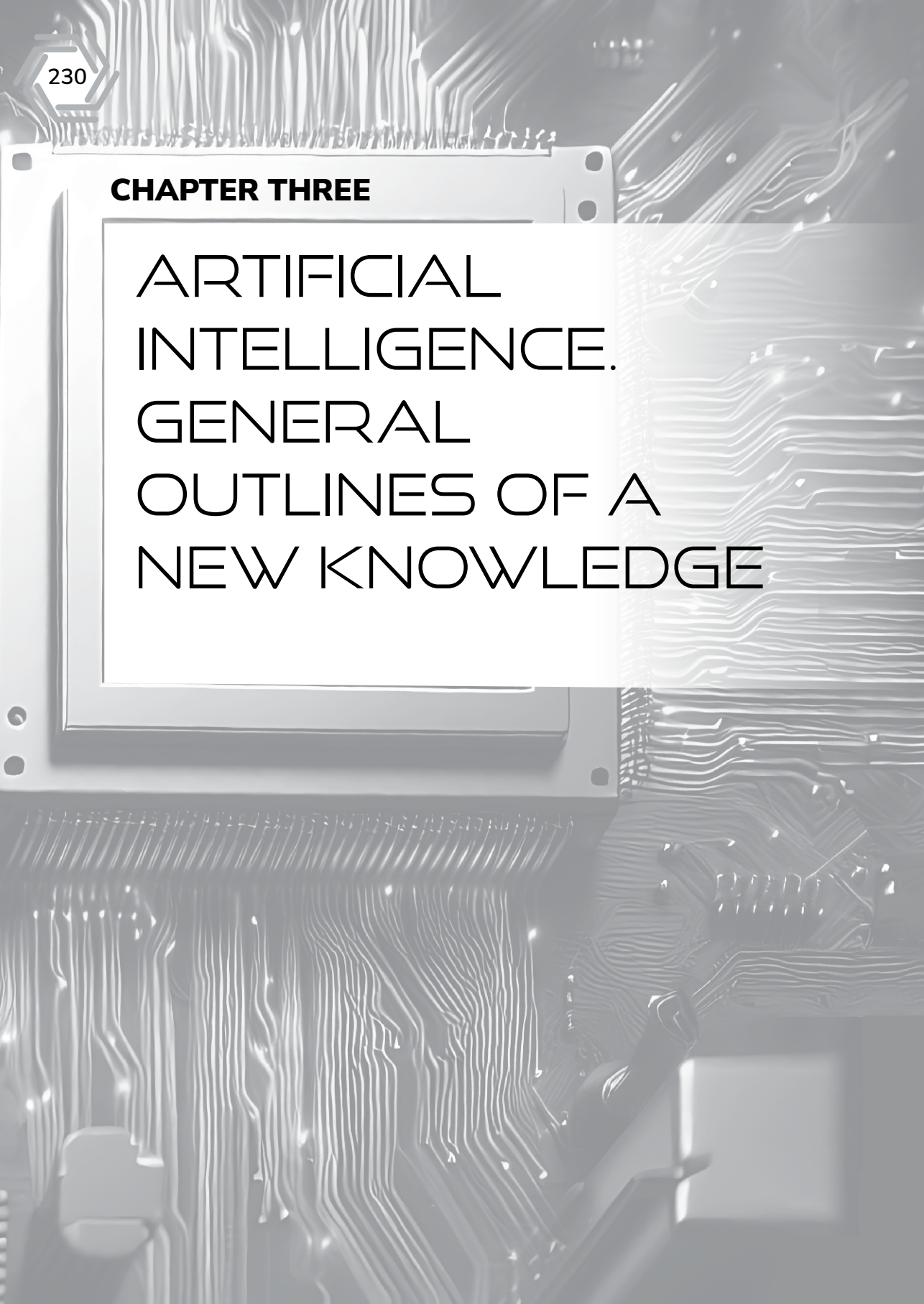
The White House (07/2025b). Winning the Race America's AI Action Plan. *The White House*. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

Xinhua News Agency **新华社** (20/05/2024). The National Standing Committee deliberated and approved the Action Plan for Digital Transformation of Manufacturing Industry, **国常会审议通过制造业数字化转型行动方案**. http://www.echinagov.com/news/356599.htm

UNCTAD (8/2025). Policy Brief 120, Preparing to seize artificial intelligence opportunities with strategic national policies. UNCTAD/PRESS/PB/2025/3 No. 120. https://unctad.org/system/files/official-document/presspb2025d3_en.pdf

CHAPTER THREE

ARTIFICIAL INTELLIGENCE. GENERAL OUTLINES OF A NEW KNOWLEDGE



PAST, PRESENT, AND FUTURE OF ARTIFICIAL INTELLIGENCE

CONTENTS:

I. Introduction **II**. Artificial Intelligence Forerunners **III.** A new discipline is born **IV.** The First Achievements of Artificial Intelligence **V.** The First AI Winter (1974-1980) **VI.** The Re-birth of AI (1980-1987) **VII.** The Second AI Winter (1987-1993) **VIII.** The Promised Results (1993-2011) **IX.** The Use of Large quantities of data (2011-data) **X.** The Future of Artificial Intelligence **XI.** Should we be afraid of Artificial Intelligence? **XII.** Conclusions **XIII.** Bibliography

> CARLOS A. COELLO COELLO
Member of El Colegio Nacional, Researcher at the Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN) in Mexico City, and Visiting Professor at the Basque Center for Applied Mathematics in Spain.

I. INTRODUCTION

Currently, Artificial Intelligence (AI) is a very popular field and it is clear that its use is already having a significant impact on our everyday lives. In fact, it is expected that such impact increases even more in the next few years.

As has happened before with other technologies, AI initially raised amazement that eventually transformed into fear. Such fear is grounded on the fact that a large portion of the population worldwide has a very limited knowledge about AI and of its potential and limitations. Furthermore, science fiction has fed, during many years, this fear to AI technologies in a significant number of novels, movies and TV series.

In this chapter, we will discuss the development of AI from a historical perspective, starting from its inception (in the 1950s) up to our days. Throughout this quick historical overview of AI, we will witness the way in which AI algorithms have evolved, starting with relatively simplified models to the sophisticated approaches available today which are able to defeat the (human) world champion of Go and to chat (in a very realistic way) with a human. Our historical trip will take us from the so-called symbolic AI which was found to be too limited to solve realistic problems to the sub-symbolic AI that is so popular today, which relies on huge amounts of data and lots of computer power.

Clearly, AI has revolutionized science and has had a significant impact even in the way in which research is being conducted. This was recognized by the 2024 Nobel Prize committees. But in addition to the scientific progress that is being fueled by AI today, this technology is also changing our everyday lives with increasingly powerful applications that allow us to potentiate our creativity.

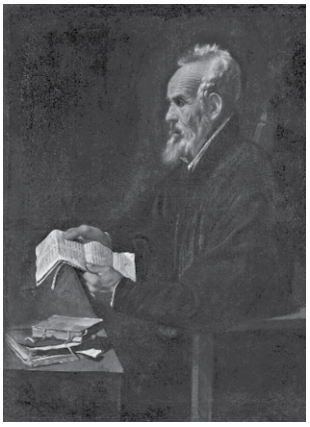
Today, newborns will grow up with increasingly powerful and sophisticated AI algorithms, and they will certainly be able to assimilate this new technology a lot easier that today's adults, analogously to what happened with smart mobile phones and smart TVs.

But perhaps the most valuable lesson that remains to be learned is that, whether we like it or not, AI is a technology that will remain with us and, therefore, it will become increasingly important in the years to come, to try to adapt to it, by learning at least about its basic use. This should help us to reduce our fear when reading any sensationalist note in social media about the dangers of AI. Simultaneously, our knowledge of this technology will allow us to make a more rational and appropriate use of it. In fact, this chapter aims to contribute to a certain extent to acquiring some of this (necessary) basic knowledge of AI technology.

II. ARTIFICIAL INTELLIGENCE FORERUNNERS

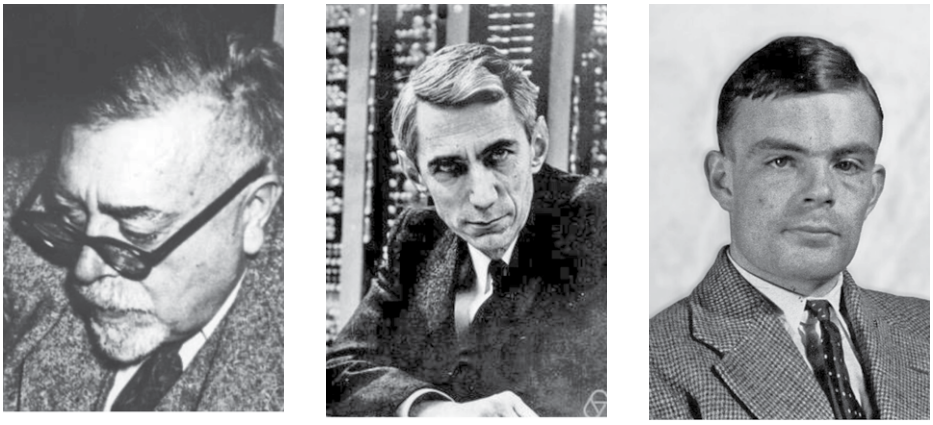
The study of human intelligence has fascinated humankind for many years. The fundamental premise of artificial intelligence is that the process of generating thoughts can be mechanized [Russell and Norvig, 2021]. The study of mechanical (or formal) reasoning is very old and is well-documented since at least 1000 B.C. in China, India and Greece. The Spanish philosopher Ramón Llull (see Figure 1) designed and built (in the Thirteenth Century) a logical machine in which theological theories, subjects and predicates were represented by using geometrical shapes that he considered to be "perfect" (e.g., circles, squares and triangles). Through the use of levers, cranks and a wheel, it was possible to use this machine to demonstrate that a certain postulate was either true or false [Mata, 2006]. The work of Llull had a great influence on Gottfried Leibniz, who speculated (in the Seventeenth Century) about the future existence of a universal language to express reasoning. Leibniz believed that in the future, philosophers would not need to debate, since it would be possible to mathematically prove if a certain postulate was either correct or incorrect.

FIGURE 1 Ramón Llull.



SOURCE Wikipedia (image in the public domain).

FIGURE 2 From left to right: Norbert Wiener, Claude Shannon and Alan Turing.



SOURCE Wikipedia (images in the public domain).

All these ideas had a great influence on the development of the "hypothesis of the physical symbols" formulated by Allen Newell and Herbert Simon. This hypothesis was the main foundation of the so-called symbolic artificial intelligence, which was the predominant branch of AI in the early days of this discipline.

Historically, however, three are the characters who are considered to be the main forerunners of the development of artificial intelligence (see Figure 2): (1) Norbert Wiener, (2) Claude Shannon and (3) Alan Turing. Next, we will briefly describe the main contributions (as related to AI) made by each of them.

Norbert Wiener was a prominent American mathematician who proposed, in the 1930s (together with the Mexican physiologist Arturo Rosenblueth) a discipline that was called ***cybernetics*** [Wiener, 1948], which studies (in a transdisciplinary way) circular processes such as feedback systems in which the outputs also act as inputs. In its early days, this discipline was applied to a wide variety of areas, including biological, cognitive, mechanical and even social processes.

Claude Shannon was an American engineer who published a paper in 1948 [Shannon, 1948] which originated a discipline that would be later called "information theory" that consists of a mathematical theory of communication. In this paper, Shannon presents the basic concepts of digital communication and then develops concepts such as information's entropy and redundancy. Additionally, Shannon introduced the term *bit* (which would be later attributed to John Tukey) which he used to denote a unit of information.

Alan Turing was an English mathematician who published a paper in 1950 [Turing, 1950] in which he proposed an operational definition of intelligence. Turing's proposal was very simple. He proposed a test (which is known today as the "Turing test"). This test consists in having a person in front of a computer terminal. In a separate room there would be a human and in another one there would be a computer operating some software without human assistance. This person would be interacting through a computer without knowing who replied to him/her (either the other human or the computer). If, at some point, the human could not distinguish who replied, then, according to Turing, we could say that the computer is "intelligent", because its answers were indistinguishable from those of a human.

Nevertheless, Turing is better remembered for his work in which he defined a mathematical model of a universal computer (which is known today as "Turing universal machine"). Turing defined computable problems (i.e., problems that can be solved in a computer) and non-computable problems, originating a field which is called today "theory of computation".

In the 1940s, there were several important developments within this discipline which, at that point in time, didn't have a name yet. From them, perhaps the most influential work was the paper entitled "*A logical calculus of the ideas immanent in nervous activity*" [McCulloch and Pitts, 1943]. In this paper, Warren McCulloch and Walter Pitts propose a model of artificial neurons in which each neuron is characterized for being either "on" or "off". In this model, a neuron is activated (i.e., it's in an "on" state) as a response to the stimulus of a sufficient number of neighboring neurons. Additionally, McCulloch and Pitts managed to prove that any computable function could be modelled by some network of neurons connected with each other and that all the logical operators (e.g., AND, OR, NOT) could be implemented through the use of simple structures in such a network. They also suggested that networks of neurons that were properly configured would be able to learn. For developing this model, McCulloch and Pitts relied on three main sources: (1) their knowledge about basic physiology and the way in which neurons actually work in our brain, (2) a formal analysis

FIGURE 3 Marvin Minsky.



SOURCE Wikipedia (image available under a *Creative Commons* license).

of propositional logic [Whitehead and Russell, 1910] and (3) the recent theory of computation proposed by Alan Turing. This paper is considered as the first ever to be fully focused on artificial intelligence [Rusell and Norvig, 2021]. In fact, this work constitutes a seminal contribution for a research area which would become very popular many years later: artificial neural networks.

Also during the 1940s, Donald Hebb [1949] posed a hypothesis about learning based on the mechanism of neural plasticity. In this work, he proposes a very simple update rule to modify the strength of a connection between neurons, which is known today as "Hebb learning" and which is still used today for the so-called unsupervised learning.

In 1951, influenced by the paper from McCulloch and Pitts that was mentioned before, Marvin Minsky (see Figure 3), who at that time was a PhD student at the Math Department of Princeton University, built, with the help of Dean S. Edmonds (who was a PhD student at the Physics Department of Princeton University), the first physical model of a neural network. This device was called *Stochastic Neural Analog Reinforcement Computer* (SNARC) and it used 300 bulbs to simulate a network of 40 neurons which adopted Hebb learning. SNARC was able to simulate mice that successfully found their way out of a maze (this problem was originally posed by Claude Shannon). The simulation was visualized using "light arrangements". A circuit was reinforced each time that one of the simulated mice reached the goal. SNARC surprised its creators who would later say that the mice helped among themselves in the simulation, so that if one of them was able to find a good path, the other would follow it. In this PhD thesis, Marvin Minsky studied universal computation in neural networks, which was a topic that got him in trouble. His PhD committee was rather skeptical about Minsky's research work, since they were not sure that this could be considered as a PhD thesis in mathematics. However, it is known that the legendary Hungarian mathematician John von Neumann stepped in to settle this matter. When he was asked if he thought that Minsky's research work was part of mathematics, he answered: "if not now, one day this work will become part of mathematics". Over the years, Marvin Minsky became one of the most influential AI pioneers in the world.

III. A NEW DISCIPLINE IS BORN

In the 1950s, Princeton University had a faculty member who would become one of the founding fathers of artificial intelligence: John McCarthy (see Figure 4). After obtaining a PhD in mathematics from Princeton University in 1951, McCarthy worked there as a lecturer during two years. Then, and after a short period at Stanford University, he was hired as Assistant Professor at Dartmouth College. McCarthy convinced Marvin Minsky, Claude Shannon and Nathaniel Rochester to help him to bring together a group of American researchers interested in automata theory, neural networks and the study of intelligence. These efforts led to a two-month workshop that took place in the Summer of 1956 at Darmouth College. This event is considered the official birthplace of Artificial Intelligence, which is a term coined by John McCarthy in the grant proposal that he wrote to get funding to finance this workshop. In this proposal, he said the following [Russell and Norvig, 2021]:

FIGURE 4 John McCarthy.



SOURCE Wikipedia (available under a *Creative Commons* license).

> We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Darmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

McCarthy avoided to use the term "computer" when naming the new discipline, since he wanted to differentiate from the work done by Norbert Wiener on cybernetics, in which both analog and digital computers were adopted. However, clearly, McCarthy also planned to use computers for the new discipline that he had in mind [Russell and Norvig, 2021]. The conjecture indicated in the proposal would be later called "the hypothesis of the systems of physical symbols" and would become the foundation of the so-called "symbolic artifi-

cial intelligence". This hypothesis would be extended, reinforced and named by Allen Newell and Herbert Simon in their *ACM Turing Award* lecture in 1976 [Newell and Simon, 1976].

The attendants to this workshop would eventually write computer programs that would become milestones for the initial development of artificial intelligence: Arthur Samuel, Nathaniel Rochester, Ray Solomonoff, Oliver Selfridge, Trenchard More, Allen Newell and Herbert A. Simon. However, the last two were the super stars of the workshop, since by then, they had already developed a program called *Logic Theorist* [McCorduck, 2004], which had been used to prove most of the theorems from Chapter 2 of *Principia Mathematica* [Whitehead and Russell, 1910]. The *Logic Theorist* was able to find a shorter (and simpler) proof than the one indicated in the book for at least one problem.

IV. THE FIRST ACHIEVEMENTS OF ARTIFICIAL INTELLIGENCE

The early days of this new discipline was characterized by several success cases that generated a lot of enthusiasm among AI researchers. Many of these early AI projects were financed in the USA by the *Defense Advanced Research Projects Agency* (DARPA).

After the great success of the *Logic Theorist*, Allen Newell and Herbert Simon proposed the *General Problem Solver* (GPS), which was designed to imitate the protocols used by humans when solving problems [McCorduck, 2004]. The GPS was able to solve a very limited set of problems, but in them, this program considered both subgoals and possible actions, in an analogous way to humans. The success of the GPS and other further programs that operated as cognitive models led Newell and Simon to formulate the famous hypothesis of the physical symbols that states [Newell and Simon, 1976]: "a system of physical symbols has all the necessary and sufficient means to produce general intelligent actions.''

This hypothesis was the foundation for the so-called *symbolic* Artificial Intelligence and it would be challenged throughout the years. According to this hypothesis, the core idea behind AI is that any system aimed to show intelligence had to operate through the manipulation of data structures made of symbols [Russell and Norvig, 2021].

At IBM, Nathaniel Rochester and his colleagues made important contributions to Artificial Intelligence. For example, Herbert Gelernter [1959] built the *Geometry Theorem Prover*, which was able to prove theorems from Euclidean geometry which were considered difficult by several math students.

In 1952, Arthur Samuel (also working at IBM) wrote a series of computer programs to play checkers. Samuel was convinced that teaching a computer to play was a very fruitful task that would lead to the development of strategies to solve more general problems. In order to play checkers with a computer, Samuel devised search trees of the board positions that were reachable from the current state of the game. Since the computer available at that time had very limited memory, he implemented a technique that would be later called "alpha-beta pruning" [Sutton and Barto, 2018]. Additionally, instead of searching for individual paths that could reach the end of the game, Samuel developed a function that gave a score to the current board position at any given time. So, movements were based on such score function combined with a minimax strategy, which aimed to maximize the position of the player who was moving at that moment, while minimizing the score of the adversary. Samuel also proposed in 1955 several mechanisms to allow his program to improve its game level over time, after playing thousands of times against itself. In fact, the term *machine learnin*g was coined by Samuel to refer to this mechanism [Samuel, 1959].

John McCarthy left Darmouth College to move to the Massachusetts Institute of Technology (MIT) where he made three seminal contributions to AI in a single year (1958). The first, was the development of a high-level programming language called LISP, which would be used for many years to develop AI applications worldwide. The second contribution was his invention (in collaboration with other researchers from MIT) of *time sharing*, which was a mechanism that allowed to improve the access to the scarce and expensive computer resources available at that time. And the third and last of these contributions was a paper entitled "*Program with Common Sense*" [McCarthy, 1958], in which he described the *Advice Taker,* which was a hypothetical program that can be seen as the first full AI system. Same as the *Logic Theorist* and the *Geometry Theorem Prover*, McCarthy's program was designed to use knowledge that could search for solutions to arbitrary problems. However, unlike the two other systems, the *Advice Taker* aimed to gather general knowledge about the world. McCarthy showed, for example, how, through some simple axioms, his system could generate a plan to drive a car to the airport. Additionally, McCarthy showed that his program was designed to accept new axioms during its operation, which allowed it to perform actions in new areas without a need for any re-programming. Thus, the *Advice Taker* illustrated in a very accurate way, the hypothesis of the physical symbols, since it made clear the importance of having a formal representation of knowledge as well as the importance of using deductive processes to manipulate such knowledge.

It was also in 1958 that Marvin Minsky started to work at MIT. Although, at the beginning, Minsky collaborated with McCarthy (they were the founders of the Artificial Intelligence Lab of MIT in 1959), they eventually took different paths, since McCarthy had a strong interest in developing models of representation and reasoning based on formal logic, whereas Minsky was more interested in developing computer programs that led to practical applications.

In 1961, James Slagle wrote (in LISP) the first computer program for symbolic integration called SAINT (*Symbolic Automatic INTegrator*), as part of his PhD thesis [Slagle, 1961], which was supervised by Marvin Minsky at MIT.

FIGURE 5 Frank Rosenblatt.



SOURCE Cornell University division of rare and manuscript collections.

In 1964, Daniel G. Bobrow developed a computer program (written in LISP) called STUDENT, as part of his PhD thesis. This program was able to solve algebra problems at a high school level by taking their descriptions in natural language [Russell and Norvig, 2021].

In 1957, Frank Rosenblatt (see Figure 5) at Cornell University, developed the first hardware implementation of a multi-layered perceptron, which he showed to the press on June 23rd, 1960.

Rosenblatt described the details of his perceptron in a paper [Rosenblatt, 1958]. His model consisted of an input layer, a hidden layer (that was unable to learn) with random weights and an output layer with connections that were able to learn. Although Rosenblatt's goal was to produce a hardware implementation of his perceptron, he initially simulated it in software using an IBM 704 from the Aeronautics Lab at Cornell University. Rosenblatt extended his perceptron in a series of papers and a book [Rosenblatt, 1962], and was able to prove four theorems that laid the theoretical foundation for his proposal. In his book, Rosenblatt included several experiments using perceptrons with up to two layers, which he trained using "backpropagated errors". However, it is important to mention that he didn't use the algorithm that is known today as

backpropagation. In fact, Rosenblatt didn't have a general method to train multi-layer perceptrons. Although Rosenblatt's work was widely recognized and attracted a lot of interest from the press, this interest quickly decreased upon the publication of a paper written by Marvin Minsky and Seymour Papert in 1969. In this paper, they criticized the limitations of the perceptron for classification, without clarifying that the results reported in the paper were obtained using a single-layer perceptron [Kirdin *et al.*, 2022].

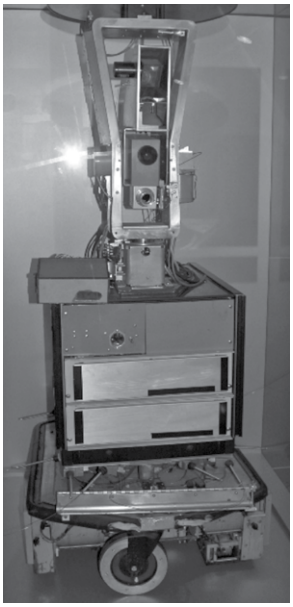
FIGURE 6 John Alan Robinson.



SOURCE Wikipedia (image in the public domain under a *Creative Commons* license).

In the same direction as Rosenblatt, there were two other important neural networks' achievements in the 1960s. First, Winograd and Cowan (1963) showed how a large number of elements could collectively represent an individual concept, giving rise to an increase in robustness and parallelism. Second, Bernard Widrow proposed, together with his PhD student Marcian Edward (Ted) Hoff, an adaptive algorithm based on a least mean squares filter (Widrow and Hoff, 1960). This algorithm allowed the development of the neural networks ADALINE (*ADAptive Linear NEuron*) and MADALINE (*Many ADALINE*). Later on, this work also led to the development of the backpropagation algorithm.

FIGURE 7 Robot *Shakey*.



SOURCE Wikipedia (available under a *Creative Commons* license).

In 1963, McCarthy created the Artificial Intelligence Lab at Stanford University. His plan to use formal logic to build an advanced version of the *Advice Taker* was catapulted by the resolution method proposed by John Alan Robinson (see Figure 6) in 1965 (Robinson, 1965), which consisted of a full algorithm that was able to prove theorems based on first order logic.

The research on AI that took place at Stanford at that time emphasized the development

of general purpose methods for logic reasoning. One of the most emblematic projects at that time was a robot named *Shakey* (see Figure 7). This robot was the first to be controlled using AI techniques and was the first to demonstrate the full integration of logic reasoning and physical action.

The rest of the decade of the 1960s was very fruitful in discoveries that would have a great impact on AI in the years to come. For example, Alexsey Ivakhnenko and Valentin Lapa developed the first deep learning algorithm for multi-layer perceptrons in the Soviet Union [Ivakhnenko and Lapa, 1967]. It was also in the decade of the 1960s that Lofti Zadeh published his seminal paper on fuzzy logic [Zadeh, 1965] and when the three main types of evolutionary algorithms were proposed: genetic algorithms [Holland, 1962], evolutionary programming [Fogel *et al.,* 1966] and evolution strategies [Rechenberg, 1965; Schwefel, 1965]. All of these contributions would originate a new discipline that would be later called "computational intelligence" [Kacprzyk and Pedrycz, 2015].

FIGURE 8 Joseph Weizenbaum.



SOURCE Wikipedia (available under a *Creative Commons* license).

FIGURE 9 Shunichi Amari.



SOURCE Wikipedia (available under a *Creative Commons* license).

Between 1964 and 1966, Joseph Weizenbaum (see Figure 8), who was a Professor at MIT, developed a program (written in a programming language called SLIP, which was created by Weizenbaum himself) named ELIZA [Weizenbaum, 1966]. This program simulated a conversation using a very simple method to recognize key words in a sentence. Although the program gave the impression of being able to understand a human conversation, it didn't even have a representation of the text that received as its input. ELIZA used different scripts. The most famous of them was called DOCTOR and simulated a psychologist from the Rogerian school (the name was adopted from the creator of this school, whose name was Carl Rogers). In the Rogerian school, the psychologist repeats with very slight variations, the words of his/her patient.

ELIZA was the first chatbot in history and raised lots of concerns in Weizenbaum when he realized that many people told it very intimate things. These concerns eventually transformed Weizenbaum into one of the main detractors of AI in his time. In 1976, he published a book entitled "*Computer Power and Human Reason: From Judgement to Calculation*" [Weizenbaum, 1976], in which he indicates that AI could become fashionable at some point, but that humans should never allow computers to make important decisions by themselves, since they lack important human qualities such as: compassion and wisdom. The core idea of his book is that there is a fundamental difference between deciding and choosing. For Weizenbaum, *deciding* is something that can be programmed in a computer, but *choosing* is something that must be done through judgement and not only based on calculations.

In 1967, Shunichi Amari (see Figura 9) reported the first multi-layer neural network trained with a stochastic gradient method [Amari, 1967]. This neural network was capable of classifying nonlinear separable patterns (something that Rosenblatt's perceptron was unable to do).

In 1968, Richard Greenblatt developed at MIT, the first computer program to play chess at a tournament level: *Mac Hack*. This was the first program to defeat a human player as well as the first in competing in a human chess tournament.

The *backpropagation* algorithm was independently developed by several researchers at the beginning of the 1970s. The first one was proposed by Seppo Linnainmaa in his Masters thesis [Linnainmaa, 1970; Linnainmaa, 1976]. Then, Paul Werbos independently developed the same algorithm in 1974 [Anderson and Rosenfeld, 2000] but due to several difficulties, he managed to publish it eight years later [Werbos, 1982]. However, it was a further paper (with one more independent discovery of this algorithm) [Rumelhart *et al.,* 1986] the one that popularized the use of the backpropagation algorithm.

V. THE FIRST AI WINTER (1974-1980)

The most prominent AI researchers in the early days of this discipline were too optimistic and made very ambitious predictions. For example, Herbert Simon said in 1957 that the world champion in chess would be defeated by a computer in 10 years. He missed by 30 years, since this prediction was fulfilled until 1997.

This optimism gave rise to the so-called "first AI winter" which occurred from 1974 to 1980. The name given to this period refers to the considerable reduction of the funding provided by the main government agencies that sponsored AI research in the USA. The main reason for this lack of financial support was that

most of the results that had been promised in the mid-1950s were still not a reality 20 years later [Russell and Norvig, 2021].

The highly optimistic prophecies that were made during the 1950s were not fulfilled due to several reasons. The first of them was that they were based on results obtained using very simple and low-scale examples. When attempting to scale up these early symbolic AI-based systems so that they could solve complex large-scale problems, they simply failed. The reason for this failure was that most of these early AI programs did not know anything about the specific domain in which they were applied. These programs performed relatively simple syntactic manipulations that were not necessarily scalable or portable to any real-world domain.

An example of these limitations occurred in the area of automated translation (called *machine translation*). In 1957, after the successful launch of the Russian satellite Sputnik, the National US Research Council attempted to speed up the translation of scientific articles written in Russian. For that sake, they asked the help of AI experts who used translation techniques based on simple syntactic transformations using grammars from Russian and English. These scientists adopted a simple word replacement algorithm that used an electronic dictionary. This was clearly an incorrect approach because text translation requires knowledge of the context of each sentence, since the same word may have several different meanings depending on the context in which is being used.

A report elaborated in 1966 by an advisory committee from the National Research Council indicated that it was not possible to use existing AI techniques to automate the translation of scientific texts from Russian to English. This report caused the cancellation of all the US government funding devoted to machine translation [Russell and Norvig, 2021].

Before complexity theory was developed within computer science, a common belief was that scaling up a program to larger problems was just a matter of having faster hardware and larger computer memories. The first AI programs used semi-enumerative approaches since they used to try many possible combinations to choose the correct one. These techniques only work properly in small problems but are not suitable for large scale problems since in those cases the number possible options grows exponentially, making impossible the application of semi-enumerative approaches. This is known as the "curse of dimensionality".

The third problem was the sort of structures adopted by the early AI techniques to generate intelligent behaviors. Many of these techniques had limitations inherent to the structure itself, which did not depend on the scale

of the problem. That was, for example, the case of the one-layer perceptron developed by Rosenblatt that was unable to classify classes that could not be divided by a straight line. Although this problem doesn't arise in multi-layer perceptrons, such perceptrons were rather uncommon in the early days of neural networks because their training was considerably more complex.

VI. THE RE-BIRTH OF AI (1980-1987)

In the period from 1980 to 1987, there were three important events that allowed a re-birth of AI research:

1. The success of the expert systems developed by Edward Feigenbaum.
2. The "fifth generation" project developed by the government and industrial sectors of Japan.
3. John Hopfield and David E. Rumelhart made seminal contributions that manage to re-activate the interest of scientists on the so-called connectionism (i.e., a new name given to neural networks). This interest would never go away.

FIGURE 10 Edward Feigenbaum.



SOURCE Wikipedia (available in the public domain).

Next, we will briefly discuss each of these events.

EXPERT SYSTEMS

In 1965, Edwad Feigenbaum (see Figure 10), who was a faculty member of the Computer Science Department at Stanford University, got interested in a problem that was proposed to him by Joshua Lederberg (a Nobel Prize winner in genetics): inducing the three-dimensional structure of chemical compounds from their mass spectrograms. Feigenbaum and his collaborators worked during 10 years (until 1975) in a system called Heuristic DENDRAL (its name was later shortened to DENDRAL) which was used to infer geometrical structures of complex chemical compounds given their chemical formula and the information of their mass spectrograms. DENDRAL was able to find structures that were previously unknown and these discoveries appeared in a series of papers published in the *Journal of the American Chemical Society*.

DENDRAL used rules that were obtained from chemists. These rules were related to the way in which a mass spectrometer fragmented compounds into sub-structures. Using the knowledge of the sub-structures, Heuristic DENDRAL was able to deduce the most plausible overall structure of the chemical compound. Later on, a new version of the program, called meta-DENDRAL, was even able to automatically deduce new rules from chemical data that Heuristic DENDRAL could use to improve its performance [Buchanan and Feigenbaum, 1978].

This work convinced Feigenbaum of the importance of providing AI programs of knowledge in the form of rules and procedures so that they could be properly guided for solving problems. Feigenbaum coined the name **Expert System** to denote the rule-based system adopted by DENDRAL.

Expert systems were the first AI tool that was successfully applied to the solution of real-world problems. Over the years, Feigenbaum developed other expert systems for areas such as medicine (MYCIN) and molecular genetics (MOLGEN).

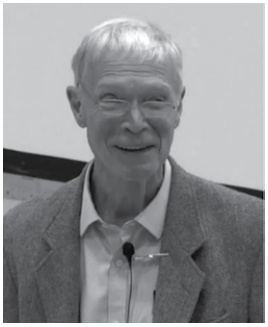
THE FIFTH GENERATION PROJECT

In 1981, the japanese government in collaboration with its industrial sector, introduced a project that was called "Fifth Generation Computer Systems". The original plan indicated a duration of 10 years and it was very ambitious but not too realistic. The original proposal was modified in 1982 by another one with more realistic objectives.

Among other things, this project indicated that fifth generation computer systems would have: (1) mechanisms for problem resolution and for performing inference, (2) very large knowledge bases, and (3) intelligent interfaces that would use natural language both in printed and in spoken form.

Clearly, these features would require a significant computer power and it was proposed to use hardware for managing parallel databases and software for the knowledge bases. Additionally, hardware would be adopted for performing parallel inference and for managing abstract data types and data flow machines. Furthermore, there would be

FIGURE 11 John J. Hopfield.



SOURCE Wikipedia (available under a *Creative Commons* license).

specialized hardware for speech recognition as well as interfaces that would operate with natural language.

Instead of choosing LISP for this ambitious project, the Japanese decided to adopt PROLOG, which is a programming language that was developed in France by Alain Colmerauer and Philippe Roussel at the beginning of the 1970s.

Although in its first 10 years, the fifth generation project was far from fulfilling its original goals [Feigenbaum and Shrobe, 1993], its inception attracted a lot interest from the US Government and raised concerns about the possibility that Japan would lead the computer and the intelligent systems industry. In order to prevent this from happening, the US Government decided to invest large amounts of money to finance research on parallelism at the beginning of the 1980s. This funding contributed to the re-birth of AI in a very significant way.

THE RETURN OF NEURAL NETWORKS

In 1982, the physicist John Hopfield (see Figure 11) published a paper [Hopfield, 1982] in which he proposed a type of recurrent neural network that is now known as Hopfield network (it is also called associative memory) in which neurons are modelled to interact according to the spin dynamics of physical systems such as magnetic materials. This sort of neural network is tolerant to noise in the data because Hopfield was able to prove that the training of his network can be understood in terms of the energy of the system. Therefore, the network iterates when there is noise, aiming to reach the state with minimum energy.

Hopfield played a key role in the re-activation of the interest on neural networks, because besides showing the way in which concepts from statistical mechanics could be applied to neural networks, this contribution somehow legitimated research on neural networks, since Hopfield was already a well-established physicist at that time. His work would be crucial to the re-birth of neural networks, which would come back to stay. In 2024, Hopfield received the Nobel Prize in Physics. The Nobel committee indicated that Hopfield "invented in 1982 a network that uses

FIGURE 12 David E. Rumelhart.



SOURCE Wikipedia (available under a *Creative Commons* license).

a method for saving and recreating patterns. He found inspiration in physics' models of how many small parts in a system affect the system as a whole. The invention became important in, for example, image analysis".

Also in 1982, David E. Rumelhard (see Figure 12) developed the so-called backpropagation algorithm in an independent way (as we mentioned before, this algorithm had been already proposed by other researchers, but Rumelhart was not aware of such proposals). Rumelhart showed the algorithm to Terrence Sejnowski who tried it and realized that it was faster than the Boltzmann machines developed by Sejnowski and Geoffrey Hinton (who was then a postdoctoral researcher under Rumelhart's supervision at the University of California in San Diego). Hinton became one of the first users of the backpropagation algorithm developed by Rumelhart, although at the beginning he was a bit reluctant to abandon Boltzmann machines. Nevertheless, after some time, Hinton realized that the backpropagation algorithm had a better performance than Boltzmann machines.

The publication of this algorithm (Rumelhart *et al.,* 1986) was another key event that contributed to the re-birth of neural networks. The publication of the book entitled *Parallel Distributed Processing* (Rumelhart and McClelland, 1987), edited by David Rumelhart and James McClelland was another key event for the neural networks community. In its two volumes, this book used the term "connectionism" to refer to neural networks and other similar learning models. The book originated from a strong debate between the symbolic AI community and the neural networks community. It is said that at some point, Geoffrey Hinton said that symbols were the "luminous ether of AI" (Russell and Norvig, 2021). In other words, Hinton said that symbols are a model of intelligence that doesn't work and that takes us in the wrong direction.

In 1983, Geoffrey Hinton published (together with Terrence Sejnowski) a seminal paper on Boltzmann machines, that were able to learn to recognize characteristic elements in a set of data. That same year, Hinton and Sejnowski applied Boltzmann machines for the first time to cognitive science problems (Hinton and Sejnowski, 1983).

In 2024, Hinton was the co-recipient of the Nobel Prize in Physics. The Nobel committee indicated that Hinton "used tools from statistical physics to create the Boltzmann machine, which can learn to recognize characteristic elements in a set of data. The invention became significant, for example, for classifying and creating images".

Hinton had several other seminal contributions to neural networks including: distributed representations, neural networks with a delay and the Holmholtz

machines. Because of his many contributions that span over 40 years, he is frequently called the "grandfather" of deep learning. In recent years, his warnings about the dangers of AI have become very famous.

In 1987, a program called NETtalk was released. This program was developed to learn to pronounce written text in English using a neural network (Sejnowski and Rosenberg, 1987). NETtalk was the outcome of the research work conducted by Terrence Sejnowski and Charles Rosenberg in the mid-1980s. The goal of this research work was to build simplified models that were able to provide some sort of evidence about the complexity of cognitive tasks involved in human learning. The aim was to develop a connectionist model that was able to learn to perform a task comparable to the ones developed by humans. The neural network adopted in this case had 3 layers and 18,629 adjustable weights, which was considerably large at that time. The training data consisted of 20,000 words that were provided with their corresponding pronunciation. This neural network was trained in two different ways: first using a Boltzmann machine and then using the backpropagation algorithm. The success of NETtalk led to the development of much more complex learning models based on neural networks.

VII. THE SECOND AI WINTER (1987-1993)

Although we already mentioned that the first AI winter took place during the 1970s, this expression was really coined in 1984, during a debate that took place during the annual meeting of the *American Association of Artificial Intelligence*. During this event, Roger Schank and Marvin Minsky (two senior AI researchers that had suffered the consequences of the first AI winter) warned a group of businessmen about the irrational growth of the enthusiasm in AI that took place at the beginning of the 1980s. Schank and Minsky warned that after so much enthusiasm there would be disappointment and they described a chain reaction similar to a "nuclear winter" that would start with the pessimism of the AI community followed by the pessimism of

FIGURE 13 A *LISP Machine.*



SOURCE Wikipedia (available under a *Creative Commons* license).

the press, which would be joined by several cuts to the funding to conduct AI research (Crevier, 1993). This prediction was indeed correct, since the second AI winter started only three years after this event, in 1987.

The enthusiasm for AI that reigned at the beginning of the 1980s seemed to be well justified. The first commercial expert system, called XCON, developed by Carnegie-Mellon University in 1978 for *Digital Equipment Corporation* (DEC) had a great success. This program was an assistant to process orders of VAX computers. The system allowed the automatic selection of system components based on the customer requirements. XCON started to operate in 1980 and only six years later, DEC estimated that it had produced savings of over 40 million dollars.

Several corporations followed DEC's example and started to install expert systems. By 1985, these companies had invested over 1 billion dollars in AI. Several companies were created to provide support to these AI applications. This included several software companies such as *Teknowledge* and *Intellicorp*, as well as hardware companies such as *Symbolics* and *LISP Machines Inc.*, which built specialized computers called *LISP Machines* (see Figure 13).

These computers used the LISP programming language as a native language (i.e., it was provided in hardware), which allowed them to execute LISP programs in a faster way. At that time, LISP was a very popular programming language to develop AI applications.

But in 1987, the market of *LISP Machines* collapsed. The reason was that recently created companies such as *Sun Microsystems* started to offer very powerful hardware. Soon, there were companies such as *Lucid* that offered LISP environments that were highly optimized for the workstations sold by *Sun Microsystems*.

The performance of the workstations from *Sun Microsystems* started to get in trouble to the developers of the *LISP Machines*. Soon, other companies such as Apple and IBM started to offer simpler and more popular architectures to execute LISP programs. By 1987, these new computers were already more powerful than the expensive *LISP Machines*.

Additionally, several rule-based engines for desk computers (e.g., CLIPS) started to become available. Thus, companies did not have any reason to pay more for computers that had special hardware to execute LISP in a more efficient way. Consequently, this industry that had an estimated value of 500 million dollars vanished in only one year (Crevier, 1993).

At the beginning of the 1990s, companies realized that their expert systems, although successful, had maintenance costs that were becoming too

high. Furthermore, these expert systems were difficult to update and were prone to errors if the rules were not carefully encoded. So, slowly, the companies devoted to the development of expert systems started to disappear or had to diversify to survive.

By 1993, more than 300 companies that developed AI-based solutions had gone bankrupt.

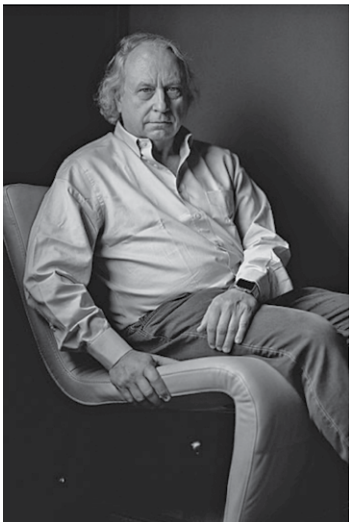
A second event that contributed to the second AI winter was the failure of the Fifth Generation project that the Japanese government had pushed so hard. After investing $850 million dollars over 10 years, the Japanese government realized that the results that they had obtained were still far from what they had originally envisioned.

In 1983, in response to the Fifth Generation project, DARPA started to finance AI research through the *Strategic Computing Initiative*. From its inception, this program was designed with pragmatic and reachable goals which even included general artificial intelligence (i.e., the development of AI software capable of performing different tasks with the same programming) as a long-term goal. This program was under the responsibility of the *Information Processing Technology Office* (IPTO) and also included funding for research in supercomputing and in microelectronics. By 1985, the US Government had already spent $100 million dollars in 92 projects that were being developed in 60 institutions (half of them in companies and the other half in universities and Government labs in the USA).

In 1987, Jack Schwarz was appointed as Head of the IPTO and that's when problems started. Schwarz did not like expert systems and applied brutal cuts to AI funding. He was convinced that AI was not the most appropriate research strategy and thought that DARPA had to invest in other technologies that he considered to be more promising (McCorduck, 2004).

Another interesting phenomenon that occurred during the second AI winter was the "rebellion" of several scientists who started to strongly question the "hypothesis of the physical symbols". The central

FIGURE 14 Rodney Brooks.



SOURCE Wikipedia (available under a *Creative Commons* license).

figure of this rebellion was Rodney Brooks (see Figure 14), who was a researcher at the Massachusetts Institute of Technology. Brooks published in 1990 a paper in which he strongly questioned what he considered to be the "dogma" of the hypothesis of the physical symbols (Brooks, 1990). In this paper, Brooks argues that symbols are not always necessary to produce AI, because the world is our best model. These claims led to the development of different approaches to develop AI, mainly in areas such as robotics, in which Brooks was particularly successful.

Also in 1990, Yann Le Cun (see Figure 15), at Bell Labs, used a convolutional neural network to recognize handwritten digits. This system was widely used in the 1990s to read ZIP codes and personal bank checks. Many consider this as the first practical application of neural networks that was really useful (Russell and Norvig, 2021). Yann LeCun received, together with Yoshua Bengio and Geoffrey Hinton, the *ACM Turing Award* in 2018, for their contributions to deep learning. This is considered to be the most important award granted to a computer scientist.

FIGURE 15 Yann LeCun.



SOURCE Wikipedia (available under a *Creative Commons* license).

VIII. THE PROMISED RESULTS (1993-2011)

In the period from 1993 to 2011, AI finally started to fulfill some of its old promises, mainly because of the greater computer power that was available, but also because of the development of more powerful algorithms.

Deep Blue was the first computer program able to defeat the world champion in chess, in 1997. In a now famous series of six matches, *Deep Blue* was able to defeat Garry Kasparov in two, it lost in one and it tied in three more. These matches became a landmark in AI history.

After this historical victory, IBM decided that *Deep Blue* would never play again and sent it to the Smithsonian Museum in Washington, D.C.

The development of *Deep Blue* (see Figure 16) started in 1985 at Carnegie-Mellon University and the project moved to IBM in 1989. *Deep Blue*'s algorithm depended on brute force and it was executed in a supercomputer IBM RS/6000 SP that had 32 processors designed to execute an expert system for

playing chess. *Deep Blue*'s algorithm was able to evaluate 200 million chess positions per second.

During the 1990s, a new paradigm emerged. It was called "intelligent agents". An intelligent agent is a system that perceives its environment and takes actions that maximize its probabilities of success.

But perhaps one of the most relevant events during this period was the incorporation of multiple disciplines (e.g., mathematics, economics, operations research and electrical engineering) into AI.

A book published by Judea Pearl in 1988 incorporated probability theory into AI (Pearl, 1988). This gave rise to new AI tools such as Bayesian networks, hidden Markov models, information theory, stochastic models and classical optimization.

FIGURE 16 Deep Blue.



SOURCE Wikipedia (available under a *Creative Commons* license).

In a paper published in 1994, Lofti Zadeh (see Figure 17) from the University of California at Berkeley, coined the term **Soft Computing** to include: fuzzy logic, neural networks, evolutionary algorithms, probabilistic reasoning and chaos theory (Zadeh, 1994).

Also in 1994, James Bezdek (Siddique and Adeli, 2013) used for the first time the term **Computational Intelligence** to refer to a system that deals with low-level data such as numbers and which does not use a traditional representation of knowledge (as required by symbolic AI). Bezdek would later indicate that Computational Intelligence is based on the use of *Soft Computing* techniques, while traditional AI uses *Hard Computing* techniques.

In 2003, the Institute of Electrical and Electronics Engineers (IEEE) approved that the IEEE *Neural Networks Council* changed its name into *IEEE Computational Intelligence Society*. This new Society indicated that *Computational Intelligence* included fuzzy logic, neural networks, evolutionary algorithms and hybrid systems that combine any of these techniques. Computational Intelligence has played a key role within AI during this century.

In 1997, there was another relevant event for the development of deep learning: Sepp Hochreiter and Jürgen Schmidhuber (see Figure 18) proposed the so-called *Long Short-Term Memory* (LSTM). This is a type of recurrent neural

network that is designed to mitigate a problem known as the "vanishing of the gradient", that avoided the training of deep neural networks using gradient methods or backpropagation (Hochreiter and Schmidhuber, 1997). This mechanism provides to neural networks a short term memory that can last during thousands of iterations (therefore the name). The intuition behind the architecture of an LSTM is to create an additional model in a neural network that is able to learn when to remember and when to forget relevant information. In other words, the neural network learns in a very effective way what information could need later on in a particular sequence as well as the moment in which this information is no longer required.

FIGURE 17 Lofti A. Zadeh.



SOURCE Wikipedia (available under a *Creative Commons* license).

Schmidhuber is a very controversial character, who has expressed in public that he and other AI pioneers have not received the proper credit for their contributions to deep learning and has indicated that Yoshua Bengio, Yann LeCun and Geoffrey Hinton have been unfairly favored.

In 2003, Bengio published a paper in which he used neural networks for natural language processing (Bengio *et al.,* 2003). Training a neural network for distinguishing sentences that have a meaning from those who don't have a meaning is a very difficult task because there are many different ways of expressing the same idea using natural language. This gives rise to the "curse of dimensionality" that had been mentioned before, which demands very large amounts of training data.

FIGURE 18 Jürgen Schmidhuber.



SOURCE Wikipedia (available under a *Creative Commons* license).

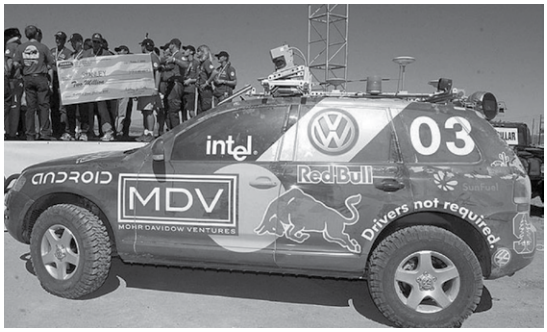
In the previously indicated paper, Bengio and his colleagues introduce a new type of representation for words (called *high-dimensional word embeddings*), which allowed a change of paradigm in natural language processing and automated translation.

Later on, Bengio proposed "attention" mechanisms that make a neural network

to focus only on a context that is relevant to a particular sentence. These mechanisms gave rise to a new generation of natural language processors as well as much more powerful automated translators.

In 2004, DARPA decided to establish a competition of autonomous vehicles with cash prizes. In its first edition, called *DARPA Grand Challenge*, no vehicle was able to finish the competition. But in 2005, a team from Stanford University was able to win this competition, since they were able to drive their car, called *Stanley* (see Figure 19) in an autonomous way, through the desert for more than 200 kilometers.

FIGURE 19 Stanley.



SOURCE Wikipedia (available under a *Creative Commons* license).

In 2007, the *DARPA Urban Challenge* was established. In this case, the winner was a team from Carnegie-Mellon University. This team managed to get their autonomous vehicle to navigate in an urban environment, following transit regulations and being exposed to the many risks associated to urban traffic for almost 90 kilometers.

With this, another AI promise was finally fulfilled.

IX. THE USE OF LARGE QUANTITIES OF DATA (2011-DATA)

At the beginning of this century, the access to large quantities of data and to a considerably larger computer power changed the way in which AI algorithms are developed. Advancements in deep learning (particularly, in convolutional neural networks and in recurrent neural networks) allowed great progress in image and video processing, as well as in text analysis and speech recognition.

The availability of large quantities of data (something known as big data) had made necessary the development of new processing models that allow the extraction of useful information from them. The availability of large volumes of data allowed deep neural networks to produce results that still surprise us today. Next, we will briefly describe a few examples of this.

ALEXNET

AlexNet is the name of a convolutional neural network architecture designed by Alex Krizhevsky in collaboration with Ilya Sutskever and Geoffrey Hinton (who was the PhD thesis supervisor of Krizhevsky). AlexNet competed in the *ImageNet Large Scale Visual Recognition Challenge* on 30 September 2012. AlexNet had an error of only 15.3% which placed it almost 11 points above the second place. AlexNet had 60 million parameters and 650,000 neurons. In the paper in which AlexNet is described (Krizhevsky *et al.,* 2017), its authors indicate that the depth of the model was essential to achieve these results. This required a high computational cost for which they used Graphics Processing Units (GPUs) during the training of the neural network.

AlexNet is considered a landmark regarding the success of deep learning in complex classification tasks.

GENERATIVE ADVERSARIAL NETWORKS

In 2014, Yoshua Bengio (see Figure 20) developed with Ian Goodfellow (who was Bengio's PhD student at that time) the concept of Generative Adversarial Networks (GANs) (Goodfellow *et al.,* 2014).

While most neural networks were designed to recognize patterns, a generative network learns to generate objects that are difficult to distinguish from among those available in the training set. This network is called "adversarial" because it is possible to train it to generate false outputs that are believable against another network that learns to identify false outputs. This allows a dynamic learning process inspired on game theory. This process is frequently used in unsupervised learning. In recent years, generative adversarial networks have been used, for example, to generate very realistic photos that contain non-existent people.

FIGURE 20 Yoshua Bengio.



SOURCE Wikipedia (available under a *Creative Commons* license).

ALPHAGO

AlphaGo is a program designed by *DeepMind Technologies* to play Go, which is a very ancient game with a search space larger than that of chess. To learn to play,

AlphaGo uses as its input a description of the board of Go which it processes using a deep neural network. Using the so-called "policy network", AlphaGo selects the next movement. Then, another network (called "value network") predicts the winner of the game. AlphaGo was trained with a very large number of matches between humans so that it could develop a game style similar to that of humans. Then, it played many times against itself and through the use of reinforcement learning, it managed to reach a very high gaming level. In October 2015, AlphaGo played its first match against a human: the three times European champion, Fan Hui. AlphaGo defeated Hui in the first matches that they played.

In 2016, AlphaGo faced the legendary Korean Go player Lee Sedol, winner of 18 world titles and considered as the greatest player of the decade. AlphaGo won 4 matches and lost only one in the tournament played in Seoul, South Korea, in March 2016. This tournament was seen by more than 200 million people and marked a new landmark in computer games. Unlike Deep Blue, AlphaGo uses machine learning techniques that have a much more general purpose and many consider this program to be a very important step towards general AI.

DeepMind Technologies is an English company that was created in 2010 by Demis Hassabis, Shane Legg and Mustafa Suleyman. In 2014, this company was bought by Google, which paid 400 million sterling pounds for it. Hassabis remained as its Vicepresident of Engineering.

The predecesor of AlphaGo was AlphaZero, which, in 2017, was able to defeat StockFish 8 which, at that time, was the world champion in computer chess (with an ELO of 3400). AlphaZero only required to be trained during 4 hours playing against itself.

In 2020, Demis Hassabis and John Jumper presented AlphaFold2. This neural network was able to predict the structure of virtually all the known proteins (200 million). This will provide scientists a better understanding of things such as antibiotic resistance and has allowed to create enzymes that decompose plastics.

This work made them win (together with David Baker) the Nobel Prize in Chemistry in 2024.

CHATGPT

ChatGPT is a generative AI chatbot designed by *OpenAI* which belongs to the family of language models known as *Generative Pre-Trained Transformer* (GPT). ChatGPT is based on a Large Language Model (LLM) called GPT-4o and can generate responses and can sustain conversations similarly to those held by a human.

ChatGPT was trained using humans and reinforcement learning adopting the supercomputer Azure, from Microsoft. It was released to the general public in November 2022, causing a great shock all over the world. Version 4.0 was released on March 14th, 2023 for the users of ChatGPT plus (the commercial version of the program). ChatGPT 5.0 was released on August 7th, 2025.

Some people say that ChatGPT played a key role in raising a general interest in AI and attracting higher investments and lots of interest from the general public in this discipline. But it has also raised concerns because its inappropriate use could give rise to disinformation and plagiarism, among many other negative things.

OpenAI was created in 2015 as a non-profit organization with the mission of ensuring that general AI would benefit humankind. Some of its founders are: Elon Musk, Sam Altman, Ilya Sutskever, Greg Brockman and other AI experts. One of its investors is Microsoft which in addition to providing billions of dollars, gave access to the company to its supercomputing platform (Azure).

THE DREAM OF GENERAL ARTIFICIAL INTELLIGENCE

Advancements in machine learning have generated a lot of interest in the so-called **general artificial intelligence**, which is defined as the ability to solve any problem with a single AI algorithm, instead of being limited to finding the solution to a particular problem.

Foundational models, which are large scale AI models, trained with very large quantities of unlabeled data that can be used for a large number of tasks, started their development in 2018. Models such as **GPT-3**, released by *OpenAI* in 2020 and **Gato**, released by *DeepMind* in 2022, have been described as seminal contributions leading to general AI.

In 2023, *Microsoft Research* tested **GPT-4** in a large variety of tasks and concluded that this system "could be seen, in a reasonably way, as a first version (still incomplete) of a general AI system".

Gato consists of a multimodal deep neural network developed by *DeepMind*, which uses the same network with the same weights to play Atari, to label images, to pile blocks with a robot arm, to chat with a human and to perform around 600 different tasks, making decisions based on the context. This neural network uses 1.2 billion parameters.

Flamingo is a language model with visual support that was developed by *DeepMind* and has 80 billion parameters. Flamingo takes an image or a video as an input and can sustain a conversation with a human (using a chat) based on

its contents. It can answer non-trivial questions about the image or the video, providing very accurate responses.

X. THE FUTURE OF ARTIFICIAL INTELLIGENCE

Clearly, the growing popularity of AI has generated a lot of optimism within the specialists in this area, who foresee great changes in the next few years, impacting several areas, such as:

- Education: Students could receive personalized contents for their lectures, tailored to their particular needs. AI could also determine the optimum teaching strategies for each student. Some people believe that, by 2028, the education system will be radically different from the current one.
- Health: Some people believe that it will be possible to have personalized health treatments (based on our DNA) in the future. Today, there are already some personalized therapies that have been generated using AI, but much more radical advancements are expected in the next few years.
- Finance: Natural language processing combined with machine learning could allow banks and financial advisors help customers in a more effective and efficient way in areas such as: fraud detection, finance planning, customer service, insurance services, and credit evaluation among many others.
- New ways of doing scientific research: Very significant scientific breakthroughs are expected in the next few years with the help of AI due to its great capacity to analyze very large quantities of data and to describe patterns and complex relationships among them.
- Transportation: Some people believe that autonomous vehicles will become popular both for commercial transportation and for public transportation, producing a radical change in the way in which goods and people and transported.

But some people also envision negative consequences associated to the misuse of AI in the future. For example:

- The end of privacy: AI systems require very large quantities of private information to personalize an application. Thus, the development of more sophisticated applications that make our everyday life easier will require of constant invasions to our privacy. This is a topic that has already generated very intense discussions in some countries.

- **Military applications:** Will someone develop AI-guided missiles or AI-designed bacteriological weapons that could kill millions of people? What are the ethical limits of AI?
- **Loss of jobs:** Many people are afraid of losing their jobs due to the popularity of AI algorithms that are able to perform many tasks which are performed today by humans. AI experts have indicated that only jobs in which creativity is not required are at risk, but it is still unclear what will be the real impact of AI in the economy of each country.
- **Environmental impact:** A relatively recent concern has been the environmental impact of AI applications that require large computational resources (e.g., generative AI). A recent study conducted by researchers at the University of California at Riverside indicated that, for example, Microsoft used approximately 700,000 liters of water during the training of GPT-3. This amount of water would be enough to produce 370 BMW cars or 320 Tesla cars. The reason for this environmental impact is that many AI applications require very large quantities of energy which, when transformed into heat, requires very large quantities of water to lower the temperature of the computer equipment involved. ChatGPT also consumes important quantities of water. An average conversation in ChatGPT involves 20 to 50 questions, which requires the equivalent of a 500 ml bottle of water. Although apparently small, this water consumption becomes relevant when we consider the millions of users of ChatGPT in the world.

XI. SHOULD WE BE AFRAID OF ARTIFICIAL INTELLIGENCE?

Since its inception, AI has generated fear not only among the general public, but also among experts (a remarkable example of this is Geoffrey Hinton, who has expressed in public his fear of the high power and potential misuse of AI).

The greatest fear is perhaps that AI algorithms go out of control and decide to exterminate humankind. Most of these fears are unsubstantiated, but it is clear that there are several risks associated with the use of new technologies and AI is, by no means, an exception. Because of this, the type of tasks in which AI is adopted is a key element to reduce risks. Clearly, AI, as any other advanced technology, can be dangerous if it is used in certain critical tasks (e.g., to develop biological weapons or to respond to a nuclear attack without human intervention).

It is however important to consider that computers and AI algorithms are not biological entities and have no awareness of themselves. So, in practical terms,

a human could unplug a computer or to delete a dangerous computer program that could potentially damage a human being.

Most of the current fears to AI have been caused by science fiction which tends to exaggerate technological advancements, enhancing their capabilities and, therefore, the risks associated with their use.

Nevertheless, it is important to have a legal framework that allows to regulate the use of our personal data as well as the use of potentially dangerous AI algorithms. This could reduce risks as well as a misuse of our personal data. Equally important is to introduce courses of ethics since elementary school, because the new generations will grow up with AI as an everyday technology and must have a clear picture of the risks associated to its misuse.

XII. CONCLUSIONS

We live today in an era in which we have very powerful computer programs that rely on very large amounts of data and on very significant computer power to perform tasks using AI algorithms, with impressive results. Some people believe that in the future, AI could help us to solve problems which are critical to our survival such as: climate change, drought, lethal diseases, among many others.

Clearly, subsymbolic AI (which uses numbers to store knowledge and relies on very large amounts of data) has defeated symbolic AI in the long term. However, one of the greatest challenges faced by current AI researchers is to be able to explain the way in which AI algorithms reach a solution. And symbolic AI could be very useful for this.

Evidently, the fear to AI will continue and may even increase over the years, as more powerful algorithms become available.

XIII. BIBLIOGRAPHY

[Amari, 1967] Amari, Shunichi (June 1967). "A Theory of Adaptive Pattern Classifiers", *IEEE Transactions on Electronic Computers*, **EC-16**(3): 299-307.

[Anderson and Rosenfeld, 2000] Anderson, James A.; Rosenfeld, Edward, Eds. (2000). *Talking Nets: An Oral History of Neural Networks*, The MIT Press, USA, ISBN 978-026251111-7.

[Bengio *et al.*, 2003] Bengio, Yoshua; Ducharme, Réjean; Vincent, Pascal ; Janvin, Christian (March 2003). "A Neural Probabilistic Language Model", *The Journal of Machine Learning Research*. **3**:1137-1155.

[Brooks, 1990] Brooks, Rodney A. (1990). "Elephants don't play chess", *Robotics and Autonomous Systems*, **6**(1-2): 3-15.

[Buchanan and Feigenbaum, 1978] Buchanan, B.G.; Feigenbaum, E.A. (1978). "DENDRAL and Meta-DENDRAL: Their Applications Dimension", *Artificial Intelligence*, **11**: 5-24.

[Crevier, 1993] Crevier, Daniel (1993). *AI: The Tumultuous Search for Artificial Intelligence*, Basic Books, New York, USA. ISBN 978-0-465-02997-6.

[Feigenbaum and Shrobe, 1993] Feigenbaum, Edward; Shrobe, Howard (1993). "The Japanese Fifth Generation Project: Introduction, survey, and evaluation", *Future Generation Computer Systems*, **9**: 105-117.

[Fogel *et al.,* 1966] Fogel, L.J.; Owens, A.J. ; Walsh, M.J. (1966). *Artificial Intelligence Through Simulated Evolution*, Wiley, New York, USA.

[Gelernter, 1959] H. Gelernter (1959). "Realization of a Geometry-Theorem Proving Machine". In *Proceedings of an International Conference on Information Processing*, pp. 273-282. UNESCO House.

[Goodfellow *et al.,* 2014] Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014). "Generative Adversarial Nets", In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014)*, pp. 2672-2680.

[Hebb, 1949] Hebb, Donald (1949). *The Organization of Behavior. A Neuropsychological Theory*, Wiley, New York, USA, ISBN 978-1-135-63190.

[Hinton and Sejnowski, 1983] Hinton, Geoffrey E.; Sejnowski, Terrence J. (June 1983). "Optimal Perceptual Inference", In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Washington, D.C. USA, IEEE Computer Society, pp. 448-453.

[Hochreiter and Schmidhuber, 1997] Hochreiter, Sepp; Schmidhuber, Jürgen (1997). "Long short-term memory", *Neural Computation*, **9**(8): 1735-1780.

[Holland, 1962] Holland, J.H. "Outline for a logical theory of adaptive systems", *Journal of the Association for Computing Machinery*, **3**: 297-314.

[Hopfield, 1982] Hopfield, J.J. (April 15, 1982). "Neural networks and physical systems with emergent collective computational abilities", *Proceedings of the National Academy of Sciences*, **79**(8): 2554-2558.

[Ivakhnenko and Lapa, 1967] Ivakhnenko, A.G.; Lapa, V.G. *Cybernetics and Forecasting*. Elsevier, New York, USA.

[Kacprzyk and Pedrycz, 2015] Kacprzyk, Janusz; Pedrycz, Witold, Eds. (2015). *Springer Handbook of Computational Intelligence*, Springer, ISBN 978-3662435045.

[Kirdin *et al.,* 2022] Kirdin, Alexander; Sidorov, Sergey; Zolotykh, Nikolai (2022). "Rosenblatt's First Theorem and Frugality of Deep Learning", *Entropy*, **24**(11): 1635.

[Krizhevsky *et al.,* 2017] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E. (24 May 2017). "ImageNet Classification with Deep Convolutional Neural Networks", *Communications of the ACM*. **60**(6): 84-90.

[Linnainmaa, 1970] Linnainmaa, Seppo (1970). *The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors*, Tesis de Maestría, University of Helsinki (en Finlandés).

[Linnainmaa, 1976] Linnainmaa, Seppo (1976). "Taylor expansion of the accumulated rounding error", *BIT Numerical Mathematics*, **16**(2): 146-160.

[McCulloch and Pitts, 1943] McCulloch, Warren S.; Pitts, Walter (1943). "A logical calculus of the ideas immanent in nervous activity", *The Bulletin of Mathematical Biophysics*, **5**(4). 115-133.

[Mata, 2006] Mata Alonso-Lasheras, Santiago (2006). *El hombre que demostró el cristianismo: Ramon Llull*. Ediciones Rialp, España, ISBN 978-84-321-6573-3.

[McCarthy, 1958] McCarthy, John (1958). "Programs with Common Sense". In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, pp. 75-91, Her Majesty's Stationary Office, London, UK.

[McCorduck, 2004] McCorduck, Pamela (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. Second Edition, A.K. Peters, ISBN 978-1-56881-205-2.

[Newell and Simon, 1976] Newell, Allen; Simon, Herbert A. (1976). "Computer science as empirical inquiry: symbols and search", *Communications of the ACM*, **19**(3): 113-126.

]Pearl, 1988] Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, California, USA. ISBN 978-0-08-051489-5.

[Rechenberg, 1965] Rechenberg, I. (1965). "Cybernetic solution path of an experimental problem", Technical Report Library 34 Translation 1122, Royal Air Force Establishment.

[Robinson, 1965] Robinson, John Alan (1965). "A Machine-Oriented Logic Based on the Resolution Principle", *Journal of the ACM*, **12**(1): 23-41.

[Rosenblatt, 1958] Rosenblatt, Frank (1958). "The perceptron: A probabilistic model for information storage and organization in the brain", *Psychological Review*, **65**(6): 386-408.

[Rosenblatt, 1962] Rosenblatt, Frank (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, D.C., USA.

[Rumelhart *et al.*, 1986] Rumelhart, David E. ; Hinton, Geoffrey E. ; Williams, Ronald J. (October 1986). "Learning representations by back-propagating errors". *Nature*. *323*(6088): 533-536.

[Rumelhart and McClelland, 1987] Rumelhart, David E.; McClelland, James L. (Editors). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition: Foundations*. MIT Press, USA, ISBN 978-0-26-268053-0.

[Russell and Norvig, 2021] Stuart J. Russell; Peter Norvig (2021). *Artificial Intelligence. A Modern Approach*, Fourth Edition, Pearson, USA, ISBN 978-0-13-461099-3.

[Samuel, 1959] Samuel, Arthur (1959). "Some Studies in Machine Learning Using the Game of Checkers", *IBM Journal of Research and Development*. **3**(3): 210-229.

[Schwefel, 1965] Schwefel, H.-P. (1965). *Kybernetische Evolution als Strategie der experimentellen Forschung* 36 *in der Strömungstechnik*, Dipl.-Ing. Thesis, Technical University ofBerlin, Hermann Föttinger--37 Institute for Hydrodynamics.

[Sejnowski and Rosenberg, 1987] Sejnowski, Terrence J. ; Rosenberg, Charles R. (1987). "Parallel networks that learn to pronounce English text", *Complex Systems*, **1**(1):145-168.

[Shannon, 1948] Shannon, Claude E. (Julio 1948). "A Mathematical Theory of Communications", *The Bell System Technical Journal*, **27**(3): 379-423.

[Siddique and Adeli, 2013] Siddique, Nazmul; Adeli, Hojjat (2013). *Computational Intelligence: Synergies of Fuzzy Logic, Neural Networks and Evolutionary Computing*. John Wiley & Sons. ISBN 978-1-118-53481-6.

[Slagle, 1961] Slagle, James (1961). A Heuristic Program that Solves Symbolic Integration Problems in Freshman Calculus, Symbolic Automatic INTegrator (SAINT), PhD thesis, Massachusetts Institute of Technology, June 1961.

[Sutton and Barto, 2018] Sutton, Richard S.; Barto, Andrew G. (1990). *Reinforcement Learning: An Introduction*. Second Edition MIT Press, Cambridge, Massachusetts, USA. ISBN 978-0-262-03924-6.

[Turing, 1936] Turing, A.M. (1936). "On computable numbers, with an application to the Entscheidungsproblem", *Proceedings of the London Mathematical Society*, **42**:230-265.

[Turing, 1950] Turing A.M. (1950). "Computing machinery and intelligence", *Mind*, **59**: 433-460.

[Weizenbaum, 1966] Weizenbaum, Joseph (January 1966). "ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM, 9(1): 36-45.

[Weizenbaum, 1976] Weizenbaum, Joseph (1976). *Computer Power and Human Reason: From Judgement to Calculation*, W.H. Freeman & Co., USA, ISBN 978-0716704638.

[Werbos, 1982] Werbos, Paul (1982). "Applications of Advances in Nonlinear Sensitivity Analysis", In R.F. Drenick and F. Kozin, Eds., *System Modeling and Optimization. Proceedings of the 10th IFIP Conference, New York City, USA, USA, August 31-September 4, 1981*, Springer-Verlag, Lecture Notes in Control and Information Sciences.

[Whitehead and Russell, 1910] Whitehead, Alfred North; Russell, Bertrand (1910). *Principia Mathematica*, Vol. 1, Cambridge University Press, Cambridge, UK.

[Widrow and Hoff, 1960] Widrow, B.; Hoff, M.E. (1960). "Adaptive switching circuits", in *1960 IRE WECON Convention Record*, pp. 96-104.

[Wiener, 1948] Wiener, Norbert (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*, MIT Press, Cambridge, Massachusetts, Estados Unidos, ISBN 978-0-262-73009-9.

[Winograd and Cowan, 1963] Winograd, S.; Cowan, J.D. (1963). *Reliable Computation in the Presence of Noise*, MIT Press, USA.

[Zadeh, 1965] Zadeh, L.A. (June 1965). "Fuzzy Sets", *Information and Control*, **8**(3): 338-353.

[Zadeh, 1994] Zadeh, Lofti A. (1 March 1994). "Fuzzy Logic, Neural Networks, and Soft Computing", *Communications of the ACM*, **37**(3): 77-84.

THE CONSTITUTION OF THE ALGORITHM AND THE RUPTURES IN DIGITAL SOCIETY

CONTENT

I. Crises, transitions and ruptures **II.** The ruptures in digital society **III.** The constitution as a cultural and civilisational factor **IV.** Technology as a cultural factor **V.** The constitution of the algorithm

> FRANCISCO BALAGUER CALLEJÓN

Professor of Constitutional Law. University of Granada, Spain

I. CRISES, TRANSITIONS AND RUPTURES

The concept of transition is a step forward from the crises we legal scholars have analysed many times in the first quarter of the 21st century, such as financial[1], democratic[2], health[3] and geopolitical crises[4]. This is because it not only focuses on the changes taking place, but also on their outcome and possible future scope. Crises place us in the present and make us reflect on the difficulties we face, which are reflected in constitutionalism, as was the case with the financial and health crises. While analysing these issues, we could not predict how they would develop, whether they would gradually subside or persist into the future. In truth, both crises ended, although they did not cease to have consequences and leave a significant mark on the evolution of our societies.

1 Cf. Balaguer Callejón, F., "Crisis económica y crisis constitucional en Europa", *Revista Española de Derecho Constitucional*, 98, 2013; Italian version: "Crisi economica e crisi costituzionale in Europa" *KorEuropa*, 1/2012; French version: "Crise économique et crise constitutionnelle en Europe", *Constitutions*, avril-juin 2013.

2 Cf. Balaguer Callejón, F., "Las dos grandes crisis del constitucionalismo frente a la globalización en el siglo XXI", *Revista de Derecho Constitucional Europeo*, 30, 2018. Italian version: "Le due grandi crisi del costituzionalismo di fronte alla globalizzazione nel XXI secolo", in *Passato, presente e futuro del costituzionalismo e dell'Europa*, a cura di F. Lanchester, CEDAM, 2019, pp. 59-82; Portuguese version: "As duas grandes crises do constitucionalismo diante da globalização no século XXI", *Espaço Jurídico Journal of Law [EJJL]*, 19 (3), 2018, pp. 681-702.

3 Cf. Balaguer Callejón, F., "Crisi sanitaria, globalizzazione e diritto costituzionale", in *Scritti in onore di Paolo Ridola, Rivista italiana per le scienze giuridiche*, 2020 and Balaguer Callejón, F., "Crisis sanitaria y derecho constitucional en el contexto global", *Teoría y Realidad Constitucional*, 46, 2020.

4 Cf. Balaguer Callejón, F., "Il futuro del costituzionalismo in un mondo (ri)globalizzato: mediazioni negative nella globalizzazione frammentata", *Nomos. Le attualità nel diritto*, 3/2023.

When we speak of 'transition', we take a step forward from the crisis because we can envisage the future in its essential features. In crisis situations, we are uncertain about the possible scope of change and cannot rule out the possibility that things will more or less return to normal. In transitions, however, we can predict the outcome and the evolution of the new order or 'new normal' that awaits us. While crises generate expectations about the scope of transformations and whether they will occur at all, transitions allow us to anticipate those changes and visualise their outcome with a certain degree of certainty.

Ruptures, on the other hand, have a different meaning to transitions. There isn't enough information to imagine the scope of the changes, but we can venture that they will be highly significant. We cannot predict the outcome because ruptures are not like bridges; they do not lead us to a place we could previously have envisaged. Ruptures are characteristic of revolutionary processes that can have a historical dimension, developing fully over the years to define a new social, economic or political order.

To understand the difference between a transition and a rupture, we can think of a door we have to open. In transitions, we know with greater or lesser certainty what we will find behind the door. It is something predictable, whose specific configuration we don't know, but we can guess. In ruptures, however, we don't know what will be behind the door. We simply don't know the path we will follow or where it will lead us.

Examining the historical evolution of constitutionalism, we find examples of both transitions and ruptures. The establishment of constitutionalism following the French Revolution in the 19th century was a moment of rupture that gave rise to the modern world, the scope of which was unforeseeable even by those who promoted it.

However, in the 20th century, the transition from the legal State to the constitutional State did not represent a rupture because it had already been foreshadowed by doctrinal and political debates during the interwar period. This configuration was influenced by the ideas of H. Kelsen on the normativity of the constitution, constitutional justice and pluralist democracy[5]; H. Heller on

5 Cf. Kelsen, H., *Reine Rechtslehre*, 2nd edition, Verlag Franz Deuticke, Wien, 1967 reprint; Kelsen, H., *Vom Wesen und Wert der Demokratie*, 1920, Spanish version of the 2nd edition 1929, *Esencia y valor de la democracia*, Labor, Barcelona, 1934 and Kelsen, H., "La garantie juridictionnelle de la Constitution (La Justice constitutionnelle)", 1928, Spanish version, *Anuario Iberoamericano de Justicia Constitucional*, 15, Madrid, 2011.

the social State[6]; and F. Neumann on the economic constitution[7], to name just a few relevant examples.

In the 21st century, the third wave of globalisation and the ongoing accelerated technological development do indeed represent a rupture. These transformations are related to the consolidation of a global market –despite tendencies towards the fragmentation of globalisation, whose scope is currently limited[8]– and the impact of technological development on all aspects of life, especially communication processes and our ever-increasing dependence on artificial intelligence.

We are entering a new era in which emerging global power factors are not only transforming politics, economics and science, but also interpersonal relationships and the cultural patterns that shape societies. Society is experiencing changes of historical significance that have never before occurred with such depth and speed.

We cannot predict what awaits us behind this open door, nor what the future holds for constitutionalism. The modern constitution is a *historical product*. Unlike religious documents, a constitution does not connect human beings to an authority that transcends them. Therefore, a constitution can remain valid as long as people believe in its authority. Constitutions, on the other hand, connect people to one another by establishing the agreed rules that make a free society possible.

During the historical rupture that gave rise to the modern world, the constitution served a specific purpose: to liberate the forces of capital and labour, thereby facilitating the establishment of a new economic and social order based on a legal State and the rule of law. Through the creation of national markets that replaced the estate system of absolutism, the constitution moved in the direction of history, in favour of progress. Jurists played a crucial role in this era, as it were they who paved the way for the modern world through legislation. This role became even more significant in the constitutional State that followed

6 Cf. Heller, H., "Rechtsstaat oder Diktatur?", 1929, Spanish version in Heller, H., *Escritos políticos*, Alianza Editorial, Madrid, 1985.

7 Cf. Neumann, F., *Über die Voraussetzungen und den Rechtsbegriff einer Wirtschaftsverfassung*, 1931, Spanish version, *Revista de Derecho Constitucional Europeo*, 9, 2008.

8 Cf. Balaguer Callejón, F., "Il futuro del costituzionalismo in un mondo (ri)globalizzato: mediazioni negative nella globalizzazione frammentata", *Nomos. Le attualità nel diritto*, 3/2023.

the legal State because national markets were regulated through the economic constitution, fostering competition –the equivalent of political pluralism in the economic sphere. The constitution reached its zenith as a necessary social and economic reference point. We can speak of *a golden age of constitutionalism*[9].

However, in the third wave of globalisation, the constitution plays no relevant role. In the face of new global legitimising factors (economic and technological), the constitution is moving against the tide of history, as its system of democratic processes and fundamental rights is perceived as an obstacle to the free operation of the global economy and technology. Furthermore, as the market is global rather than national, there is an asymmetry between the constitution's State-based scope of effectiveness and the global scope of market regulation. Just as the *Constitution* played a key role in the transformation to the modern world *through law*, *technology* is now doing the same *through algorithms*.

What will happen to the constitution in the future? Forty years ago, when I was a Visiting Professor in the United States, I saw a television commercial for Pepsi Cola that struck me. In it, a futuristic archaeologist was showing his students artefacts from the past, such as an electric guitar and a ball. Suddenly, a Coca-Cola bottle appeared. A student asked, 'Professor, what is this?' After staring at the bottle, he replied, "I have no idea". For some populist leaders of our time, the constitution is already a troublesome artefact. They would love nothing more than for some future archaeologist to give the same answer when presented with a copy of the Constitution of the United States, Argentina or any other country: 'I have no idea what it was'.

However, the Constitution is not yet part of history, much less of archaeology. It remains the most sublime manifestation of the development of our civilisation. We cannot assume that it will remain unchanged: to survive, it must adapt; it must learn to regulate new power dynamics; it must experiment with new ways of protecting rights and democracy in an increasingly hostile global context. As legal scholars, our work is to facilitate this adaptation, ensuring that the advances made possible by technological development and globalisation are commensurate with the dignity of the civilisational progress that the constitution has enabled.

9 Cf. Balaguer Callejón, F. "El final de una época dorada. Una reflexión sobre la crisis económica y el declive del Derecho constitucional nacional", in *Estudos em Homenagem ao Professor Doutor José Joaquim Gomes Canotilho*, Vol. II, *Constituição e Estado: entre Teoria e Dogmática*, Coimbra Editora, Coimbra, 2012, pp. 99-122.

II. THE RUPTURES IN DIGITAL SOCIETY

Today's world is essentially governed by algorithms designed by technological companies that configure communication processes largely based on their economic interests. From a constitutionalist perspective, these algorithms and the virtual reality they generate are causing major ruptures that could be characterised as follows:

1. The first rupture is that of the *cultural context of the constitution*, with the emergence of new legitimising factors corresponding to the global power factors of our time, particularly technological ones[10]. New cultural patterns and paradigms are emerging and transforming constitutional culture[11]. This results in the erosion of constitutional rights and institutions, making it difficult for the constitution to fulfil its functions of controlling power, guaranteeing rights, and channelling fundamental social and political conflicts.
2. The second rupture is that of the *constitution as a unitary cultural reference* (analogical constitution versus digital constitution). This split generates many other ruptures that affect the legal system and the system of sources of law, as well as constitutional rights. In general, the existence of a certain incompatibility between algorithms and the procedural dimension of pluralist democracy and constitutional law, as a public, plural, and participatory process, cannot be ignored[12].
3. The third rupture is between *physical and virtual reality*. The former is still regulated by the State through public law. The latter has a global scope and is essentially regulated by tech companies through private law. Therefore, not only are these two realities different in their configuration, but they are also different in their ordering: physical reality is essentially State-regulated and

10 Cf. Balaguer Callejón, F. "Constitution, démocratie et mondialisation. La légitimité de la Constitution face à la crise économique et aux réseaux sociaux", in *Mélanges en l'honneur du Professeur Dominique Rousseau. Constitution, justice, démocratie*. L.G.D.J., Paris, 2020.

11 Cf. Balaguer Callejón, F., "La constitución del algoritmo y las transformaciones culturales en la sociedad digital", in Caamaño, F./Jove Villares, D. (Dirs.), *Tecnologías abusivas y Derecho*, Tirant lo Blanch, Valencia, 2024. Italian version: "La costituzione dell'algoritmo e le trasformazioni culturali nella società digitale", *lceonline* (www.lceonline.eu), 3/2024.

12 Cf. Balaguer Callejón, F. "La constitución del algoritmo. El difícil encaje de la constitución analógica en el mundo digital", in Gomes, Ana Cláudia Nascimento; Albergaria, Bruno; Canotilho, Mariana Rodrigues (Coord.). *Direito Constitucional: diálogos em homenagem ao 80° aniversário de J. J. Gomes Canotilho*. Belo Horizonte: Fórum, 2021.

public, whereas virtual reality is essentially global and private[13]. Furthermore, virtual reality does not accurately reflect physical reality; rather, it distorts it and subjects it to the economic interests of the large tech companies designing and configuring it to suit their own needs[14].

4. The fourth rupture occurs in *the configuration of reality* itself, as a shared social perception of reality is destroyed by big technology companies. These companies are the new mediators controlling communicative processes, as opposed to traditional media as the old mediators. By promoting fake news and alternative realities, the new mediators create tension within reality, which has great destructive potential for the public sphere[15].
5. The rupture of *the economic constitution*, generated by globalisation and driven by technological development, is also an element to be considered insofar as it deprives the State of a very important ability to organise social life and submits it the major global players[16]. It also affects the status of fundamental rights, depriving them of their connection with the dignity of the person and orienting them towards rights that are more closely linked to the requirements of economic traffic (like consumers' and users' rights or the right to the protection of personal data) [17].

All these ruptures are very different from those that have historically shaped constitutionalism. Previous ruptures had to do with the revolutionary implementation of the constitution in the modern world or with its defence against dictatorial systems. All of them occurred *within the State* and in relation

13 Cf. Balaguer Callejón, F., "La cultura constitucional en la era digital", in Pérez Collados, J. M. (Coord.) *La cultura jurídica en la era digital*, Aranzadi, Cizur Menor, 2022.

14 Cf. Balaguer Callejón, F., "Inteligencia artificial y cultura constitucional", in Balaguer Callejón, F., Wolfgang Sarlet, I. (Dirs), Derechos fundamentales y democracia en el constitucionalismo digital, Aranzadi, Cizur Menor, 2023. Italian version: "Intelligenza artificiale e cultura costituzionale", in Francisco Balaguer Callejón - Ingo Wolfgang Sarlet (a cura di) *Diritti fondamentali e democracia nel costituzionalismo digitale*, *lceonline* Numero Speciale 2/ Maggio-Agosto), 2023.

15 Cf. Balaguer Callejón, F., "The Impact of the New Mediators of the Digital Age on Freedom of Speech" in Hindelang, S. (editor) Moberg, A. (editor) *YSEC Yearbook of Socio-Economic Constitutions*, Springer 2022.

16 Cf. Balaguer Callejón, F., "Costituzione economica e globalizzazione", *federalismi.it*, |n. speciale. 5, 2019, pp. 42-54.

17 Cf. Balaguer Callejón, F., "Los contextos ordinamentales de la tutela multinivel de los derechos", *lceonline 3/2023*, Italian version: "I contesti ordinamentali della tutela multilivello dei diritti", *La cittadinanza europea*, 2/2023.

to the organisation of State power. By contrast, the ruptures that are occurring in the 21st century are *outside the State* and shape a world in which the power of the State no longer reaches a large part of the reality previously ordered by it. We saw this with the financial crisis, which demonstrated the impossibility of recovering by the citizens a power that was located outside the State[18].

For these purposes, it is useful to compare some of the democratic revolutions within the State with the subsequent situation in eurozone countries during the financial crisis. In the case of the Portuguese Carnation Revolution, for example, and those that followed the fall of the Berlin Wall or the Arab Spring, the army or citizens went to the dictator's palace and overthrew him.

However, the democratic regression experienced by some eurozone countries during the financial crisis could not be resolved through internal means alone. Some governments fell, but only because of external pressures; never because of a citizen-driven democratic revolution. Where could citizens have gone to put an end to this democratic regression? Not to government palaces, as this would not have changed anything with a different government, nor to the central banks of Eurozone countries, as they lacked the power to solve the problems.

Power factors in our time, such as large digital companies, are increasingly dissociating themselves from the State and amassing power, including economic power, in relation to the State. We have seen this with the growth of tech companies during the health crisis. Some of them, such as Apple, achieved a stock market value that exceeds the GDP of countries like Italy and Brazil[19]. However, their ability to shape communicative processes and the public sphere through their internet applications is much more important than their economic value.

18 Cf. Balaguer Callejón, F., "Una interpretación constitucional de la crisis económica", *Revista de Derecho Constitucional Europeo*, 19, 2013. Portuguese version: "Uma Interpretação Constitucional da Crise Econômica", *Revista de Direito Público*, 54, 2013.

19 Apple was then the first company in the world to surpass a stock market value of three trillion dollars. Cf. *Reuters/El País*, "Apple se convierte en la primera compañía del mundo en superar los tres billones de dólares de valor en Bolsa", *El País*, 3/1/2022. Apple has already reached a stock market value of over $4 trillion, while Nvidia has surpassed $5 trillion. These two companies are currently ranked third and first respectively by stock market value worldwide, with Microsoft in second place. Alphabet, Amazon and Meta occupy the fourth, fifth and sixth positions, respectively. Cf. Millán, S. *El País*, 28 and 29 October 2025.

III. THE CONSTITUTION AS A CULTURAL AND CIVILISATIONAL FACTOR

From the French Revolution onwards, a concept of constitution was established that integrates all the legitimacies characteristic of the modern world. The constitution, as the law governing society and politics, incorporates the characteristics of a concept of law that, as García Pelayo[20] indicated, stems from developments that had previously occurred in the field of natural sciences, and which are intended to be applied to many other areas[21]. Law –and the constitution as the fundamental political law– is not only the expression of the general will[22] and, therefore, the highest democratic manifestation, but also the expression of an inherent rationality[23] that reflects, on the political plane, the characteristics of scientific law[24]. The constitution combines scientific and democratic legitimacy (understood at the time as not incompatible with census suffrage) with legitimacy derived from its role in progress and development, including the economic aspect.

Over two hundred years (from 1789 to 1989, to simplify, in two dates that changed the world), constitutionalism has modulated, but also reduced, the impressive flow of legitimacy that came from these diverse sources. The first of these, the scientific approach, was lost because of new philosophical ideas and political rea-

20 "The same role that scientific law plays in the world of the spirit will be played by legal law in the world of the State", García Pelayo, M., *Derecho constitucional comparado*, reprint of the 7th edition of 1961, Alianza Editorial, Madrid 1984, p. 70.

21 As I. Berlin indicates, with the development of the natural sciences, a new historical consciousness emerged, leading to the belief that it was possible not only to explain the past but also to predict the future. Thus, prophecy linked to religion was replaced by the philosophy of history and sociology. Cf. I. Berlin, "The Bent Twig: On the Rise of Nationalism" (1972), included in Berlin, I. *Sobre el nacionalismo. Textos escogidos*. Editorial Página Indómita, Barcelona, 2019.

22 Article 6 of the *Declaration of the Rights of Man and of the Citizen* of 1789.

23 Let us recall that for J.J. Rousseau, "'il ne faut plus demander à qui il appartient de faire des loix, puisqu'elles sont des actes de la volonté générale (...) ni si la loi peut être injuste, puisque nul n'est injuste envers lui- même", *Du contrat social ou Principes du droit politique*, 1762, Livre II, Chapitre VI.

24 For García Pelayo, this concept of law, typical of early constitutionalism, "signifies the expression, in the legal field, of a mental attitude dominant in the 19th century, which strives to encompass under the law all possible fields of the spirit", Cf. *op. cit.*, pp. 68 et seq.

lities that changed the concept of law and constitution[25]. The second approach, the democratic one, was shattered by the introduction of property-based suffrage and would ultimately be drowned in the tide of fascism during the interwar period. It would be recovered, however, with the refunding of constitutionalism that the normative constitutions represented after the Second World War.

In any case, the loss of scientific legitimacy was only partial because it was maintained through the concept of progress, with which normative constitutions were familiar. This idea of progress was also channelled through the economic constitution and the regulation of labour-capital relations contained within it. From this perspective, despite successive crises being resolved within the constitutional framework, economic legitimacy remained constant and were not perceived as systemic crises for constitutionalism until the most recent one in the 21st century.

In general, it can be said that, despite the loss of its original 'scientific' legitimacy, the constitution has been considered an instrument of progress and civilisational advancement linked to technical and economic development through the three essential functions it has fulfilled thus far. Two of these functions originate from French constitutionalism: the protection of rights and the control of power. The third function, which is derived from German and Italian constitutionalism, is the normative function of channelling fundamental social conflicts through law.

In the context of pluralist democracy, the constitution will add a new legitimacy that, in some respects, replaces the scientific legitimacy initially attributed to it as an expression of a universal and abstract reason. This new legitimacy is legal, that of Law, and it also incorporates the two prerequisites for the legitimacy of constitutionalism stemming from the French Revolution: the guarantee of rights and the control of power. From the diversity of reasons and beliefs will emerge the fundamental consensus that gives rise to normative constitutions. One of

25 The deterministic perception of history, derived from the scientific nature of law, would ultimately be shattered by Karl Popper's theory. Popper questions this determinism, arguing that it cannot prevent a democratic evolution of society through its own means, and instead proposes a principle of individual responsibility: Neither nature nor history can tell us what we should do. Facts, whether from nature or history, cannot decide for us or determine the ends we must choose. It is we who give purpose and meaning to nature and history (Popper, K. R., The Open Society and Its Enemies, 1945, Spanish edition, Paidós, Barcelona, 1989, pp. 489-490). According to Popper, historicism attempts to shift our responsibility onto history because historicism arises from a lack of faith in the rationality and responsibility of our actions (p. 490).

the essential functions of the constitution will consist precisely in preserving this diversity, guaranteeing the survival of political pluralism, which in the sphere of society is what biodiversity is in the sphere of nature.

Thus, in contrast to scientific legitimacy, there is legal legitimacy, and in contrast to limited democracy, there is pluralist democracy. Similarly, in contrast to laissez-faire economics, there is the economic constitution. The same could be said of many other areas: the social State versus the liberal State; the constitutional State of law versus the legal state of law, and so on. Despite all these changes they entail, it can be said that the sources of legitimacy for normative constitutionalism remain essentially the same as those of the first constitutionalism: The constitution accompanies the scientific and economic progress of societies, making it possible through the peace and stability that the law provides.

The belief that the constitution is a transformative and progressive force capable of fostering personal and social development projects would therefore remain constant until the financial crisis of the 21st century. Consider, for example, in its origin the American Declaration of Independence of 1776, which stated that "We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness".

Even the 2004 Draft Treaty establishing a constitution for Europe stated the following in its Preamble: "Convinced that, thus 'United in diversity', Europe offers them (the peoples of Europe) the best chance of pursuing, with due regard for the rights of each individual and in awareness of their responsibilities towards future generations and the Earth, the great venture which makes of it a special area of human hope". 'Happiness' and 'hope' represent a legitimacy that goes beyond what law or politics can offer. They imply a legitimacy that transcends these spheres and extends to society as a whole and to individuals' lives, where the constitution is one of the cultural references that define civilisation and humanity's progress.

However, the new forms of legitimacy that have developed rapidly in the 21st century are not necessarily associated with constitutionalism. In fact, they operate outside the constitution as a closed system with its own logic. This system can modulate or even materially reforming the constitutional system and can confront it and subject it to its demands. This occurs with both economic and technological legitimacy[26] which tend to converge. In both cases, a

26 Cf. Balaguer Callejón, F., "Crisis económica y crisis constitucional en Europa" cit. and "Una interpretación constitucional de la crisis económica", cit.

specific vocabulary is constructed that has a political dimension and expresses a particular worldview, with the aim of conditioning and structuring the constitutional order.

IV. TECHNOLOGY AS A CULTURAL FACTOR

Legitimacy based on progress has been a constant throughout human history and a cultural factor specifically linked to modern constitutionalism. Globalisation may have somewhat altered the terms of this relationship insofar as it has amplified insecurity and uncertainty. However, the idea of progress continues to exist as an essential instrument of legitimation. This idea of progress is closely related to scientific and technological advances because, in the context of a capitalist economy, technological legitimacy is largely intertwined with economic legitimacy. The economy also conditions the orientation and evolution of technological development. Nevertheless, technology possesses a functional autonomy over the economy insofar as, from a constitutional perspective, it can operate independently when faced with obstacles posed by constitutional rights or legal limitations derived from the constitution.

Certainly, in contrast to the revolutionary optimism of early constitutionalism, when the constitution and science were seen as driving the progress of society –the former being a manifestation of universal reason, as was the latter– many of the scientific discoveries of the 20th century foreshadowed the potentially disastrous nature of science, not only for progress, but for the survival of humanity itself. However, the historical conditions in which many of these advances occurred, whether under totalitarian regimes or amid war within democratic systems, did not imply an inherent conflict between constitutionalism and scientific and technological development. Naturally, a certain tension is inevitable, but this has generally been resolved within the framework of constitutional values and principles. Constitutional systems have adapted to address the challenges posed by scientific development, thereby safeguarding constitutional rights.

In the context of globalisation, the terms of the political and constitutional game are changing in the 21st century because of technological development[27].

27 As D. Rousseau, indicates “Au-delà même de la question du numérique, ce qui se joue en ce début de siècle, c’est la fin d’un monde fabriqué au xixe siècle et relookée au xxe”, Rousseau, D., “Le numérique, nouvel objet du droit constitutionnel”, *Les Nouveaux Cahiers du Conseil constitutionnel*, 2017/4 N° 57, p. 9

Globalisation has not only accelerated technological development but also changed the relationship between human beings and previously known dimensions of *space* and *time*. This is what is happening with constitutionalism. Its genetic relationship with space and time is very different from that which has been generated in the context of globalisation. Regarding *space*, modern constitutionalism has historically been linked to the State. It historically emerged as a movement that attempted to limit and control the power of the State and has adapted to social and political changes, always within the framework of the Nation-State. The projection of constitutionalism at the supranational level is generating significant transformations both in terms of the concept of a constitution and in terms of Nation-States themselves and their internal constitutional systems[28].

However, beyond these potential future transformations that could lead to a revitalisation of constitutionalism in the supranational sphere, the fact remains that many of the functions that the constitution could fulfil within the State are now largely ineffective against major global actors[29]. This ineffectiveness is particularly evident in the face of financial speculators and technology companies, which operate globally and encounter no restraint from State policies or legislation.

With regard specifically to the national public sphere, in which constitutionalism has developed until now, the changes introduced by technological development, especially through social networks and internet applications, are also highly relevant[30]. The progressive fragmentation and radicalisation of the

28 Cf. Balaguer Callejón, F., "El Tratado de Lisboa en el Diván. Una reflexión sobre constitucionalidad, estatalidad y Unión Europea", *Revista Española de Derecho Constitucional*, 83, 2008 and Balaguer Callejón, F., "Constitución y Estado en el contexto de la integración supranacional y de la globalización", in Miguel Carbonell, Héctor Fix-Fierro, Luis Raúl González Pérez and Diego Valadés (Coord.), *Estado constitucional, Derechos humanos, Justicia y vida universitaria. Estudios en homenaje a Jorge Carpizo*, UNAM, México, 2015, Tomo IV, Volumen 1, pp. 197-211.

29 In the words of D. Rousseau, "La nouvelle civilisation du numérique signe, en effet, la fin de l'État souverain national, la fin du contrôle de l'État sur l'économie et la société", *op. cit.*, p. 9.

30 Certainly, it is not possible to assess these changes solely from a negative perspective. As Julien Bonnet and Pauline Türk point out, "Les technologies numériques n'étant qu'un outil, c'est la façon dont elles vont être utilisées, développées et encadrées qui déterminera, dans l'avenir, leurs effets bénéfiques ou délétères, à moyen et long terme, sur la démocratie", Bonnet, J. and Türk, P., "Le numérique : un défi pour le droit constitutionnel", *Les Nouveaux Cahiers du Conseil constitutionnel*, 2017/4 N° 57, p. 18. To assess these effects, it would also be necessary to differentiate, as Coralie Richaud does, between collective and

public sphere, which is generated or at least enhanced by technology companies through the algorithms they design for internet applications, makes it more difficult to achieve a broad social consensus aimed at ordering society through constitutional instruments.

Regarding the *time*, the new rhythm set by technological development has created an understanding of time that is incompatible with certain historical features of constitutionalism. Consider constituent processes, for example, which aim to establish stable fundamental norms capable of political and social integration. The immediacy of response to political and social problems –a characteristic specific to the public sphere generated through social networks– makes it difficult to think in the terms in which constituent or constitutional reform processes have traditionally developed. The timeframe of constitutions, as we have known them until now, is inconsistent with the characteristics of the public sphere that has emerged in recent years through social networks[31].

Globalisation will generate a new structure, a new context, and a new way of thinking. We can see this by comparing Popper's thought with Bauman's: the open society versus the liquid society[32]. What for Popper was an opportunity –that is, this recovery of one's own strength in the face of determinism, this sense of personal responsibility, the very uncertainty we have about history, the fact that history is not written and depends on us[33]– for Bauman, describing the liquid society and the problems it currently faces in the context of

individual practice on social networks. Cf. Richaud, C., "Les réseaux sociaux : nouveaux espaces de contestation et de reconstruction de la politique ?, *Les Nouveaux Cahiers du Conseil constitutionnel*, 2017/4 N° 57, pp. 29-44.

31 However much social media also offers possibilities for participating in constituent and legislative processes. Cfr. Sales, E., "La transformation de l'écriture de la Constitution, l'exemple islandais", *Les Nouveaux Cahiers du Conseil constitutionnel*, 2017/4 N° 57, pp. 45-57, and Vidal-Naquet, A., "La transformation de l'écriture de la loi : l'exemple de la loi sur la République numérique", *Les Nouveaux Cahiers du Conseil constitutionnel*, 2017/4 N° 57, pp. 59-68.

32 Cf. Balaguer Callejón, F., "Europa como sociedad abierta en el contexto global" in *Europa, società aperta. volume I. Problemi attuali e prospettive future*, a cura di D'Amico, M., Liberali, B., Pérez Miras, A., Raffiotta, E., Romboli, S., Teruel Lozano, G., Editoriale Scientifica, Napoli, 2022.

33 So, if we want to remain human, then there will only be one path, that of the open society. We must proceed towards the unknown, the uncertain, and the unstable, using all the reason at our disposal, to secure the security and freedom to which we aspire, Popper, K.R. *op. cit.*, p. 216.

globalisation, is viewed with unease, understood as a threat, which society perceives as such and which makes it feel insecure. For Bauman, uncertainty is the fundamental problem of liquid society and cannot be overcome through reason or personal responsibility[34]. It is the sign of an era in which human relationships break down, the State can no longer solve problems that are global, and personal responsibility becomes a burden[35].

What has occurred in terms of social and cultural transformation is that globalisation and technological development have dramatically accelerated the conditions of historical time, and therefore, the coordinates of *time* and *space* that we controlled within the Nation-State during the 20th century no longer exist. Technology as a cultural factor has contributed to fostering a complicated environment for constitutionalism insofar as it has placed outside the State and the national public sphere *the capacity to generate cultural patterns* that significantly affect the configuration of societies.

34 "Social forms (structures that limit individual choices, institutions that guard repetitions of routines, patterns of acceptable behaviour) can no longer (and are not expected) to keep their shape for long, because they decompose and melt faster than the time it takes to cast them, and once they are cast for them to set. Forms, whether already present or only adumbrated, are unlikely to be given enough time to solidify, and cannot serve as frames of reference for human actions and long-term life strategies", Bauman, Z., *Liquid Times: Living in an Age of Uncertainty*, Polity Press Cambridge, 2007, p. 1.

35 For Bauman, "Interhuman bonds, once woven into a security net worthy of a large and continuous investment of time and effort, and worth the sacrifice of immediate individual interests (or what might be seen as being in an individual's interest), become increasingly frail and admitted to be temporary", *ibidem*, p. 2. Also: "The separation and pending divorce of power and politics, the couple that since the emergence of the modern state and until quite recently was expected to share their joint nation-state household 'till death did them part'. Much of the power to act effectively that was previously available to the modern state is now moving away to the politically uncontrolled global (and in many ways extraterritorial) space; while politics, the ability to decide the direction and purpose of action, is unable to operate effectively at the planetary level since it remains, as before, local" pp. 1 y 2. And: "The responsibility for resolving the quandaries generated by vexingly volatile and constantly changing circumstances is shifted onto the shoulders of individuals –who are now expected to be 'free choosers' and to bear in full the consequences of their choices" pp. 3-4.

V. THE CONSTITUTION OF THE ALGORITHM

The constitution of the algorithm is a way of analysing the digital reality from the perspective of the ruptures that it is generating and that have a constitutional dimension. It is also a way of proposing solutions to mitigate these ruptures and facilitate a constitutional response to *constitutionalise the algorithm* in the sense of promoting greater consistency of the algorithms with constitutional principles and values, but also to *digitalise the constitution*, updating it and adapting it to the new conditions of the digital era[36].

The digital world is uncharted territory for the constitution. This is not only because the format has changed and is no longer compatible with our previous conceptual framework, but also because the power factors of the digital world do not conform to constitutional provisions. For various reasons, the 'constitution of the algorithm' (in the dual sense of constitutionalising the algorithm and digitalising the constitution) will have to start from new bases if it wants to control these factors of power.

The constitution regulates a world that no longer exists in part or is socially irrelevant[37] –particularly regarding digital natives. Some constitutional precepts affecting areas of life today are outdated. This is not because the norms they contain should not be preserved and their effectiveness guaranteed, but because the State power that the constitution regulates is unable to do so in the digital world.

The problem is not only that the constitution addresses a part of reality that practically no longer exists –the analogue world– but also that it fails to regulate the reality that has been imposed, shaping a new type of society living in a digital world. This issue is not only related to the transformation of the fundamental rights established in the constitution, but also to the emergence of new rights

36 Cf. Balaguer Callejón, F., *La constitución del algoritmo*, 2nd edition, Fundación Manuel Giménez Abad, Zaragoza, 2023; Portuguese version: *A constituição do algoritmo*, Editora Forense, Rio de Janeiro, 2023; Italian version: *La costituzione dell'algoritmo*, Le Monnier, Mondadori, Milano, 2023.

37 The constitution was forged in an analog world and, as Juan Francisco Sánchez Barrilao rightly says, the digital world has largely changed its object. Cf. Sánchez Barrilao, J.F. "El derecho constitucional ante la era de ultrón: la informática y la inteligencia artificial como objeto cons- titucional", *Estudios de Deusto*, Vol. 64/2, Julio-Diciembre 2016, p. 239.

that must be regulated to provide legal certainty and protection against global agents that currently cause them significant harm.

Algorithms play an essential role in digital society, shaping the new reality spreading throughout the physical and virtual worlds. Algorithms are used to process data and organise reality. Sometimes this ordering is carried out directly on physical formats. In these cases, any legal issues that arise are usually related to the activity itself rather than the prior regulation carried out by the algorithm, which does not have a normative configuration but expresses an action of the computer processor that has been previously defined by human beings. If the definition incorporates inappropriate instructions that cause damage, those responsible for designing those instructions will be held accountable.

A different issue is that of algorithms that have *a normative configuration* and that move within the scheme of the sources of law, as they are inserted in public decision-making processes affecting constitutional rights. In these cases, it is necessary to consider the constitutional dimension of the algorithm from a double perspective. On the one hand, in relation to *the system of sources of law* and to the need to preserve the principles of the legal system, which do not allow for leaving areas of regulation that affect fundamental rights outside the representative circuit of pluralist democracy[38].

On the other hand, the constitutional dimension of the algorithm must also be considered in relation to *the system of constitutional rights*. This system cannot renounce the guarantee of rights simply because infringements occur through computer procedures using algorithms. An algorithm is an instrument designed for a specific purpose, used to improve processes and generate various kinds of returns (even if they are ultimately essentially economic). This purpose must be compatible with the constitution, as must the algorithms designed to achieve this compatibility.

The transformation towards a digital society is bringing about *a civilizational change* that leaves the constitution in a very difficult position regarding the power factors that are emerging in the new global order. It is not only about the new structures, procedures and techniques of digital society, which can be symbolised by the increasingly present algorithm in our lives. It is also about *new paradigms and cultural patterns* that are changing the way we think, as well as the values that inspire our societies.

38 Cf. Balaguer Callejón, F., *Fuentes del Derecho*, CEPC, Madrid, 2022.

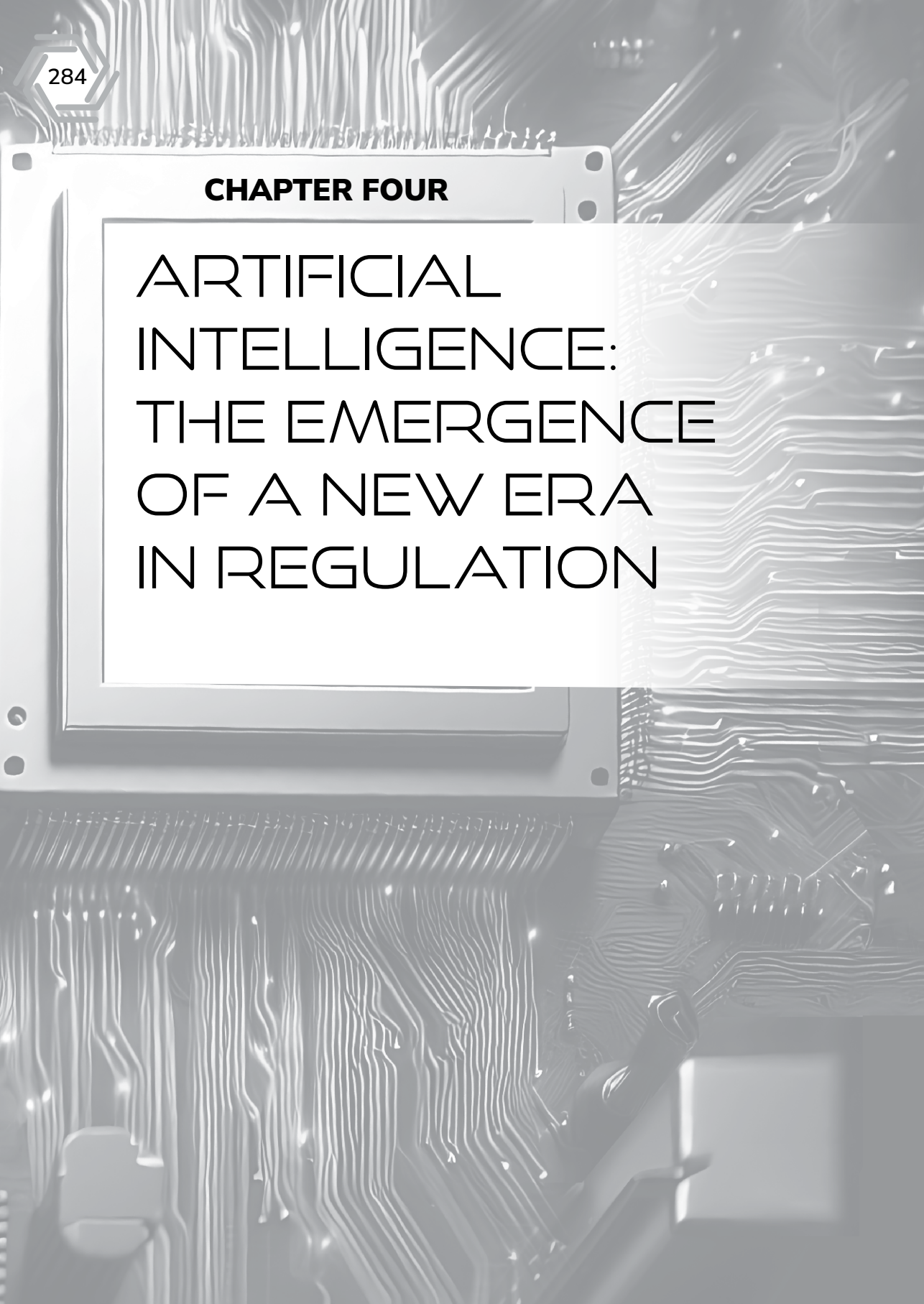
These transformations are not the result of random events or mere technological developments that serve to advance societies and contribute to general well-being. On the contrary, they are *new power factors* that are progressively destroying politics and severely weakening values and constitutional rights. Algorithms are not designed by themselves: even within the framework of machine learning, someone must consider how they will be used for purposes other than fulfilling their ultimate purpose of bringing economic benefits.

The growing incompatibility between the constitution and algorithms is not just a technical issue; it is a complex ideological construct that generates narratives aimed at increasing the profits of today's major global corporations. Rebuilding the order requires a 'constitution of the algorithm', which involves digitalising the constitution and constitutionalising technology, including artificial intelligence, to serve society and control the economic interests of large companies.

This 'constitution of the algorithm' will have to *redefine the role of new procedures* within the legal system. It will also need to *reorganise the entire system of fundamental rights* to establish a direct link between constitutional rights and new digital techniques. This will prevent fundamental rights from being violated on a massive scale or being subservient to market-related rights. The algorithm cannot serve as a means of infringing rights on a large scale. The constitution cannot continue to ignore the digital world, where much of economic activity and people's lives already take place.

CHAPTER FOUR

ARTIFICIAL INTELLIGENCE: THE EMERGENCE OF A NEW ERA IN REGULATION



THE EUROPEAN AI ACT –PROMISES, PERILS & TRADE-OFFS IN A COMPLEX REGULATORY ECOSYSTEM–

SUMMARY

Introduction **I.** Key features of the AIA **I.I** The AIA's Nature, Timeline, and Status **I.II** The AIA's Risk-based approach **I.III** Relevant actors **I.IV** The AIA's Twin aims and the need for Trade-offs **I.V** Support for innovation – The need to protect startups and SMEs **II.** Analysis/discussion **II.I** Potential for global impact **II.II** Potential to respond to future technological advances **II.III** Impact on the competitiveness of Europe **II.IV** Lessons for other countries **III.** Conclusions Bibliography

> TIMO MINSSEN
LL.D., LL.Lic., LL.M.., MICL & Dipl. Jur., Professor of Law & Director of the Centre for Advanced Studies in Bioscience Innovation Law (CeBIL) at the University of Copenhagen, Timo.Minssen@jur.ku.dk

> HANNAH LOUISE SMITH
BA, M.St, BCL, DPhil, postdoctoral researcher at CeBIL, hsm@jur.ku.dk

> SEBASTIAN PORSDAM MANN
MA (Cantab.), PhD (Cantab.), DPhil (pending, Oxon.), GDL (BPP University), Assistant Professor at CeBIL, Sebastian.porsdam.mann@jur.ku.dk

INTRODUCTION

The EU's AI Act[1] (AIA) is a groundbreaking piece of legislation that represents the world's first comprehensive legal framework for artificial intelligence (AI). Initially proposed by the European Commission in April 2021, it took years of negotiation with the European Council before a final agreement was reached with the AIA entering into force on August 1, 2024. Nevertheless, many of its provisions are not yet applicable and obligations will continue to be introduced until 2026, when it will become fully applicable, with a few exceptions lasting until 2030. For example, it is only since February 2025 that the prohibitions on AI systems that pose unacceptable risks have been applicable, yet the impact of this will be muted until August 2025, when provisions relating to governance, enforcement and penalties will also become applicable[2].

This chapter will highlight some of the AIA's features, including its objectives and provisions, as well as how it fits within a broader regulatory landscape. This provides the basis for a discussion of the promises, perils and trade-offs that the AIA is facing within a complex regulatory ecosystem and in an increasingly fierce international competition for global dominance.

1 European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AIA)*, Official Journal of the European Union, L 311/1 (July 12, 2024).

2 European Commission. *Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AIA)*. C(2025) 884 final (February 4, 2025), Brussels. https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act.

I. KEY FEATURES OF THE AIA

The AIA marks a seminal development in the legal oversight of AI technologies, having received approval on March 13, 2024. Although referred to as an "Act", it functions as an EU Regulation (Regulation EU 2024/1689), i.e., the strongest form of harmonizing legislation at the EU's disposal as it applies directly in all Member States without the need for further national implementation legislation. As such, once it enters into force, it will automatically become part of the legal framework of all 27 EU Member States. This places it on par with other significant regulations, such as the Digital Services Act[3], the Digital Markets Act[4], the Medical Device Regulation (MDR)[5], the Clinical Trials Regulation (CTR)[6], pending liability legislation[7], and the General Data Protection Regulation (GDPR)[8]. Together, these interrelated legal instruments are collectively shaping the European landscape for the regulation of AI and digital technologies.

I.1 THE AIA's NATURE, TIMELINE, AND STATUS

Despite the three years it took from the European Commission's initial proposal in 2021 to its agreement among EU Ministers in February 2024, the EU still

3 European Parliament and Council. *Regulation (EU) 2022/2065 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act)* (Text with EEA relevance), PE/30/2022/REV/1, OJ L 277, (October 27, 2022):1-102.

4 European Parliament and Council. *Regulation (EU) 2022/1925 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)* (Text with EEA relevance), PE/17/2022/REV/1, OJ L 265 (October 12, 2022):1–66.

5 European Parliament and Council. *Regulation (EU) 2017/745 on Medical Devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002, and Regulation (EC) No 1223/2009, and Repealing Council Directives 90/385/EEC and 93/42/EEC* (Text with EEA relevance), OJ L 117 (May 5, 2017):1-175.

6 European Parliament and Council. *Regulation (EU) No 536/2014 on Clinical Trials on Medicinal Products for Human Use and Repealing Directive 2001/20/EC* (Text with EEA relevance), OJ L 158, May 27, 2014, 1-76.

7 European Commission, *Commission Work Programme 2025*, last modified February 12, 2025, https://commission.europa.eu/strategy-and-policy/strategy-documents/commission-work-programme/commission-work-programme-2025_en.

8 See also: Mateo Aboy, Timo Minssen and Effy Vayena, *Navigating the EU AIA: implications for regulated digital medical products.* Nature Digital Medicine 7, 237 (2024). https://doi.org/10.1038/s41746-024-01232-3

achieved its goal of being the first region in the world to implement comprehensive AI legislation[9]. As an EU legislative instrument, the AIA comprises of both authoritative, binding articles and recitals that provide non-binding, albeit important guidance as to the AIA's core principles and aims. These encompass: (1) ensuring the safety of AI systems and safeguarding fundamental rights; (2) fostering innovation and the adoption of AI within the EU; (3) establishing legal clarity to promote investment and advancement in AI; and (4) managing risks linked to specific applications of AI, especially those that pose substantial threats to fundamental rights.

A staggered implementation timeline and severe fines

The AIA entered into force on August 1, 2024, but its provisions will apply in different phases until 2030. Non-compliance with the AIA could lead to significant penalties, with fines reaching up to €35 million or 7% of global turnover, underscoring the importance of early preparation[10].

This means that various actors must pay close attention to what is expected from them and when. For example, in accordance with Article 56 AIA, actors involved in general-purpose AI (GPAI) will have to be attentive to the evolving codes of practice, which are slated to apply by May 2025, and whose development is spearheaded by the AI Office with input from the AI board and a plethora of other stakeholders[11,12] Then, in August 2025, i.e., 12 months after the AIA entered into force, various obligations for providers of GPAI models will come into effect. Member States must then also appoint competent authorities to oversee the AIA, and the Commission will conduct it first review of the list of prohibited AI practices that will then occur annually. 18 months after the AIA's entry into force, i.e., on 2 February 2026, the Commission will implement acts relating to the provisions on post-market monitoring.

9 *Ibid.*

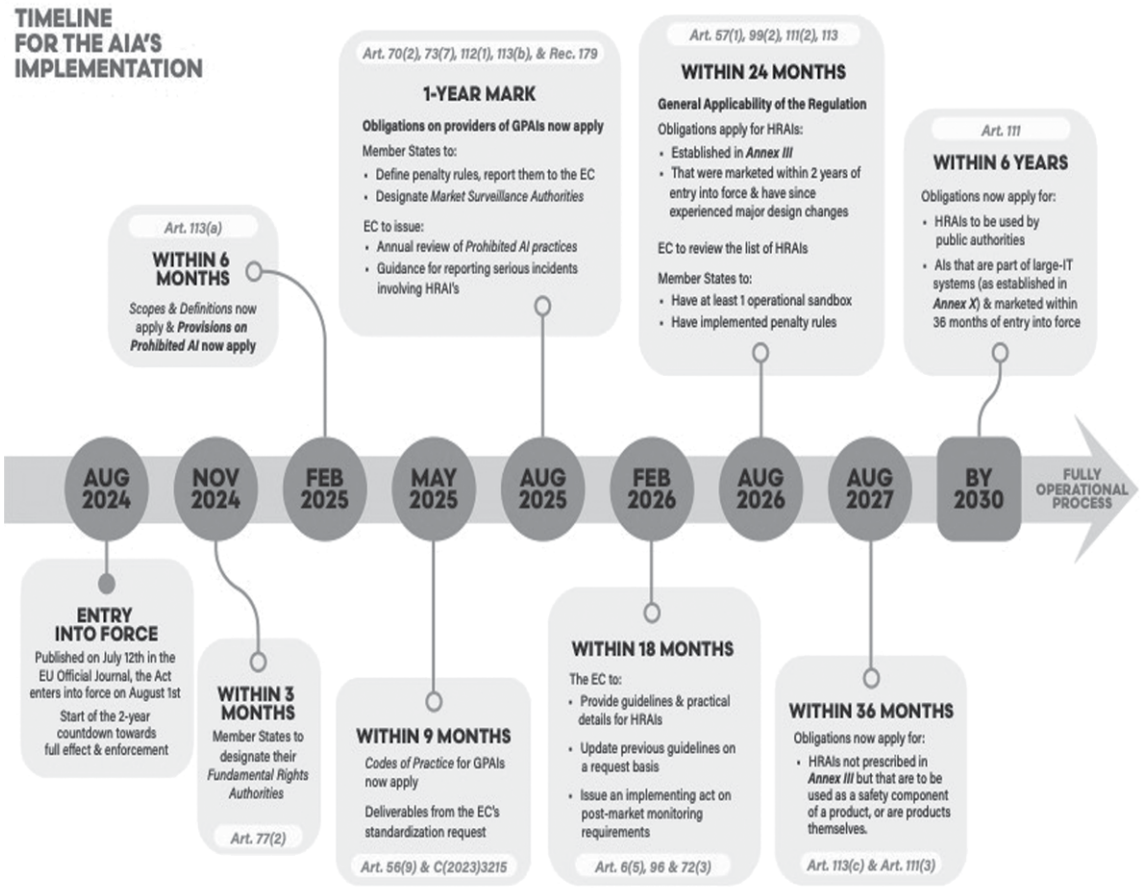
10 For an overview of the applicable penalties, see *Artificial Intelligence Act*, Art. 99.

11 For an overview on the various codes of practice under the AIA, see Jimmy Farrell and Afek Shamir, *An Introduction to Codes of Practice for the AI Act*, EU Artificial Intelligence Act, last modified October 30, 2024, https://artificialintelligenceact.eu/introduction-to-codes-of-practice/.

12 See also *Artificial Intelligence Act*, Art. 56.

Obligations for high-risk AI systems will first become applicable over the next 24 to 36 months, depending on the specific category of high-risk AI[13]. For AI systems associated with regulated digital medical products under the EU MDR/IVDR (Annex II), these obligations will start to apply 36 months after the AIA entered into force. The illustration below provides a basic overview of this implementation timeline:

FIGURE 1: Timeline for the AIA's implementation.



13 Joe Jones, *EU AIA: Next Steps for Implementation*, International Association of Privacy Professionals (January 2025), https://iapp.org/resources/article/eu-ai-act-timeline/.

Broad applicability and interactions

In contrast to sector-specific regulations that govern particular types of products, such as the EU MDR that sets requirements for the market access, safety, and efficacy of medical devices, the AIA's applicability is much broader. As AI technologies are increasingly integrated into a range of products and services that may already be regulated by sector-specific laws, one key challenge is to ensure any interactions between the AIA and other legislative instruments are harmonious. The AIA does not exist in isolation but interacts in various ways with existing legislative frameworks within the EU. Therefore, it is critical that these interactions foster an integrated regulatory environment so those developing and deploying AI understand the full range of their duties and responsibilities and individuals are aware of the protections offered to them.

Interaction with the GDPR and fundamental GDPR principles

The processing of personal data, as regulated by the GDPR, is increasingly undertaken by technologies that fall within the scope of the AIA. Therefore, the functioning and operation of these instruments cannot be evaluated in isolation. For example, both the AIA and GDPR emphasize the same principles, such as accountability and transparency. Moreover, the AIA requires that high-risk AI systems adhere to the principles of data protection as provided by the GDPR, including data minimization, purpose limitation, and storage limitation.

Nevertheless, the interplay between these instruments might not always be harmonious. For example, under Article 9 GDPR, special categories of data cannot be processed unless a specific exemption applies. In Case C-184/20[14] the EU's Grand Chamber held that if an organization could infer or deduce special categories of data, then data enabling such inferences should be treated as a special category of data. However, the AIA provides an explicit exemption from the prohibition contained in Article 9 GDPR as Article 10(5) AIA allows special categories of data to be processed where it is "strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems" so long as appropriate safeguards are in place. Hence, providers and deployers of AI systems may have to tread very carefully and examine what may be encompassed by the term "strictly necessary" to not fall foul of the GDPR's restrictions, especially after the court's expansive interpretation of special categories of data. The European Parliamentary Research Service

14 *Case C-184/20, OT v. Vyriausioji tarnybinės etikos komisija*, ECLI:EU:C:2022:601, August 1, 2022.

(EPRS) has acknowledged the potential tension between the GDPR's restrictive approach towards processing special categories of data and the AIA's provisions on processing such data to avoid discrimination. The difficulties in reconciling the two legislative frameworks means that the EPRS is even considering whether the GDPR requires reform[15].

Furthermore, the AIA builds on the concept of Data Protection Impact Assessments (DPIAs) that were established under Article 35 GDPR. DPIAs are required where processing is likely to result in a high risk to the rights and freedoms of individuals. However, where such an assessment is required by AI-related data processing, the European Data Protection Supervisor has expressed concerns that this interplay may not appropriately protect individuals. Organizations deploying high-risk AI systems are also required to conduct assessments to identify and mitigate risks related to safety and fundamental rights[16], but it may be difficult to the point of near impossibility for a provider to assess all the ways its system may be used. Therefore, an AI provider's assessment may fail to correspond to the risks that arise at the deployment stage and leave individuals exposed to harm[17].

The AIAs interaction with Product Liability and Safety Legislation

The AIA also interacts with existing (and pending) product liability[18,19] and safety legislation, including the General Product Safety Directive[20] and sector-specific

15 European Parliamentary Research Service, *Algorithmic discrimination under the AI Act and the GDPR* (February, 2025). https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2025)769509

16 *Supra* note 8.

17 European Data Protection Supervisor. *Opinion 44/2023 on the Proposal for Artificial Intelligence Act in the Light of Legislative Developments* (October 23, 2023):11. https://www.edps.europa.eu/system/files/2023-10/2023-0137_d3269_opinion_en.pdf

18 European Commission. *Liability Rules for Artificial Intelligence.* Last modified February 12, 2025. https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en.

19 European Parliament and Council. *Directive (EU) 2024/2853 on Liability for Defective Products and Repealing Council Directive 85/374/EEC* (Text with EEA relevance), PE/7/2024/REV/1, OJ L 2024/2853, (November 18, 2024).

20 European Parliament and Council. *Regulation (EU) 2023/988 on General Product Safety, Amending Regulation (EU) No 1025/2012 and Directive (EU) 2020/1828, and Repealing Directive 2001/95/EC and Council Directive 87/357/EEC* (Text with EEA relevance), PE/79/2022/REV/1, OJ L 135 (May 23, 2023):1-51.

regulations, such as the MDR[21]. The AIA does not establish a new liability regime and the recent withdrawal of a legislative draft on AI liability due to a lack of consensus[22] demonstrates the seemingly intractable tension between those favoring a precautionary approach and those favoring a permissive approach. The current geopolitical climate may support an approach that promotes the EU as an engine of AI innovation, but that is no guarantee of any future prospects. However, the AIA emphasizes that existing product liability laws continue to apply to AI systems[23], alongside any applicable liability regimes that exist within Member States. This means that manufacturers and developers of AI technologies must ensure that their products comply with safety standards and can still be held liable for certain damages caused by malfunctioning AI systems[24]. High-risk AI systems must also meet essential safety requirements, which align with existing product safety directives. This includes conducting risk assessments, ensuring human oversight, and providing necessary documentation for compliance[25].

The AIA's interaction with the Digital Services Act and Digital Markets Act

The AIA is also part of the broader legislative landscape that includes the Digital Services Act (DSA) and the Digital Markets Act (DMA), which regulate digital services and platforms within the EU[26].

The DSA addresses issues related to content moderation and the accountability of online platforms, which may intersect with AI technologies, and in particular GPAI, that are used for content curation and moderation[27]. For example, Article 13 AIA (Transparency and Provision of Information to Deployers), Article 27 AIA (Fundamental Rights Impact Assessment for High-Risk AI Systems), Article 50 AIA (Transparency Obligations for Providers and Deployers of Certain

21 Supra note 8.

22 PYMNTS, European Commission Withdraws Draft Rules on Technology Patents, AI Liability and Consumer Privacy (February 12, 2025). https://www.pymnts.com/news/regulation/2025/european-commission-withdraws-draft-rules-on-technology-patents-ai-liability-and-consumer-privacy/.

23 See e.g. the *Artificial Intelligence Act*, Recitals 9, 11, Art. 2(5), Art. 57(12), Art. 60(9).

24 *Ibid*. and *supra* note 3.

25 *Ibid*.

26 *Supra* notes 3 and 4.

27 Jordi Calvet-Bademunt and Joan Barata, *The Digital Services Act Meets the AI Act: Bridging Platform and AI Governance*, Tech Policy Press (May 29, 2024). https://www.techpolicy.press/the-digital-services-act-meets-the-ai-act-bridging-platform-and-ai-governance/.

AI Systems,) and Article 51 AIA (Classification of General-Purpose AI Models as General-Purpose AI Models with Systemic Risk) may impose additional requirements on AI systems employed in these contexts to ensure compliance with fundamental rights, ethical standards and transparency obligations[28].

The DMA focuses on ensuring fair competition in digital markets. While the AIA primarily targets the ethical and safety aspects of AI technologies, its interaction with the DMA is spurring discussions around how uses of AI could further distort competition, foster monopolistic practices, and conflict with the EU competition rules[29,30,31].

The AIA as a centerpiece of an emerging regulatory ecosystem

The AIA is thus the centerpiece of the EU's emerging regulatory ecosystem, which seeks to address the multifaceted challenges posed by AI. As the landscape evolves, several key themes and considerations are emerging.

As indicated above, the AIA emphasizes that AI developments must align with the EU's broader objectives of protecting fundamental rights and fostering trust in technology. This is supported by a focus on ethics that is interpreted through the lens of EU protections for individuals' rights and freedoms. This can be found in the European Commission's guidelines on trustworthy AI, which advocate for principles such as fairness, accountability, and transparency[32]. In the same vein, several recitals and articles in the AIA call for the establishment of guidelines and best practices for AI classification and development[33,34], and encourage organizations to adopt ethical frameworks that align with EU values.

28 *Ibid.*

29 European Union. *Consolidated Version of the Treaty on the Functioning of the European Union (TFEU)*, Articles 101 and 102. Official Journal of the European Union C 115 (May 9, 2008):94-95.

30 For further details, see, e.g., European Commission, *High-Level Group on the Digital Markets Act: Public Statement on Artificial Intelligence*, Digital Markets Act (May 22, 2024). https://digital-markets-act.ec.europa.eu/high-level-group-digital-markets-act-public-statement-artificial-intelligence-2024-05-22_en

31 Rupprecht Podszun, *From Competition Law to Platform Regulation: Regulatory Choices for the Digital Markets Act*, Economics 17, no. 1 (2023): 20220037, https://doi.org/10.1515/econ-2022-0037.

32 European Commission, *Directorate-General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI*, Publications Office (2019). https://data.europa.eu/doi/10.2759/346720.

33 See also Article 6 of the *Artificial Intelligence Act*, Article 6.

34 *Supra* note 2.

Stakeholder engagement is crucial in that regard, with input from civil society, academia, and industry being the most effective way of ensuring that diverse perspectives are considered in the development of AI technologies[35,36].

The Importance of International Cooperation

The global nature of AI technology necessitates international cooperation and a consensus on the best governance approaches as regulatory divergences may lead to trade barriers. The EU's AIA, as the first legislative instrument governing AI technologies, stands an excellent chance of serving as a reference point for other jurisdictions considering AI regulation. The EU's emphasis on a unified regulatory approach seeks to mitigate the risks of trade barriers and burdensome compliance requirements, promoting a level playing field for businesses that, most importantly, is most aligned to its rules and preferences. Therefore, the EU should continue to build links with international organizations (e.g., OECD, ISO) in the development of global standards for AI that promote ethical practices, ensure interoperability, and protect core values[37].

Adaptability and Future-Proofing

The rapid pace of AI innovation necessitates a regulatory framework that can adapt to changing technologies and market dynamics. Under Article 112, the AIA incorporates mechanisms for periodic review and evaluation, ensuring that it remains relevant and effective. The AIA also includes provisions for ongoing monitoring and evaluation of AI systems, as evidenced by Article 89, enabling regulators to respond to emerging risks and challenges. The tiered risk-based approach allows for flexibility in regulatory requirements, enabling innovation while maintaining safety and ethical standards. But there is much debate on whether the regulatory flexibility embedded in the AIA is enough, with some already finding its requirements to be outdated[38].

35 Gail E. Crawford *et al.*, *EU AI Office Seeks Stakeholder Engagement in Drafting AI Act Implementing Rules*, Latham & Watkins LLP (July 31, 2024). https://www.lw.com/en/insights/eu-ai-office-seeks-stakeholder-engagement-in-drafting-ai-act-implementing-rules.

36 International Association of Privacy Professionals, *EU AI Act Stakeholder Map*. IAPP (May 2024). https://iapp.org/resources/article/eu-ai-act-stakeholder-map/.

37 Overseas Development Institute. *International AI Governance Must Be Truly Global*, ODI (October 25, 2023). https://odi.org/en/insights/international-ai-governance-must-be-truly-global/.

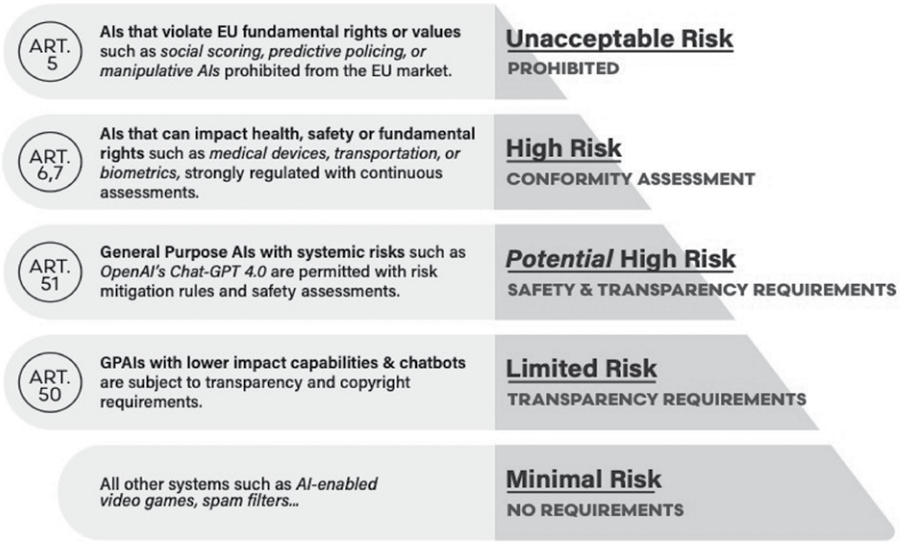
38 Sandra Wachter, *Limitations and Loopholes in the EU AIA and AI Liability Directives: What This Means for the European Union, the United States, and Beyond*, Yale Journal of Law & Technology 26, no. 3 (2024):685.

I.II THE AIA's RISK-BASED APPROACH

The AIA uses a risk-based approach and classifies AI systems into five tiers according to whether they pose unacceptable, high, systemic, limited, or minimal risks. The legal obligations differ according to the categorization to provide "proportionate and effective" rules that are tailored to the "intensity and scope" of any potential risk. Article 3(2) defines the concept of risk used in the AIA as "the combination of the probability of an occurrence of harm and the severity of that harm".

Article 5 AIA provides a list of AI systems that pose unacceptable risks and are prohibited as a result. These include AI systems that manipulate individuals to circumvent their free will (Article 5(1)(a)), make criminal risk assessments and individual predictions with no human involvement (Article 5(1)(d)), and the use of emotion recognition in the workplace or educational settings except if used for medical or safety reasons (Article 5(1)(f)), among others.

FIGURE 2: The AIA's Risk-Based Approach



Article 6 distinguishes between two types of high-risk AI systems. One type, listed in Annex I, concerns products that are already subject to ex ante conformity assessments due to existing EU harmonization legislation on health and safety. These products are AI systems that are either intended to be used as the safety component in a product or the AI system itself is the product (Article 6(1)(a)). The other type is a set of applications listed in Annex III, which provides eight

domains that cover biometrics, critical infrastructure, education and vocational training, employment, worker management, access to and enjoyment of essential private services and essential public services and benefits, law enforcement, migration, asylum, and border control management, and the administration of justice and democratic processes. These applications are lawful where certain provisions within the AIA are followed. If an AI system falls within Annex III, it is for the AI provider to assess whether the AI system does pose a significant risk to health, safety, and fundamental rights and only a positive assessment will trigger the obligations associated with high-risk systems. Article 6(3) sets out criteria that exempt AI systems from the obligations imposed upon high-risk systems. Chapter III sets out a substantial number of obligations that are imposed upon high-risk AI systems, such as those relating to risk management systems (Article 9), data governance measures (Article 10), technical documentation (Article 11), transparency provisions (Article 13), human oversight requirements (Article 14), and an appropriate level of accuracy, robustness, and cybersecurity (Article 15). Providers of high-risk AI systems must also register these systems in a public database (unless it is used for law enforcement or migration purposes), are subject to post-market monitoring by providers (Article 72) and market surveillance authorities (Article 74), and may require a fundamental rights impact assessment (Article 27). Furthermore, providers of high-risk AI systems must establish a quality management system that includes a strategy for compliance, including the need to respect fundamental rights (Article 17).

Article 50 imposes certain transparency obligations on certain AI systems. This encompasses AI systems that directly interact with individuals, produce synthetic content, or entail emotional recognition systems that are not captured by Article 5 and thus completely prohibited. These AI systems must ensure any user is informed they are interacting with an AI system, unless it is obvious to a reasonable person or where they are used to detect, prevent, investigate, or prosecute criminal offences.

A late addition to the AIA's risk classification was Article 51, which outlines the criteria for defining AI systems as general-purpose AI systems that may pose systemic risks. Recital 110 gives examples of such risks and includes "any actual or reasonably foreseeable negative effects in relation to major accidents, disruptions of critical sectors and serious consequences to public health and safety; any actual or reasonably foreseeable negative effects on democratic processes, public and economic security; the dissemination of illegal, false, or discriminatory content". Such AI systems are incongruent with a static risk classification approach, as their risk varies enormously depending upon the

prompts given by their human users. Article 53 outlines the obligations for providers of general-purpose AI models that do not pose systemic risks. The obligations include the need for up-to-date technical documentation that includes data on the model's training and testing process and results of its evaluation (Article 53(1)). Providers must also make available certain information and documentation to providers who intend to integrate the general-purpose AI model into their own AI systems (Article 53(1)(b)) unless the model is released under a free and open-source license (Article 53(2)). The public release of large language models with the ability to generate novel content, such as ChatGPT, triggered this inclusion.

The AIA implicitly classifies all other AI systems to be of minimal risk. Therefore, it imposes no obligations upon such systems but does encourage a range of stakeholders, including the European Commission, Member States, and the AI Office as the supervisory body created by the AIA, to facilitate the creation of codes of conduct (Article 95). The purpose of these instruments is to encourage voluntary compliance with the high-risk requirements, even where an AI system does not fall within the scope of Annex I or Annex III of the AIA and is thus not technically a high-risk AI system.

A regulatory approach premised upon risk is a potentially advantageous one. For example, calibrating the level of obligations based on the risk posed by AI systems helps to reduce the risk of over- and under- regulating these technologies. This provides clarity to AI providers who can accordingly tailor their research and development to align with their capacity for compliance. This will help to reduce barriers to innovation, especially for AI systems judged to not pose an unacceptable or high risk. Even for high-risk systems, prior knowledge of the obligations will allow AI producers to effectively allocate resources between developing AI systems and compliance requirements. However, the classification system does not provide absolute clarity. For example, high-risk AI systems may nevertheless be exempt from the requirements relating to such systems where they fulfil the criteria in Article 6(3). Nevertheless, these criteria have been criticized as "very vague and far-reaching"[39] that may impact negatively upon the certainty that the AIA's risk-based approach was supposed to ensure.

A risk-based regulation also aligns with other EU instruments that adopt a risk-based approach and may apply to AI systems, such as the GDPR, DSA,

39 *Ibid.*

DMA, and the MDR. A risk-based regulatory approach to AI has also been adopted by international instruments, including the G7's International Guiding Principles on AI[40] and the OECD's Framework for the Classification of AI Systems[41]. This move towards international harmonization facilitates AI's advancement by promoting a more streamlined and coherent regulatory environment. This allows AI providers to dedicate a greater proportion of their resources to research and development, rather than navigating multiple, potentially conflicting, regulatory sources.

The risk-based approach within the AIA, however, comes with various challenges. *First,* it assumes that, prior to deployment, it is possible to determine which category an AI system would fall under. The feasibility of this, given the "black box" nature of many AI systems, is questionable. Ongoing developments in the technologies and uses of AI systems, as demonstrated by the rapid rise of general-purpose AI models, may require ongoing monitoring and a re-evaluation of its classification. This promotes the protection of individuals but also undermines the clarity and certainty that a risk-based approach should foster, which may impede innovation and decrease trust in AI systems and run counter to the AIA's stated aims.

Second, there has also been significant critique of the AIA's concept of risk. As the AIA is part of the New Legislative Framework that seeks to harmonize product safety requirements, it is unsurprising that the definition resonates with a product safety conception of risk that can be measured quantitatively[42]. However, the AIA is not just about curtailing risks to individuals' health and safety but also risks posed to their fundamental rights, which are less apt to be quantified in this way. Underpinning many fundamental rights frameworks is the right to human dignity, including the EU's own EU Charter of Fundamental Rights that states in Article 1 that "human dignity is inviolable. It must be respected and protected"[43]. Questions have been raised as to how far dignity can be ensured where decisions are

40 Ministry of Foreign Affairs of Japan, *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems* (October 30, 2023).

41 Organisation for Economic Co-operation and Development, Organisation for Economic Co-operation and Development *Framework for the Classification of AI Systems*, Paris: OECD Publishing (2022).

42 Marco Almada and Nicolas Petit, *The EU AIA: A Medley of Product Safety and Fundamental Rights?*, Robert Schuman Centre for Advanced Studies Research Paper No. 2023/59 (October 18, 2023).

43 European Union, *Charter of Fundamental Rights of the European Union*, Official Journal of the European Communities, C 364/1 (December 18, 2000)

made based on quantitative data that may not consider an individual but rather derive insights from groups with characteristics that may resemble that individual[44]. Does a focus on the individual's fundamental rights neglect the risks posed to collective interests that fall outside the scope of fundamental rights yet are nevertheless vulnerable to the operation of AI systems and can entail equally deleterious consequences for individuals?[45] How can one operationalize human dignity within risk management systems and technical specifications? AI systems do not operate in a vacuum but have further ramifications and consequences, such as potentially altering our understanding of fundamental societal relationships. How can one calculate the risk posed to a general sense of social trust that underpins many social interactions considering the increased use of deepfakes that leads many to question if AI generates everything?[46]

Others have criticized the AIA's concept of risk for different reasons. Whilst the European Commission can revisit the list of high-risk systems in Annex III, the criteria listed in Article 7(2) for allowing these changes were not justified by reference to statistical data on the potential for other AI systems to cause harms captured by the AIA[47]. Furthermore, the AIA's approach to risk has fostered a very one-sided understanding of AI systems as there is an exclusive focus on the negatives associated with AI systems without a corresponding framework to articulate the benefits[48].

As the first legislative instrument governing AI, some have lamented the AIA as a missed opportunity, particularly given the possibility of it being used as a model in other jurisdictions or to at least set the tone for future global discussions on the subject. The AIA does not necessary preclude all instances of AI systems used for predictive policing where there is human input, real-time

44 Carsten Orwat *et al., Normative Challenges of Risk Regulation of Artificial Intelligence*, NanoEthics 18, no. 2 (2024): 11, https://doi.org/10.1007/s11569-024-00454-9.

45 Sue Anne Teo, *How Artificial Intelligence Systems Challenge the Conceptual Foundations of the Human Rights Legal Framework*, Nordic Journal of Human Rights 40, no. 1 (2022): 216-34, https://doi.org/10.1080/18918131.2022.2073078.

46 Atoosa Kasirzadeh, *Two Types of AI Existential Risk: Decisive and Accumulative*. arXiv preprint arXiv:2401.07836 (2024).

47 Ljupcho Grozdanovski and Jérôme De Cooman, *On the Obsolescence of Empirical Knowledge in the Age of AI*, Rutgers Computer & Technology Law Journal 49, no. 2 (2023): 123-145

48 Martin Ebers, *Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU's AIA*, European Union Law Working Papers, (October 21, 2024), Stanford-Vienna Transatlantic Technology Law Forum.

and ex-post remote biometric identification in public spaces, and biometric categorization systems[49]. The concern is that such systems have such poor accuracy rates, known issues with generating biased outputs, and obvious implications for several fundamental rights that even the obligations placed on high-risk systems are insufficient.

I.III RELEVANT ACTORS

The AIA establishes comprehensive obligations for multiple entities across the AI value chain, collectively termed 'operators' and defined in Article 3(8) as "a provider, product manufacturer, deployer, authorized representative, importer or distributor".

The Act imposes its most stringent requirements on 'providers'. According to Article 3(3), a provider is "a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge".

Deployers, defined in Article 3(4) as "a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity", also have obligations that vary by system type. Under Article 26(1), deployers must take "appropriate technical and organizational measures" to ensure they use systems in accordance with provided instructions. For emotion recognition or biometric categorization systems, Article 50(3) requires deployers to inform exposed individuals of the system's operation.

Importers, defined in Article 3(6) as "a natural or legal person located or established in the Union that places on the market an AI system that bears the name or trademark of a natural or legal person established in a third country", must verify fundamental compliance elements under Article 23(1). This includes confirming appropriate conformity procedures, technical documentation, CE marking, and authorized representative appointment. For high-risk systems, Article 23 prohibits market placement of non-compliant systems, requires proper storage conditions, mandates ten-year documentation retention, and establishes active cooperation with authorities on compliance measures.

49 *Supra* note 38.

Distributors, defined in Article 3(7) as "a natural or legal person in the supply chain, other than the provider or the importer, which makes an AI system available on the Union market", have specific obligations for high-risk AI systems under Article 24. Before making such systems available, they must verify CE marking, EU declaration of conformity, instructions for use, and compliance of providers and importers with their labeling obligations. Article 24 further requires distributors to ensure proper storage, take corrective actions for non-compliant systems, and immediately notify authorities of risks that fall within the scope of Article 79(1).

Article 25 establishes important cross-cutting provisions affecting all actors. Any operator making substantial modifications to, or changing a high-risk AI system's intended purpose, or putting their own name or trademark on an already marketed high-risk AI system, assumes full provider obligations for that system. Where this happens, the entity previously holding provider obligations ceases to do so, except where the previous provider has clearly specified that their AI system is not to be changed into a high-risk AI system.

I.IV THE AIA's TWIN AIMS AND THE NEED FOR TRADE-OFFS

The EU's legislative authority is constrained by certain principles enshrined in its founding treaties, including the need for a specific legal basis within these treaties. These legal bases provide an ongoing "steer" during the legislative process and, ultimately, influence the legislative instrument itself. The AIA rests on two such bases: Article 114 TFEU and Article 16 TFEU[50].

Article 114 TFEU enables the EU to adopt harmonizing measures across its Member States to establish and maintain the functioning of the internal market[51]. Harmonization operates to prevent fragmentation across the single market as Member States develop their own, potentially diverging, laws. Here, the consistency created by harmonization also operates to ensure legal certainty for those who produce, deploy, and use AI systems across the EU. Article 16 TFEU justifies the EU's legislative competence in the areas of personal data protection and the free movement of data. This legal basis aligns with the EU

50 European Commission. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.* COM(2021) 206 final (April 21, 2021).

51 See *TFEU*, Article 114(1).

Charter of Fundamental Rights that upholds both a right to privacy and a separate right to the protection of personal data and highlights its resonance with the GDPR.

These twin aims are reflected in Article 1(1) of the AIA, which defines the AIA's scope as the improvement of the functioning of the internal market whilst ensuring the protection of fundamental rights through the uptake of human-centric and trustworthy AI. A focus on the internal market explains the AIA's inapplicability to AI systems used exclusively for military, national security, or scientific research and development. (Articles 2(3) and 2(6)). The AIA harmonizes the rules relating to CE marking and compliance standards to integrate AI systems into the broader product safety legislative framework to provide legal certainty for producers and providers of AI systems and demonstrate AI systems' safety and trustworthiness to AI users.

These conditions are vital to the EU's efforts to promote innovation and growth within the internal market. In keeping with its fundamental rights underpinning, the AIA expands the concept of relevant risks within the existing product safety framework to incorporate risks into fundamental rights (Article 65(1)). This means developers of high-risk AI systems must conduct a pre-market assessment that incorporates potential impacts on fundamental rights and mitigation strategies (Article 6(2)). Furthermore, it requires deployers of high-risk AI systems to undertake fundamental rights impact assessments if they are governed by public law or are private entities providing public services or acting on behalf of public bodies (Article 27) and to consider the impact on fundamental rights within their risk management systems (Article 9) and data governance systems (Article 10).

Recognizing the role and relevance of fundamental rights to the legislation of AI systems is a positive development because it expands the understanding of relevant risks to better ensure AI systems can only be used for beneficial purposes. The weaving of the economic imperative of the single market and the ethics of AI shows how the development of legislation in this space is concerned with the integration of these aims, reflecting the EU's commitment to both these goals. This linkage is acknowledged in the wider corpus of EU materials concerning the governance of AI where respect for fundamental rights is not just seen as essential in itself, but also as a means of facilitating public acceptance and the competitive advantage of EU-based companies[52]. Within

52 European Commission, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the*

the AIA itself, the integration of the two aims is demonstrated in the drive for standards, a key technical means for demonstrating compliance with the AIA, to reflect fundamental rights considerations, and the introduction of a specific fundamental rights impact assessment.

In their work on the EU's Digital Single Market, Neergaard and De Vries argue that laws pertaining to the single market could act as a "normative corridor" to accommodate other objectives, including fundamental rights protections[53]. This metaphor implies a shared collective understanding of the internal market with common norms and standards that also shape the approach to fundamental rights, whilst still leaving some room for differences[54]. Within the AIA, the concept of human-centric and trustworthy AI may operate to provide a foundation for common norms that apply to the EU's economic imperative and fundamental rights protections. This would support a harmonious interpretation of the AIA that supports its twin aims.

However, the view that the AIA's twin aims may operate harmoniously in tandem is not universally held. First, some have questioned the appropriateness of a product safety framework for the protection of fundamental rights. Whilst the AIA expands the concept of risk to encompass risks to fundamental rights, the definition of risk within the AIA may fail to capture the types of risks AI systems pose to fundamental rights. Understandings of risk through a product safety lens may neglect sociopolitical risks and risks arising from the broader context of AI use, rather than just the AI system[55].

Second, there are also potential gaps and grey areas where fundamental rights protections may fall away. The impact of AI has been particularly felt within the fields of medicine and healthcare[56]. Whilst the risks to patients are numerous and potentially severe[57], in contrast to the internal market, the EU has a much more limited competence to legislate over health-related matters

Regions: Building Trust in Human-Centric Artificial Intelligence, COM(2019) 168 final (April 8, 2019).

53 Ulla Neergaard and Sybe de Vries, *The Interaction between Free Movement Law and Fundamental Rights in the (Digital) Internal Market*, Renforce Working Paper 1-2023 (2023): 1-30.

54 Cathleen Kantner, *What Is a European Identity? The Emergence of a Shared Ethical Self-Understanding in the European Union*, EUI RSCAS Working Papers 2006/28 (2006).

55 *Supra* note 42.

56 *Supra* note 9.

57 W. Nicholson Price II and I. Glenn Cohen, *Locating Liability for Medical AI*, DePaul Law Review 73, no. 2 (2024): 339-368.

to the potential detriment of patients. Furthermore, while the AIA references fundamental rights, the legal basis concerns the protection of personal data. This basis is unlikely to cover all the risks posed to patients in a healthcare context, such as infringements on their rights to access healthcare, non-discrimination, and freedom from inhumane treatment. This suggests the AIA may promote the wider use of medical AI within the EU that fully promotes the EU's economic aims but does not offer the same level of fundamental rights protections[58]. The requirement to conduct a fundamental rights impact assessment only applies to certain deployers that are either public bodies or private entities providing a public service or acting on behalf of a public body. This potentially leaves a range of AI systems not subject to this requirement, including those relating to employment and certain aspects of the financial sector[59].

Finally, the incorporation of fundamental rights into a product safety framework may lead regulators to underestimate the need to develop new capabilities. Under Article 40, fundamental rights may fall within scope of the harmonized standards that are developed by European standardization organizations (ESOs). These are private, albeit non-profit, entities that now play a critical role in translating fundamental rights into technical standards. This plays a vital role in operationalizing the AIA, despite their lack of democratic legitimacy and the fact that most stakeholders within these groups are from industry[60]. This development in their role raises questions over whether they have sufficient expertise and appetite to appropriately respond to the challenges that AI may pose to fundamental rights. The role of ESOs in operationalizing fundamental rights protection under the AIA has also been criticized for lacking the transparency and openness of the EU legislative process. Despite the essential role these standards will play in ensuring AI systems do not impact fundamental rights to the detriment of individuals, there is no guarantee of free and easy access to the standards once they are agreed[61]. This is not the only source of opacity that runs counter to the effective operation of fundamental rights.

58 Hannah van Kolfschooten, *EU Regulation of Artificial Intelligence: Challenges for Patients' Rights*, Common Market Law Review 59, no. 1 (2022): 81-112.

59 *Supra* note 38.

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

61 Marion Ho-Dac, *Considering Fundamental Rights in the European Standardisation of Artificial Intelligence: Nonsense or Strategic Alliance?*, arXiv (January 23, 2024).

Such opacity triggers uncertainty that undermines the efforts to create human-centric and trustworthy AI that promotes the economic aims of the EU and fundamental rights.

I.V SUPPORT FOR INNOVATION – THE NEED TO PROTECT STARTUPS AND SMEs

The relationship between regulation and innovation is complex[62]. While regulation can act as a powerful stimulus for innovation by creating market certainty and driving improvements in products and processes, it can simultaneously impose compliance costs and restrictions that may discourage innovative activities. The potential impacts of regulation on innovation are particularly salient for emerging technologies, like AI, as well as for small and medium-sized enterprises (SMEs), who are less able to bear the costs of regulatory compliance. Experience with previous regulatory interventions suggests that the impact of regulation on innovation is highly context-dependent and varies significantly based on factors including timing, stringency, flexibility, and certainty[63].

Several provisions of the AIA could potentially constrain innovation through both direct and indirect mechanisms. First, the Act's stringent requirements for high-risk AI systems impose substantial financial compliance costs. The European Commission estimates compliance costs for providers at approximately €6,000-€7,000 per AI system, with deployers facing additional human oversight costs of €5,000-€8,000 per year[64]. For systems requiring third-party conformity assessment, providers face further verification costs of €3,000-€7,500, assuming they already have a Quality Management System (QMS) in place. These costs could be significantly higher –with estimates of ca. €193,000-€330,050 for larger enterprises and €144,350-€247,150 for SMEs– if a new QMS needs to be established[65]. Second, the extensive testing and validation requirements for high-risk AI systems under Articles 9 and 10 may significantly extend development cycles and increase costs. While the European Commission estimates that some 15%

62 Jacques Pelkmans and Andrea Renda, *Does EU Regulation Hinder or Stimulate Innovation?*, CEPS Special Report No. 96 (November 19, 2014), Brussels: Centre for European Policy Studies.

63 *Ibid.*

64 Meeri Haataja and Joanna J. Bryson, *What Costs Should We Expect from the EU's AIA?*, SocArXiv (2021).

65 *Ibid.*

of AI systems would fall in this category, a survey conducted by the appliedAI initiative found that at least a third of AI startups would classify their products as high-risk AI systems[66].

Beyond these direct financial impacts, the Act may constrain innovation through several indirect mechanisms. The transparency and interpretability requirements in Article 13 present technical challenges, since explainability methods have sometimes been found to be imprecise, inconsistent, and vulnerable to attacks[67]. Meeting these requirements may therefore force providers to favor simpler, more interpretable models over potentially more effective but less transparent approaches. The lack of clear definitions for key terms like 'sufficient transparency' and 'interpretability' creates legal uncertainty, and the extensive documentation requirements may divert resources away from R&D activities, especially affecting SMEs and startups with limited staff and budgets.

While there are thus substantial reasons to worry about the effect of the AIA on innovation, the Act also incorporates several measures specifically designed to support innovation. Most notably, Articles 57-63 establish a comprehensive framework centered around regulatory sandboxes and other support mechanisms. Regulatory sandboxes are controlled environments where companies can develop and test innovative AI systems under regulatory supervision[68]. According to Article 57, Member States must ensure that at least one AI regulatory sandbox is established at national level by August 2026. These sandboxes serve multiple objectives laid out in Article 57(9): improving legal certainty for AI development, supporting best practice sharing through cooperation between authorities, fostering innovation and competitiveness, facilitating market access for AI systems (particularly from SMEs and startups), and contributing to evidence-based regulatory learning. The sandbox framework includes several specific provisions to ensure accessibility and effectiveness, with Article 58(2) mandating broad and equal access through transparent selection criteria and requiring participation to be free of charge for SMEs and startups.

66 appliedAI Initiative, *AIA Impact Survey: Pan-European Survey of AI Startups and VCs about the Impact of the AIA*, AppliedAI.De, (December 12, 2022). https://www.appliedai.de/assets/files/AI-Act-Impact-Survey_Slides_Dec12.2022.pdf.

67 Luca Holst *et al.*, *The Impact of the EU AIA's Transparency Requirements on AI Innovation*, in Proceedings of the 19th International Conference on Wirtschaftsinformatik, Würzburg, Germany (2024).

68 Thomas Buocz, Sebastian Pfotenhauer, and Iris Eisenberger, *Regulatory Sandboxes in the AIA: Reconciling Innovation and Safety?*, Law, Innovation and Technology 15, no. 2 (2023): 357-389, https://doi.org/10.1080/17579961.2023.2245678.

While the AIA itself does not define SMEs, startups, and microenterprises, it builds upon established EU definitions, particularly Commission Recommendation 2003/361/EC[69]. The AIA's provisions for these entities thus likely follow these established thresholds, though notably the term 'startup' remains undefined in both instruments.

Article 59 lays out provisions for further processing of personal data within regulatory sandboxes for specific public interest purposes, including public health, environmental protection, and energy sustainability. These provisions contain multiple safeguards intended to ensure compliance with the GDPR, further supporting innovation by providing a clear path for complying with two distinct but relevant regulatory instruments. This is complemented by Article 60's framework for real-world testing outside sandboxes, subject to strict safeguards including market surveillance authority approval and informed consent requirements detailed in Article 61.

Article 62 mandates that Member States provide SMEs and startups with priority access to regulatory sandboxes and establish specific support measures including awareness-raising, training activities, and dedicated communication channels. The article also stipulates that the needs and interests of SMEs shall be taken into account when the fees for conformity assessment under Article 43 are set, "reducing those fees proportionately to their size, market size and other relevant indicators". Finally, Article 62 requires the AI Office to develop standardized templates and maintain a single information platform to ease compliance burdens. Article 63 further allows microenterprises to comply with quality management system requirements in a simplified manner, though without compromising on core obligations regarding high-risk AI systems.

Article 2 of the AIA includes several important exemptions related to research and development. Most notably, the Act "does not apply to AI systems or AI models, including their output, specifically developed and put into service for the sole purpose of scientific research and development". Additionally, any

69 European Commission, *Commission Recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises* (Text with EEA relevance) (notified under document number C(2003) 1422) OJ L 124, (20 May 2003):36-41. This Recommendation defines: -Microenterprises as having fewer than 10 employees and annual turnover/balance sheet total not exceeding €2 million, - Small enterprises as having fewer than 50 employees and annual turnover/balance sheet total not exceeding €10 million, - Medium-sized enterprises (completing the SME category) as having fewer than 250 employees and annual turnover not exceeding €50 million and/or balance sheet total not exceeding €43 million.

"research, testing or development activity regarding AI systems or AI models prior to their being placed on the market or put into service" is exempted. What this will mean in practice is, however, not easy to predict[70]. In particular, it is unclear what will count as development for the 'sole purpose' of research and development.

Much of contemporary research and development, in AI as well as more broadly, is a hybrid endeavor mixing private and public actors and scientific and commercial motivations[71]. Similarly, the Act's explicit exclusion of real-world testing from this exemption introduces questions related to the ability to neatly separate pure research from ongoing development efforts[72]. Finally, scholars have identified potential mechanisms for 'regulatory arbitrage' whereby AI system providers might structure their activities to fall within the research exemption while actually pursuing commercial ends[73].

While there are thus clear risks that the AIA could constrain innovation through both direct compliance costs and indirect effects, regulation can also facilitate innovation by creating clear market signals, fostering predictable investment environments, and driving improvements in products and processes to meet new standards[74]. While compliance costs and technical constraints may create headwinds for certain types of AI development, particularly among resource-constrained SMEs, the Act's establishment of clear standards and support mechanisms could help create the predictability and trust necessary for AI innovation.

70 Sebastian Porsdam Mann, I. Glenn Cohen, and Timo Minssen, *The EU AIA: Implications for U.S. Health Care*, NEJM AI 1, no. 11 (October 2024). https://ai.nejm.org/doi/pdf/10.1056/AIp2400449.

71 Liane Colonna, *The AIA's Research Exemption: A Mechanism for Regulatory Arbitrage?* in YSEC Yearbook of Socio-Economic Constitutions 2023: Law and the Governance of Artificial Intelligence, ed. Andreas Moberg and Eduardo Gill-Pedro (Cham: Springer Nature Switzerland, 2024)

72 *Ibid.*

73 *Ibid.*

74 Shuai Shao *et al.*, *Environmental Regulation and Enterprise Innovation: A Review*, Business Strategy and the Environment 29, no. 3 (2020): 1465-1478, https://doi.org/10.1002/bse.2446.

II. DISCUSSION

Our above analysis revealed that the AIA is a very sophisticated and ambitious legislation. But what promises, perils and trade-offs is the AIA now facing within a complex regulatory ecosystem and in an increasingly fierce international competition? Will the AIA really serve as a regulatory north star both within the EU and beyond, or does its intended operation risk being undermined by implementation challenges and concerns? These are complex questions which need to be discussed from a variety of angles.

II.1 POTENTIAL FOR GLOBAL IMPACT

Like other major EU Regulations, such as the GDPR, Digital Services Act, and Digital Markets Act, the AIA forms part of the growing EU regulatory infrastructure aimed at protecting fundamental rights while promoting responsible innovation.

While the Act will therefore have a direct impact on the legislation of its Member States, its effect will also be felt beyond EU borders. This is due, in part, to its scope, as defined in Article 2, which stipulates that the Act applies to any provider placing AI systems on the EU market or whose systems' outputs are used in the EU, regardless of their location. The AIA can therefore create obligations and responsibilities for deployers, importers, and distributors of AI systems, as well as manufacturers incorporating AI into their products irrespective of their location. This extraterritorial reach means that any organization whose AI systems touch the EU market must comply with the Act's requirements. Notably, the Act applies even when only the output of an AI system is used in the Union, a particularly far-reaching provision that could capture many systems operated entirely outside EU territory.

This broad territorial scope reflects a well-documented phenomenon in EU regulation known as the "Brussels Effect", whereby EU regulations can become de facto global standards through market mechanisms[75]. This is, in part, because global companies must meet EU standards to launch their products in this substantial market and, rather than adhere to multiple compliance regimes in different territories, companies may prefer to apply the EU rules to all their operations. Furthermore, the adoption of a unified approach may be commercially attractive where this adherence to stringent regulations can signal quality and

75 Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (New York: Oxford University Press, 2020), https://doi.org/10.1093/oso/9780190088583.001.0001.

trustworthiness, and because it may be easier to design products for compliance with the strictest standards rather than maintain separate products for different markets. The GDPR illustrates this effect well –while technically only governing data protection in the EU, its requirements have shaped global privacy practices as companies found it more efficient to apply GDPR standards universally rather than maintain different standards for different markets. We may also see a *de jure* Brussels Effect where countries outside of the EU may adopt regulations very similar to the AIA, either due to commercial lobbying from businesses who would prefer a unified regulatory regime to reduce compliance costs or through the EU's own political influence[76].

Whether the AIA will lead to a Brussels Effect is not a straightforward question. Neither is the distinct, albeit related, question of whether such an influence is desirable. On the one hand, some have pointed out that the Act's complexity –compared to the GDPR's relatively more straightforward principles and obligations– makes wholesale adoption by other jurisdictions unlikely[77]. However, others have noted that the Act seems to broadly satisfy the five criteria originally laid out by Bradford as necessary for a Brussels Effect[78]: significant market power (450 million EU consumers), regulatory capacity (as demonstrated by past digital regulations), stringent standards (particularly for high-risk and general-purpose AI), inelastic targets (companies cannot easily avoid compliance), and the regulation targeting non-divisible conduct (as creating separate AI systems for different markets is likely to be impractical or costly).

Yet even if the Act does produce a Brussels Effect, Almada and Radu warn that this could be accompanied by what they term a "Brussels Side Effect" –a diminution in fundamental rights protection globally[79]. While the EU's competence to regulate fundamental rights –founded upon the Charter of Fundamental Rights and Article 16 TFEU specifically enabling data protection regulation –Almada and Radu argue that the Act's product safety framework may inadequately capture certain fundamental rights dimensions.

76 *Ibid.*

77 Christopher T. Marsden and George Christou, *Artificially Intelligent Regulation: Global Norms, International Political Economy and the Brussels Effect*, in Proceedings of the 2024 International Conference on Artificial Intelligence and Regulation (2024)

78 Charlotte Siegmann & Markus Anderljung, *The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market*, Centre for the Governance of AI (Aug. 2022). https://cdn.governance.ai/Brussels_Effect_GovAI.pdf

79 Marco Almada and Anca Radu, *The Brussels Side-Effect: How the AIA Can Reduce the Global Reach of EU Policy*, German Law Journal (2024).

This limitation could lead to the global adoption of the Act's mechanisms actually reducing overall fundamental rights protection by promoting an incomplete framework for protecting rights. Thus, on their analysis, a substantial Brussels Effect is not necessarily to be desired –due to the potential for this to undermine, rather than advance, the EU's regulatory policy goals.

The Act's global impact will also be shaped by the alternative regulatory approaches that are emerging worldwide. The United States does not, so far, have a comparatively systematic regulatory framework for AI. President Biden's Executive Order on AI[80], which was binding only on executive branch staff and agencies, emphasized safety testing for advanced AI systems while preserving innovation. However, this Order has now been revoked and replaced by Trump's executive order[81,82], which may undermine arguments of the existence of a Brussels Effect. Trump's Executive Order emphasizes the US's desire to "act decisively to retain global leadership in artificial intelligence" and removes all references to guardrails rooted in ethical principles that resonate with the EU's governance approach. China has taken a more comprehensive approach through various regulations including the Algorithm Recommendation Management Provisions, though these reflect different priorities around social control and national security. At the international level, both the OECD AI Principles and UNESCO's Recommendation on the Ethics of AI have established influential soft law frameworks emphasizing human-centric AI development, though these do not have directly binding legal force. Most notably, the Council of Europe is developing its own AI Convention, which takes a fundamentally different approach centered on human rights, democracy, and the rule of law rather than product safety[83].

---

80 The White House. *Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.* Statements and Releases (October 30, 2023). https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

81 Joseph R. Biden Jr., *Executive Order 14141: Advancing United States Leadership in Artificial Intelligence Infrastructure*, Federal Register 90, no. 5469 (January 17, 2025): 5469-5489.

82 Donald J. Trump, *Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence*, Federal Register 90, no. 5469 (January 23, 2025): 5469-5471. https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence.

83 Hannah van Kolfschooten and Carmel Shachar, *The Council of Europe's AI Convention (2023-2024): Promises and Pitfalls for Health Protection*, Health Policy 138 (December 2023): 104935, https://doi.org/10.1016/j.healthpol.2023.104935.

Given these dynamics, the AIA's global impact is likely to be significant but more nuanced than previous instances of the Brussels Effect. Much depends on the extent to which its technical requirements will diffuse through market mechanisms, as well as on how key provisions and exemptions are interpreted. Ultimately, the extent of the AIA's global influence will depend largely on how effectively the Act balances competing demands, such as protecting fundamental rights while fostering innovation, and providing clear rules while maintaining flexibility. Given the rapid pace of both the underlying technology as well as global regulatory responses, the ultimate impact of the AIA is not predictable at present.

II.II POTENTIAL TO RESPOND TO FUTURE TECHNOLOGICAL ADVANCES

There are several aspects that can be considered to evaluate the extent to which the AIA may be able to respond to future technological advances. The *first* is to consider the ability of the provisions to respond to novel and potentially unexpected technological developments. Will the risk-based system be able to accommodate technological advancements? Do the provisions allow the relevant actors to gather high-quality information to determine how the AIA operates in relation to technological advances? The *second* is related to the actors involved in the implementation, use, and enforcement of the provisions – how far are they able and willing to respond to future technological developments within the AIA? *Finally*, the AIA does not operate in a vacuum so the potential to respond to future technological developments will also depend on its ability to form a part of a wider coherent regulatory regime.

With regard to the *first* aspect, the provisions show some potential to adapt to future technological advances. The AIA uses a broad, internationally recognized definition of AI that should help ensure that it captures future AI approaches and systems. Furthermore, the AIA's focus is often on broad use cases and applications that further support its ability to capture future technological advances where they are deployed in areas covered by the AIA. In addition, Article 112 outlines the Commission's ability to assess the need for it to make amendments to Annex III, which provides one source of high-risk AI systems, and Article 5, which sets out prohibited practices. The Commission will undertake the assessment annually and report its findings to the European Parliament and Council. Indeed, Article 112 sets out detailed procedures for the review and evaluation of the AIA on several points relevant to the AIA's ability to respond to future technological developments, including amendments to provisions concerning the definition

of high-risk AI systems, AI systems requiring additional transparency measures, and the effectiveness of the supervision and governance system (Article 112(2)). Nevertheless, many of the mechanisms for evaluation and review listed in Article 112 are only required to take place every four years, which may lead to these evaluations lagging behind technological developments and responding too late. Moreover, some provisions have already been criticized for being out-of-date and not responding to current technological trends, let alone future ones. One of the means of defining general-purpose AI is threshold of 10^{25} FLOPS (floating point operations). The AIA assumes that AI models that use more than this amount of computation are a general-purpose AI model, unless a developer can demonstrate otherwise (Article 52(2)). The assumption that computing power is a defining feature of when a general-purpose AI system will lead to systemic risks has been criticized as "unconvincing"[84] and efforts to optimize AI models further undermine the relevance of the 10^{25} FLOPS threshold. This prompts concern that AI developers will focus on making AI models that fall below this threshold, rather than developing ones that are less risky. The recent release of models well under this threshold which nevertheless demonstrate state-of-the-art performance, such as DeepSeek-R1 and GPT-o3-mini, further challenges the relevance of this threshold as a proxy for systemic risk.

Now we turn to the *second* aspect, i.e., the ability of the actors involved in the application of the AIA to respond appropriately to technological developments relevant to the AIA's operation. These relevant actors include AI providers and deployers that implement the measures required to reduce the risk posed by AI systems, ESOs that develop appropriate standards to guide AI development in a way that fosters the competitiveness of the single market whilst respecting fundamental rights, and public institutions tasked with the monitoring of these actors and any necessary enforcement action. The more one focuses on technical standards to ensure the AIA responds to technological advances, the more regulatory choices are depoliticized and discussions as to values and norms, led by public institutions, are replaced with statistics and technical facts by private institutions. This leads to public institutions being largely in the dark about the day-to-day operations of the AIA, a state of affairs that becomes very problematic if technological advances then require them to intervene. The AIA contains provisions that enable public authorities to receive information on conformity assessments for high-risk AI systems (Article 21), on the operation of general-purpose

84 *Supra* note 38.

AI models (Articles 52-55), information on how far high-risk AI systems impact fundamental rights (Article 77), and to supervise, investigate, enforce, and monitor general-purpose AI risks (Section 5, Chapter IX). However, the ability to use that information to respond to potential technological developments is a distinct matter with many questioning the capability of public authorities involved in the monitoring and enforcement of the provisions[85], both of which are a valuable source of information on how to respond to novel AI technologies.

The *final* aspect concerns how well the AIA will interact with the wider regulatory landscape to ensure it responds to all relevant risks without the duplication of efforts or resources across different, albeit related, regulatory schemes. The AIA forms part of the New Legislative Framework and interacts with other EU legislation that governs digital technologies including the General Data Protection Regulation, the Digital Service Act, the Digital Markets Act, the European Health Data Space, and the Medical Devices Regulation. Questions remain as to how far these instruments will co-exist on the broader regulatory landscape concerning data, technology, and data-driven technology companies. Questions as to how far a GDPR-compliant data protection impact assessment will align with an AIA-required fundamental risks impact assessment already demonstrates the scope for duplication or, worse, allowing risks to fall between the cracks of different regulations.

II.III IMPACT ON THE COMPETITIVENESS OF EUROPE

The intentions of the AIA, in seeking to minimize the risks of AI and maximize the potential benefits in a way that respects fundamental values and broader societal goals, are commendable.

However, as stressed in the so-called Draghi Report[86], overly excessive and complex regulations also pose significant implications and risks for the technological competitiveness of the EU in the rapidly evolving global AI landscape.

85 *Supra* note 38.

86 Mario Draghi, The Future of European Competitiveness (September 2024), available at: https://commission.europa.eu/topics/eu-competitiveness/draghi-report_en ; see also European Commission, A Competitiveness Compass for the EU, Brussels, 29.1.2025 COM(2025) 30 final, available at: https://commission.europa.eu/document/download/10017eb1-4722-4333-add2-e0ed18105a34_en .

Recent examples, such as the call by the EU Parliament to reform the MDR[87] the recent withdrawal of forthcoming EU legislation on AI liability[88], plans to simplify and streamline the GDPR[89] and the AIA[90], and the so-called OMNIBUS packages[91] have demonstrated that it is a very delicate task to simultaneously balance the risks of over- and under-regulation, protect citizens' rights and other EU values, and enhance the competitiveness of the EU[92]. In other words, it is not easy to find a "regulatory sweet spot" that promotes the beneficial application of AI innovation and reduces the potentially harmful risks that this technology entails.

One of the primary concerns regarding the AIA is the potential for increased regulatory burdens on businesses, particularly startups and SMEs. As previously discussed, the classification of AI systems into risk categories –ranging from minimal to unacceptable risk– imposes varying levels of compliance requirements. For instance, high-risk AI applications will face stringent obligations, including robust documentation, risk assessments, and ongoing monitoring. This could create a challenging environment for smaller entities that may lack the resources to navigate complex regulatory requirements. Consequently, as compliance costs rise, there is a risk that innovation may be stifled and the development of groundbreaking AI technologies within the EU could grind to a halt[93].

87 European Parliament resolution on the urgent need to revise the Medical Devices Regulation (2024/2849(RSP)), available at: https://www.europarl.europa.eu/doceo/document/RC-10-2024-0123_EN.html .

88 *Supra* note 22.

89 Ellen O'Regan, Europe's GDPR privacy law is headed for red tape bonfire within 'weeks', Politico (3 April 2025), available at: https://www.politico.eu/article/eu-gdpr-privacy-law-europe-president-ursula-von-der-leyen/.

90 Pieter Haeck, EU opens door to reworking AI rulebook, Politico (9 April 2025), https://www.politico.eu/article/how-eu-did-full-180-artificial-intelligence-rules/; see also European Commission, Commission sets course for Europe's AI leadership with an ambitious AI Continent Action Plan, press release (9 April 2025), available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1013

91 European Commission, Directorate General for Communication, Commission proposes to cut red tape and simplify business environment (26 February 2025), available at: https://commission.europa.eu/news/commission-proposes-cut-red-tape-and-simplify-business-environment-2025-02-26_en .

92 See e.g. Jonas Tallberg, Magnus Lundgren, and Johannes Geith, *AI Regulation in the European Union: Examining Non-State Actor Preferences*, Business and Politics 26, no. 2 (2024): 241-267, https://doi.org/10.1017/bap.2024.10 .

93 See Aboy, Vayena, Minssen, *supra* note 8.

Moreover, the potential for regulatory divergence between the EU and other regions –such as the United States or China– could further undermine the EU's competitive position. As AI technologies continue to advance at an unprecedented pace, the more flexible regulatory environments in other jurisdictions may attract investments and talent away from Europe. Companies may choose to relocate their AI research and development activities to regions with fewer restrictions, undermining the EU's position as a global leader in AI innovation. This is particularly concerning given that the AI sector is characterized by rapid iterations and the need for agile responses to market demands.

On the other hand, the AIA also has the potential to enhance the EU's competitive edge in certain areas. By establishing clear ethical guidelines and standards, the Act could foster trust among consumers and businesses in AI technologies. In a market increasingly concerned about privacy, security, and the ethical implications of emerging technologies, the EU's proactive stance on regulating AI may serve as a differentiating factor. European companies that comply with these regulations can position themselves as leaders in responsible AI, appealing to consumers who prioritize ethical considerations. This trust can translate into market advantages, particularly in sectors like healthcare, finance, and autonomous systems, where ethical concerns are paramount.

Furthermore, the AIA emphasizes the importance of human oversight and accountability in AI systems. This focus on responsible AI practices could drive innovation in developing transparent and explainable AI models. Such advancements may not only benefit European companies but could also set a global standard for AI practices. If the EU can successfully cultivate a reputation for high-quality, ethical AI, it may attract international collaborations and partnerships, reinforcing its position in the global AI ecosystem.

In conclusion, the AIA presents a double-edged sword for the technological competitiveness of the EU. While the potential for regulatory burdens may hinder innovation and drive talent elsewhere, the emphasis on ethical standards and consumer trust could establish Europe as a leader in responsible AI. Striking the right balance between regulation and innovation will be crucial in determining whether the EU can maintain its technological competitiveness in an increasingly AI-driven world. As global dynamics shift, ongoing dialogue among policymakers, industry stakeholders, and the public will be essential to ensure that the AIA serves as a catalyst for growth rather than a barrier to progress.

II.IV LESSONS FOR OTHER COUNTRIES

Countries around the world can learn several important lessons from the EU's AIA as they navigate the complexities of AI regulation and development. One of the primary lessons is the necessity of finding a balance between regulatory oversight and fostering innovation. While it is critical to establish safety and ethical standards for AI technologies, overly burdensome regulations can stifle creativity and deter investment. Countries should aim to create flexible regulatory frameworks that can adapt to the fast-paced nature of AI development, allowing for innovation while ensuring public safety and upholding ethical principles.

The EU's emphasis on ethical AI and transparency highlights the importance of building public trust in AI systems. Countries should prioritize developing guidelines that promote responsible AI use, which includes ensuring that AI applications are explainable and accountable. Establishing clear ethical standards can not only enhance consumer confidence but can also position nations as leaders in responsible AI, attracting businesses and talent seeking to operate in a trustworthy environment.

The AIA's development illustrates the value of engaging a broad range of stakeholders, including industry leaders, academics, and civil society. Inclusive dialogue can ensure that regulations are grounded in practical realities and address the concerns of all parties affected. Countries should adopt collaborative approaches to policymaking, ensuring that regulations reflect diverse perspectives and foster a healthy ecosystem for AI innovation.

As AI technologies transcend borders, international cooperation is essential. Countries should seek to harmonize regulations and share best practices to avoid fragmentation. Collaborative efforts can enhance global standards, encourage knowledge sharing, and facilitate cross-border investments in AI research and development.

Finally, investing in regulatory capacities, education, and workforce training is crucial. As AI continues to evolve, countries must equip their authorities and populations with the skills needed to thrive in an AI-driven economy. This proactive approach will ensure that nations remain competitive and can adapt to the changing technological landscape. This also means that sufficient funds will have to be directed to the authorities, named bodies and stakeholders that are supposed to monitor, support, and enforce complex regulatory set-ups. As other countries consider how the EU's approach to AI governance pans out, part of this evaluation must consider if its complexity is nevertheless deftly handled by a set of well-resourced and highly competent actors. If the actors and

mechanisms involved in the monitoring, supporting, and enforcement of the AIA fall short of expectations, other countries might prefer more streamlined and simplified regimes. It will be for other countries to decide whether a mismatch between the regulatory burden and institutional capacity, in the hopes this would be a teething problem during the Regulation's infancy, would be a preferable approach to having fewer rules that are easier to comply with and more likely to be enforced.

By learning from the EU's experience, other countries can strategically navigate the challenges and opportunities presented by AI, fostering a balanced, ethical, and innovative environment with a feasible and enforceable regulatory set-up.

III. CONCLUSIONS

Any number of criticisms should not detract from the fact that the EU's AIA is a groundbreaking legislation that will profoundly shape the future of medical AI, both within and potentially beyond Europe. Its role within the EU's regulatory ecosystem is multifaceted as it interacts with existing legislation, promotes ethical AI development, and encourages international cooperation. Its extraterritorial reach, the size of the EU market, and the benefits of harmonized global best practices, make it difficult for foreign stakeholders to ignore its implications. Whilst the likelihood of it being used as a template for other jurisdictions to copy near verbatim seems to be increasingly small, there are valuable lessons for other jurisdictions concerning the development and implementation of comprehensive AI regulation.

The drafting of the AIA demonstrates that the universally accepted desirability of promoting innovation and protecting individual rights does not create a consensus of how to achieve these aims. The AIA, as a pioneering regulatory framework, offers valuable lessons for other countries navigating the complexities of AI governance. By learning from both the advances and challenges encountered in the EU's approach, jurisdictions around the world can develop more effective, ethical, and adaptable regulatory frameworks that promote innovation while safeguarding public interests.

Being the first to regulate does not necessarily mean the EU created the best AI regulation possible, but how actors cope with these rules will mean the AIA will continue to play a crucial role in shaping the future of AI governance in Europe and beyond. By prioritizing flexibility, support for research, and workforce development, countries can position themselves to harness the transformative

potential of AI while ensuring that its deployment aligns with societal values, legal framework and ethical considerations. Ultimately, the goal should be to foster an environment that encourages innovation while safeguarding public interests, thereby ensuring a competitive edge in the rapidly evolving global AI landscape.

For the time being, foreign stakeholders that want to invest in European markets, including companies from Latin America, should be proactive and stay informed about the AIA's implications for AI systems. To successfully navigate the AIA's impact and to avoid harsh fines, they should carefully assess their obligations, invest in compliance measures, and engage with EU regulators during its implementation phases. Even after the AIA fully enters into force, the European Commission may adopt implementing acts and delegated acts that may impact key provisions[94]. The regime will thus continue to evolve through interpretation, application, guidelines, clarifying exceptions, and any potential amendments. Stakeholders, including Latin American companies, should actively provide input to the development of adaptive governance frameworks, i.e., not only to tackle the challenges and seize the opportunities presented by the AIA while maintaining access to the EU market, but also to contribute to developing their own domestic and regional approaches for the governance of AI.

Last, but not least, there remains an urgent need to establish international treaties and agreements on the most fundamental guardrails that should apply to the governance and deployment of AI[95]. While national approaches might differ with regard to regulatory details that make a globally aligned approach difficult, this does not make the need for agreement on international obligations to protect and promote fundamental values any less important nor any less pressing. The need for such basic "guardrails" is more urgent than ever when considering the enormous advances, opportunities and risks of dual use scenarios with other emerging technologies, such as neurotechnologies, synthetic biotechnologies, quantum technologies, and the combination of these with AI.

94 *Supra* note 70.

95 See also Urs Gasser and Viktor Mayer Schönberger. *Guardrails: Guiding Human Decisions in the Age of AI*. Princeton University Press, 2024. https://doi.org/10.2307/jj.6605394.

Acknowledgements: The authors would like to thank Emma Kondrup for her assistance with this chapter, both editorial and in developing the illustrations for this manuscript. Any use of generative AI in this manuscript adheres to ethical guidelines for use and acknowledgment of generative AI in academic research, as outlined in Porsdam Mann, S., Vazirani, A.A., Aboy, M.; Minssen, T. *et al.* Guidelines for ethical use and acknowledgement of large language models in academic writing. *Nat Mach Intell* 6, 1272-1274 (2024). https://doi.org/10.1038/s42256-024-00922-7. Each author has made a substantial contribution to the work, which has been thoroughly vetted for accuracy, and assumes responsibility for the integrity of their contributions. T.M.; H.L.S. & S.P.M. were supported in part by a Novo Nordisk Foundation grant for a scientifically independent International Collaborative Bioscience Innovation & Law Programme (Inter-CeBIL Programme, grant NNF23SA0087056). T.M. was further funded in part by the European Union Classica project (grant 101057321). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency.

BIBLIOGRAPHY

Aboy, M., Minssen, T. & Vayena, E. *Navigating the EU AIA: implications for regulated digital medical products. npj Digit. Med.* 7, 237 (2024). https://doi.org/10.1038/s41746-024-01232-3.

Almada, Marco and Radu, Anca *The Brussels Side-Effect: How the AIA Can Reduce the Global Reach of EU Policy. German Law Journal* (2024).

Almada, Marco and Petit, Nicolas, *The EU AIA: A Medley of Product Safety and Fundamental Rights? Robert Schuman Centre for Advanced Studies Research Paper* No. 2023/59 (October 18, 2023).

Applied AI Initiative, *AIA Impact Survey: Pan-European Survey of AI Startups and VCs about the Impact of the AIA*, AppliedAI.De, (December 12, 2022). https://www.appliedai.de/assets/files/AI-Act-Impact-Survey_Slides_Dec12.2022.pdf.

Biden Jr., Joseph R., *Executive Order 14141: Advancing United States Leadership in Artificial Intelligence Infrastructure,* Federal Register 90, no. 5469 (January 17, 2025): 5469-5489.

Bradford Anu, *The Brussels Effect: How the European Union Rules the World*, New York: Oxford University Press (2020). https://doi.org/10.1093/oso/9780190088583.001.0001.

Buocz, Thomas, Pfotenhauer, Sebastian, and Eisenberger, Iris, *Regulatory Sandboxes in the AIA: Reconciling Innovation and Safety? Law, Innovation and Technology* 15, no. 2 (2023): 357-389. https://doi.org/10.1080/17579961.2023.2245678.

Cantero Gamito, Marta, and Marsden, Christopher T. *AI Co-Regulation? The Role of Standards in the EU AIA*. International Journal of Law and Information Technology 32 (2024). https://doi.org/10.1093/ijlit/eaae011.

Case C-184/20, OT v. Vyriausioji tarnybinės etikos komisija, ECLI:EU:C:2022:601, August 1, 2022.

Colonna, Liane, *The AIA's Research Exemption: A Mechanism for Regulatory Arbitrage?* in *YSEC Yearbook of Socio-Economic Constitutions 2023: Law and the Governance of AI*, ed. Andreas Moberg and Eduardo Gill-Pedro (Cham: Springer Nature Switzerland, 2024).

Crawford, Gail E. *et al., EU AI Office Seeks Stakeholder Engagement in Drafting AI Act Implementing Rules*, Latham & Watkins LLP (July 31, 2024). https://www.lw.com/en/insights/eu-ai-office-seeks-stakeholder-engagement-in-drafting-ai-act-implementing-rules.

Creative Commons *et al., Supporting Open Source and Open Science in the EU AIA* (July 2023).

Draghi, Mario, The future of European competitiveness, available at: https://commission.europa.eu/topics/eu-competitiveness/draghi-report_en (EU Commission September 2024)

Ebers, Martin, *Truly Risk-Based Regulation of AI: How to Implement the EU's AIA*, *European Union Law Working Papers* (October 21, 2024), Stanford-Vienna Transatlantic Technology Law Forum.

European Commission. *Commission Guidelines on Prohibited AI Practices Established by Regulation (EU) 2024/1689 (AIA).* C(2025) 884 final, Brussels (February 4, 2025), https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act.

European Commission, *Commission Recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises* (Text with EEA relevance) (notified under document number C(2003) 1422) OJ L 124, (20 May 2003):36-41.

European Commission, *Commission Work Programme 2025,* last modified February 12, 2025, https://commission.europa.eu/strategy-and-policy/strategy-documents/commission-work-programme/commission-work-programme-2025_en.

European Commission, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Building Trust in Human-Centric AI*, COM(2019) 168 final (April 8, 2019).

European Commission, *Directorate-General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI,* Publications Office (2019). https://data.europa.eu/doi/10.2759/346720.

European Commission, *First Rules of the AI Act Are Now Applicable*, *Shaping Europe's Digital Future* (February 3, 2025), https://digital-strategy.ec.europa.eu/en/news/first-rules-artificial-intelligence-act-are-now-applicable.

European Commission, *Guidelines on the Definition of an AI System Established by the AIA* (2024), https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application.

European Commission, *High-Level Group on the Digital Markets Act: Public Statement on Artificial Intelligence*, Digital Markets Act (May 22, 2024). https://digital-markets-act.ec.europa.eu/high-level-group-digital-markets-act-public-statement-artificial-intelligence-2024-05-22_en

European Commission. *Liability Rules for Artificial Intelligence.* Last modified February 12, 2025. https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en.

European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on AI (AI Act) and Amending Certain Union Legislative Acts*, COM(2021) 206 final (April 21, 2021).

European Commission, A Competitiveness Compass for the EU, Brussels, 29.1.2025 COM(2025) 30 final, available at: https://commission.europa.eu/document/download/10017eb1-4722-4333-add2-e0ed18105a34_en .

European Commission, Commission sets course for Europe's AI leadership with an ambitious AI Continent Action Plan, press release (9 April 2025), available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1013 .

European Commission, Directorate General for Communication, Commission proposes to cut red tape and simplify business environment (26 February 2025), available at: https://commission.europa.eu/news/commission-proposes-cut-red-tape-and-simplify-business-environment-2025-02-26_en .

European Data Protection Supervisor. *Opinion 44/2023 on the Proposal for Artificial Intelligence Act in the Light of Legislative Developments* (October 23, 2023):11. https://www.edps.europa.eu/system/files/2023-10/2023-0137_d3269_opinion_en.pdf

European Parliament and Council. *Regulation (EU) No 536/2014 on Clinical Trials on Medicinal Products for Human Use and Repealing Directive 2001/20/EC* (Text with EEA relevance), OJ L 158, May 27, 2014, 1-76.

European Parliament and Council. *Directive (EU) 2024/2853 on Liability for Defective Products and Repealing Council Directive 85/374/EEC* (Text with EEA relevance), PE/7/2024/REV/1, OJ L 2024/2853, (November 18, 2024).

European Parliament and Council. *Regulation (EU) 2023/988 on General Product Safety, Amending Regulation (EU) No 1025/2012 and Directive (EU) 2020/1828, and Repealing Directive 2001/95/EC and Council Directive 87/357/EEC* (Text with EEA relevance), PE/79/2022/REV/1, OJ L 135, May 23, 2023, 1-51.

European Parliament and Council. *Regulation (EU) 2022/1925 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)* (Text with EEA relevance), PE/17/2022/REV/1, OJ L 265, October 12, 2022, 1-66.

European Parliament and Council. *Regulation (EU) 2022/2065 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act)* (Text with EEA relevance), PE/30/2022/REV/1, OJ L 277, October 27, 2022, 1-102.

European Parliament and Council. *Regulation (EU) 2017/746 on In Vitro Diagnostic Medical Devices and Repealing Directive 98/79/EC and Commission Decision 2010/227/EU* (Text with EEA relevance), OJ L 117, May 5, 2017, 176-332.

European Parliament and Council. *Regulation (EU) 2017/745 on Medical Devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002, and Regulation (EC) No 1223/2009, and Repealing Council Directives 90/385/EEC and 93/42/EEC* (Text with EEA relevance), OJ L 117, May 5, 2017, 1-175.

European Parliamentary Research Service, *Algorithmic discrimination under the AI Act and the GDPR* (February, 2025). https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2025)769509.

European Union, *Charter of Fundamental Rights of the European Union*, *Official Journal of the European Communities*, C 364/1 (December 18, 2000).

European Union, *Consolidated Version of the Treaty on the Functioning of the European Union (TFEU) Official Journal of the European Union* C 115 (May 9, 2008):94-95.

European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council on AI (AIA)*, *Official Journal of the European Union,* L 311/1 (July 12, 2024).

Farrell, Jimmy and Shamir, Afek, *An Introduction to Codes of Practice for the AI Act,* EU Artificial Intelligence Act, last modified October 30, 2024, https://artificialintelligenceact.eu/introduction-to-codes-of-practice/.

Gasser, Urs, and Mayer-Schönberger, Viktor. *Guardrails: Guiding Human Decisions in the Age of AI*. Princeton University Press, 2024. https://doi.org/10.2307/jj.6605394.

Gornet, Mélanie, and Maxwell, Winston. 2024. *The European Approach to Regulating AI through Technical Standards*. Internet Policy Review 13 (3), https://policyreview.info/articles/analysis/regulating-ai-through-technical-standards.

Grozdanovski, Ljupcho and De Cooman, Jérôme, *On the Obsolescence of Empirical Knowledge in the Age of AI, Rutgers Computer & Technology Law Journal* 49, no. 2 (2023): 123-145.

Haataja, Meeri and Joanna J. Bryson, *What Costs Should We Expect from the EU's AIA?*, *SocArXiv* (2021).

Haeck, Pieter EU opens door to reworking AI rulebook, Politico (9 April 2025), https://www.politico.eu/article/how-eu-did-full-180-artificial-intelligence-rules/.

Ho-Dac, Marion, *Considering Fundamental Rights in the European Standardisation of AI: Nonsense or Strategic Alliance?*, *arXiv* (January 23, 2024).

Holst, Luca *et al., The Impact of the EU AIA's Transparency Requirements on AI Innovation*, in *Proceedings of the 19th International Conference on Wirtschaftsinformatik*, Würzburg, Germany (2024).

International Association of Privacy Professionals, *EU AI Act Stakeholder Map*. IAPP (May 2024). https://iapp.org/resources/article/eu-ai-act-stakeholder-map/.

Jones, Joe, *EU AIA: Next Steps for Implementation*, *International Association of Privacy Professionals* (January 2025), https://iapp.org/resources/article/eu-ai-act-timeline/.

Kantner, Cathleen, *What Is a European Identity? The Emergence of a Shared Ethical Self-Understanding in the European Union*, *EUI RSCAS Working Papers* 2006/28 (2006).

Kasirzadeh, Atoosa, *Two Types of AI Existential Risk: Decisive and Accumulative*. arXiv preprint arXiv:2401.07836 (2024).

van Kolfschooten, Hannah, *EU Regulation of AI: Challenges for Patients' Rights,* Common Market Law Review 59, no. 1 (2022): 81-112.

van Kolfschooten, Hannah, and Shachar, Carmel, *The Council of Europe's AI Convention (2023-2024): Promises and Pitfalls for Health Protection,* Health Policy 138 (December 2023): 104935, https://doi.org/10.1016/j.healthpol.2023.104935.

Mann, Sebastian Porsdam, Cohen, I. Glenn and Minssen, Timo, *The EU AIA: Implications for U.S. Health Care,* NEJM AI 1, no. 11 (October 2024), https://ai.nejm.org/doi/pdf/10.1056/AIp2400449.

Marsden, Christopher T. and Christou, George *Artificially Intelligent Regulation: Global Norms, International Political Economy and the Brussels Effect*, in *Proceedings of the 2024 International Conference on AI and Regulation* (2024).

Ministry of Foreign Affairs of Japan, *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems* (October 30, 2023).

Moille, Céline, Popescu, Adelina and Scott, Gretchen, *EU AIA Implementation Timeline*, *Goodwin Insights* (October 10, 2024), https://www.goodwinlaw.com/en/insights/publications/2024/10/insights-technology-aiml-eu-ai-act-implementation-timeline.

Neergaard, Ulla and de Vries, Sybe, *The Interaction between Free Movement Law and Fundamental Rights in the (Digital) Internal Market,* Renforce Working Paper 1-2023 (2023): 1-30.

O'Regan, Ellen, Europe's GDPR privacy law is headed for red tape bonfire within 'weeks', Politico (3 April 2025), available at: https://www.politico.eu/article/eu-gdpr-privacy-law-europe-president-ursula-von-der-leyen/.

Organisation for Economic Co-operation and Development, *OECD Framework for the Classification of AI Systems*, Paris: *OECD Publishing* (2022).

Orwat, Carsten *et al., Normative Challenges of Risk Regulation of AI*, NanoEthics 18, no. 2 (2024): 11, https://doi.org/10.1007/s11569-024-00454-9.

Overseas Development Institute. *International AI Governance Must Be Truly Global*, ODI (October 25, 2023). https://odi.org/en/insights/international-ai-governance-must-be-truly-global/.

Pelkmans, Jacques and Renda, Andrea *Does EU Regulation Hinder or Stimulate Innovation?*, *CEPS Special Report No. 96* (November 19, 2014), Brussels: *Centre for European Policy Studies*.

Podszun, Rupprecht, *From Competition Law to Platform Regulation: Regulatory Choices for the Digital Markets Act,* Economics 17, no. 1 (2023): 20220037, https://doi.org/10.1515/econ-2022-0037.

Price II, W. Nicholson and Cohen, I. Glenn, *Locating Liability for Medical AI,* DePaul Law Review 73, no. 2 (2024): 339-368.

PYMNTS, *European Commission Withdraws Draft Rules on Technology Patents, AI Liability and Consumer Privacy* (February 12, 2025), https://www.pymnts.com/news/regulation/2025/european-commission-withdraws-draft-rules-on-technology-patents-ai-liability-and-consumer-privacy/.

Shao, Shuai *et al., Environmental Regulation and Enterprise Innovation: A Review,* Business Strategy and the Environment 29, no. 3 (2020): 1465-1478, https://doi.org/10.1002/bse.2446.

Siegmann, Charlotte and Anderljung, Markus, *The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market*, Centre for the Governance of AI (Aug. 2022). https://cdn.governance.ai/Brussels_Effect_GovAI.pdf

Tallberg, Jonas, Lundgren, Magnus and Geith, Johannes, *AI Regulation in the European Union: Examining Non-State Actor Preferences*, Business and Politics 26, no. 2 (2024): 241-267, https://doi.org/10.1017/bap.2024.10.

Teo, Sue Anne, *How Artificial Intelligence Systems Challenge the Conceptual Foundations of the Human Rights Legal Framework*, Nordic Journal of Human Rights 40, no. 1 (2022): 216-34, https://doi.org/10.1080/18918131.2022.2073078.

The White House. *Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.* Statements and Releases (October 30, 2023). https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

Trump, Donald J., *Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence*, Federal Register 90, no. 5469 (January 23, 2025): 5469-5471. https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence.

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

Wachter, Sandra*, Limitations and Loopholes in the EU AIA and AI Liability Directives: What This Means for the European Union, the United States, and Beyond,* Yale Journal of Law & Technology 26, no. 3 (2024):685.

GENERAL FRAMEWORK OF REGULATION IN THE PEOPLE'S REPUBLIC OF CHINA ON ARTIFICIAL INTELLIGENCE AND DIGITAL TRANSFORMATION

SUMMARY

I. Introduction **II.** Laws regulating AI in China **III.** Measures regulating AI in China **IV.** Provisions on the management of algorithmic recommendations in internet information services **V.** Other sectoral AI regulations and technical standards in China **VI.** AI legislation proposals by leading experts **VII.** Conclusion

> HUI ZHOU

Deputy director, associate professor. Department of cyber and information law, Institute of law, Chinese Academy of Social Sciences

> YUE ZHU

Assistant professor of law at Tongji University

I. INTRODUCTION

The rapid advancement of artificial intelligence (AI) has significantly transformed various sectors worldwide, prompting governments to develop regulatory frameworks that address both the opportunities and challenges presented by this technology. In China, the government has been proactive in establishing various laws and regulations to guide the development, deployment, and use of AI. These legal instruments aim to promote innovation while safeguarding national security, protecting personal rights, and ensuring ethical standards are upheld.

This brief provides an in-depth examination of China's AI regulatory landscape, beginning with foundational laws such as the CSL, DSL, and PIPL. These laws set the groundwork for data handling, personal information protection, and cybersecurity, which are crucial for the responsible development of AI technologies.

Further, the brief explores specific measures tailored to address the nuances of AI technologies, including the Provisions on the Management of Algorithmic Recommendations in Internet Information Services, the Provisions on the Administration of Deep Synthesis Internet Information Services, the Interim Measures for the Management of Generative Artificial Intelligence Services, and the Measures for the Labeling of AI-Generated Synthetic Content (official draft for suggestion). These regulations focus on algorithmic transparency, content authenticity, and the ethical deployment of AI applications.

Additionally, the brief quite sketchily presents sector-specific regulations and technical standards that guide AI applications in industries such as finance, healthcare, transportation, and education. These sectoral regulations complement the broader legal framework by addressing industry-specific risks and requirements, ensuring that AI technologies are applied responsibly within different contexts.

Finally, the brief discusses proposed AI legislation, namely the MAIL and AIL, which offer insights into the potential future direction of AI governance in China. These proposals similarly emphasize ethical considerations, innovation promotion, and the protection of individual rights, reflecting China's consensual commitment to shaping a legal environment conducive to responsible AI development.

By analyzing all these instruments, the brief aims to provide a comprehensive understanding of how China navigates the complex landscape of AI governance. It highlights the country's efforts to balance the encouragement of technological innovation with the imperative to protect various rights and interests of state, society, and individuals, thereby contributing to the global discourse on AI regulation.

II. LAWS REGULATING AI IN CHINA

In China, several laws have been enacted to regulate various aspects of AI, particularly focusing on network and data security, and personal information protection. This section examines the AI-related provisions within three key Chinese laws: As mentioned above, the CSL, the DSL, and the PIPL. These laws collectively shape the regulatory landscape for AI in China by establishing guidelines for data handling, personal information protection, and the legitimate application of AI.

1. CSL

The Cybersecurity Law, effective since June 2017, serves as the foundational legislation governing cybersecurity in China. While it does not explicitly mention AI, its provisions have significant implications for AI systems, particularly concerning data security, network security, and the obligations of network operators.

Key AI-Relevant Provisions:

Obligations of Network Operators: Network operators must implement security measures to protect their networks and data from unauthorized access, interference, or destruction. They are required to adopt technical measures to prevent computer viruses and cyber-attacks. For AI systems, this means that certain entities developing and/or deploying AI technologies must ensure robust cybersecurity measures are in place to protect data integrity and prevent malicious exploitation of AI algorithms.

Network Product Security: Providers of network products and services shall ensure compliance with mandatory requirements set forth by national standards and must not include malicious programs. When security flaws or vulnerabilities are discovered, they must immediately adopt remedial measures and inform users. AI products and services covered herein shall adhere to these security standards, ensuring the integrity and reliability of AI applications.

Prohibition of Malicious Activities: Individuals and organizations are prohibited from engaging in certain activities that harm cybersecurity, including providing programs or tools used for network intrusions. This may encompass AI tools, including frontier AI tools, that could be used maliciously, potentially placing legal responsibilities on developers and distributors of AI technologies.

Critical Information Infrastructure Security: Operators of critical information infrastructure (CII) have additional security obligations, including conducting security assessments and protecting important data. AI systems classified as part of CII must implement enhanced security measures, potentially affecting sectors like finance, transportation, and healthcare where AI is increasingly integrated.

Implications for AI:

The CSL establishes a legal framework that affects AI development and deployment in China. AI systems shall therefore comply with stringent cybersecurity measures. AI companies must ensure that their network and data handling practices align with the obligations set forth in the law. The emphasis on network and data security directly impacts how AI technologies are developed, implemented, and managed.

2.DSL

Effective from September 2021, the Data Security Law provides a comprehensive framework for data handling and security in China. While not exclusively focused on AI, the law's emphasis on data governance directly impacts AI systems that rely heavily on data collection and processing.

Key AI-Relevant Provisions:

Definition of Data Handling: The law defines data handling as the collection, storage, use, processing, transmission, provision, and disclosure of data. This broad definition may encompass many AI activities involving data, making AI systems subject to the law's requirements.

Categorized and Graded Data Protection System: The law establishes a system to categorize and grade data based on its importance to national security, public interests, or individual rights. AI systems processing important or core national data must adhere to stricter security measures, including enhanced protection protocols and potential restrictions on data handling practices.

Promotion of Data Development and Use: The law encourages the advancement of data infrastructure and supports innovative applications of data in all industries. It promotes research into data development and use technologies, fostering AI innovation while ensuring data security. AI companies can potentially leverage state support for research and development, aligning their innovation strategies with national priorities.

Data Security Obligations: Entities handling data must establish data security management systems and adopt necessary technical measures to ensure data security. AI companies must implement robust data protection mechanisms throughout the data lifecycle, including data encryption, access controls, and regular security assessments.

The law also establishes systematic rules for data security risk assessment, reporting, information sharing, monitoring, and early warning.

Implications for AI:
The Data Security Law imposes requirements on data handling practices, directly affecting AI systems that depend on large datasets. AI operators covered shall accordingly implement comprehensive data security measures and comply with data classification and grading requirements. The law seeks to balance the promotion of data-driven innovation with the protection of national security and individual rights. AI companies must be vigilant in their data governance strategies to align with these legal obligations.

3. PIPL

In effect since November 2021, the PIPL is China's first comprehensive law dedicated to personal data protection. It introduces specific provisions related to AI, particularly concerning automated decision-making and the rights of individuals.

Key AI-Relevant Provisions:
Definition of Personal Information and Automated Decision-Making: Personal information is defined as information related to identified or identifiable natural

persons. Automated decision-making is defined as the use of \programs to automatically analyze or assess personal behaviors and make decisions based on that analysis. AI systems that process personal data for decision-making probably fall under these definitions and are subject to the law's provisions.

Transparency and Fairness in Automated Decision-Making: Personal information handlers using automated decision-making must ensure transparency, fairness, and justice in their processes. They must not engage in unreasonable differential treatment of individuals in terms of trade conditions such as pricing. This provision addresses concerns about AI algorithms potentially leading to biased or unfair outcomes, mandating that AI systems be designed and operated in a manner that is equitable.

Right to Explanation and Refusal: When automated decision-making produces decisions with a significant impact on individuals' rights and interests, individuals have the right to request an explanation and to refuse the processing based solely on automated decision-making. This empowers individuals to challenge AI-driven decisions that affect them, compelling AI developers to provide mechanisms for human oversight and intervention.

Consent for Sensitive Personal Information: Handling sensitive personal information generally requires obtaining separate consent from individuals. AI systems processing such data must ensure they have the necessary consent and implement strict protection measures, including enhanced security protocols and access restrictions.

Obligations for Personal Information Handlers: Entities handling personal information must adopt measures to prevent unauthorized access, leaks, or misuse. They must conduct personal information protection impact assessments, particularly when using personal data for automated decision-making. Large platform operators with complex business models have additional obligations, such as establishing independent bodies to supervise personal information protection and publishing social responsibility reports. This promotes greater accountability and transparency in AI operations.

Cross-Border Data Transfer: The PIPL imposes conditions on transferring personal information outside of China, including passing security assessments or obtaining certifications. AI systems that rely on cross-border data processing must comply with these requirements, potentially impacting international AI collaborations and data exchange.

Implications for AI:
The PIPL has significant implications for AI systems that handle personal data. It emphasizes individuals' rights over their personal information, including the right to explanation and refusal of automated decision-making. AI developers in many circumstances must ensure transparency in their algorithms, avoid biases, and implement mechanisms to allow individuals to exercise their rights. The stringent consent requirements for sensitive data necessitate careful data handling practices. Furthermore, AI companies must establish robust compliance programs, conduct regular audits, and stay informed of legal developments to mitigate risks.

Combined Implications for AI:
The interplay of these three laws creates a comprehensive regulatory environment for AI in China. AI systems must navigate complex requirements. Key considerations may include:

Data governance. AI relies on vast amounts of data. Companies must implement robust data management systems, ensure data quality, and comply with data classification and protection requirements. This includes adhering to the categorized and graded protection system for data and ensuring secure data storage and processing.

Algorithmic transparency and fairness. The emphasis on transparency and fairness in automated decision-making requires AI developers to address algorithmic biases and provide explanations for AI-driven decisions, aligning with global concerns about ethical AI. AI systems must be designed to prevent discrimination and allow individuals to understand and challenge decisions affecting them.

Cross-border data challenges. Restrictions on cross-border data transfer impact AI collaborations and operations involving international data flows. Companies must assess their data transfer practices and explore mechanisms such as security assessments or obtaining certifications to legally transfer data abroad if necessary for global collaboration.

Innovation, While the laws impose obligations, they also support AI innovation through state initiatives. Companies can align their research and development efforts with national strategies to leverage support and contribute to China's AI development goals. The promotion of data development and use technologies provides opportunities for AI advancement within the legal framework.

III. MEASURES REGULATING AI IN CHINA

In addition to the foundational laws governing cybersecurity, data security, and personal information protection in China, several specific measures have been introduced to further regulate the development and deployment of AI. These measures address AI in specific contexts such as algorithmic recommendation, deep synthesis, generative AI, and the labeling of AI-generated content. This section provides a comprehensive summary of four key regulatory instruments: the Provisions on the Management of Algorithmic Recommendations in Internet Information Services, the Provisions on the Administration of Deep Synthesis Internet Information Services, the Interim Measures for the Management of Generative Artificial Intelligence Services, and the Measures for Labeling of AI-Generated Synthetic Content (official draft for suggestion). These measures collectively aim to promote the healthy and orderly development of AI in China while safeguarding national security, social order, and individual rights.

IV. PROVISIONS ON THE MANAGEMENT OF ALGORITHMIC RECOMMENDATIONS IN INTERNET INFORMATION SERVICES

Effective from March 1, 2022, the Provisions on the Management of Algorithmic Recommendations in Internet Information Services establish a regulatory framework for the use of algorithmic recommendation technologies in internet information services. These technologies may include algorithms that generate and synthesize content, provide individualized pushes, filter searches, and make scheduling decisions to offer users information tailored to their preferences.

Key AI-Relevant Provisions:

Compliance and Ethical Use: Service providers must adhere to laws and regulations, respect social norms and ethics, and follow principles of fairness, transparency, and accountability. They are required to prevent the use of algorithmic recommendations for activities that endanger national security, disrupt social order, or infringe on the lawful rights of others.

Entity Responsibility and Management Systems: Providers must implement entity responsibility for algorithm security, establishing management systems for algorithm mechanism checks, scientific and technological ethics reviews, user registration, data security, and personal information protection. They must

allocate professional staff and technical support proportional to the scale of their services.

Algorithm Review and Assessment: Providers are required to periodically check, assess, and verify algorithm mechanisms, models, data, and outcomes. They must ensure that algorithms do not violate laws or ethical standards, such as inducing users to become addicted or overspend.

Information Security Management: Providers must strengthen the management of information security by creating a database of characteristics used to identify illegal and negative content. They must prevent the transmission of prohibited information and ensure that algorithmically generated or synthesized content is conspicuously labeled if required.

User Modeling and Tag Management: Providers must manage user models and tags responsibly, ensuring that unlawful or negative information is not used to profile users or push content to them.

Transparency and User Choice: Providers must inform users conspicuously about the algorithmic recommendation services they offer, including basic principles and main operation mechanisms. They must provide users with options to disable personalized recommendations and to select or delete user tags that influence the algorithmic recommendations.

Protection of Special Groups: Special considerations are mandated for services provided to minors, seniors, and laborers. For minors, providers must prevent content that could negatively impact their physical and psychological health and avoid inducing addiction. For seniors, services must consider their needs and safeguard their rights, including protection against fraud. For laborers, providers must ensure that algorithms related to work assignments and compensation protect lawful rights such as fair wages and reasonable working hours.

Fair Transaction Rights: In commercial transactions, providers must not use algorithms to conduct unreasonable differential treatment of consumers, such as discriminatory pricing based on consumer preferences or behavior.

Oversight and Filing Requirements: Certain poviders must perform filing procedures, including submitting details about their algorithmic services. Authorities establish a management system for oversight.

Implications for AI:

This measure places significant responsibility on AI developers and service providers to ensure that their algorithmic recommendation systems are transparent, ethical, and compliant with legal standards. It emphasizes user rights,

including the right to understand and control how algorithms influence the information they receive. AI systems must also avoid discriminatory practices and protect vulnerable groups, aligning with broader concerns about algorithmic fairness and accountability.

Provisions on the Administration of Deep Synthesis Internet Information Services

Effective from January 10, 2023, the Provisions on the Administration of Deep Synthesis Internet Information Services specifically regulate technologies known as "deep synthesis", which include deep learning and virtual reality techniques used to generate or manipulate content such as text, images, audio, and video. This includes technologies that can create realistic virtual scenes or manipulate or alter biometric information, often referred to in the context of "deepfakes".

Key AI-Relevant Provisions:

Prohibition of Illegal Content and Activities: Service providers and users are prohibited from using deep synthesis technologies to produce or disseminate content that is illegal, harmful to national security, or infringes upon the rights of others. Specifically, they must not create or spread fake news using deep synthesis technologies.

Information Security Responsibility: Providers must implement primary responsibility for information security, establishing management systems for user registration, algorithm review, data security, and personal information protection. They must have safe and controllable technical safeguard measures in place.

Real Identity Verification: Providers must verify the real identity of users through methods such as mobile numbers or identification documents, ensuring accountability and traceability of users who access deep synthesis services.

Content Review and Management: Providers must strengthen the management of content generated through deep synthesis, reviewing user inputs and synthesis outcomes. They must establish a database of characteristics to identify illegal and negative content and take appropriate measures when such content is discovered.

Mechanisms for Dispelling Rumors: Providers must establish mechanisms to address and dispel misinformation or rumors generated through deep synthesis technologies, reporting such incidents to relevant authorities.

Training Data Management: Providers and technical supporters must manage training data responsibly, ensuring data security and compliance with

personal information protection laws. If biometric information is involved, they must obtain consent from the individuals concerned.

Security Assessments: Providers offering tools with certain functionalities, such as generating or editing biometric information, must conduct security assessments either independently or through professional bodies.

Mandatory Labeling: Providers must employ technical measures to attach identifiers to content generated or edited using deep synthesis technologies. Prominent labels are required for certain types of content that might cause confusion or mislead the public, such as AI-generated text, voice, or images that simulate real people.

User Appeals and Complaints: Providers must set up accessible channels for user appeals, complaints, and reports, and must handle them promptly, providing feedback on the outcomes.

Implications for AI:

The provisions address the potential misuse of AI technologies that can create realistic synthetic content, especially so-called "deepfake" content, which poses risks such as misinformation, fraud, and infringement of personal rights. AI developers and service providers must ensure transparency through labeling, obtain necessary consents, and implement robust security and ethical measures in their technologies. The measures aim to balance innovation with the need to prevent the negative societal impacts of deep synthesis technologies.

Interim Measures for the Management of Generative Artificial Intelligence Services

Effective from August, 2023, the Interim Measures for the Management of Generative Artificial Intelligence Services provide regulatory guidelines for the development and use of generative AI technologies that produce content such as text, images, audio, and video for public consumption.

Key AI-Relevant Provisions:

Scope and Application: The measures apply to the use of generative AI technologies to provide services to the public within China. They emphasize the balance between promoting innovation and ensuring security, indicating a tolerant and cautious approach to regulation.

Compliance and Ethical Standards: Providers must adhere to laws and administrative regulations, respect social ethics, and uphold mainstream social values. They must prevent the generation of prohibited content, such as that

which endangers national security, promotes extremism, or contains misinformation.

Non-Discrimination: During all stages of AI development and deployment, providers must prevent discrimination based on factors like race, ethnicity, gender, or health. This aligns with global concerns about AI fairness and the prevention of biased outcomes.

Intellectual Property and Fair Competition: Providers must respect intellectual property rights and commercial ethics, avoiding monopolistic practices or unfair competition by exploiting advantages in algorithms or data.

Protection of Personal Rights: Providers must prevent the infringement of others' lawful rights, including rights to image, reputation, privacy, and personal information. They must avoid content that could harm the physical or psychological well-being of individuals.

Data and Model Requirements: Providers must ensure that the data and foundational models used in training have lawful sources. They must respect intellectual property rights and obtain consent when personal information is involved. Measures must be taken to improve the quality, truthfulness, and diversity of training data.

User Privacy and Data Protection: Providers are responsible for the confidentiality of user input and usage records. They must not collect unnecessary personal information or unlawfully provide user data to others. They are obligated to respond to user requests regarding personal information access, correction, or deletion.

Content Labeling: Providers must label AI-generated content, especially images and videos, in accordance with existing provisions on deep synthesis services. This ensures transparency and allows users to distinguish between AI-generated and human-generated content.

Complaint and Reporting Mechanisms: Providers must establish mechanisms for handling public complaints and reports, ensuring that issues are addressed promptly and feedback is provided.

Security Assessments and Filings: Certain providers must conduct security assessments and comply with filing requirements.

Implications for AI:

These measures place comprehensive obligations on providers of generative AI services to ensure legal compliance, ethical conduct, and the protection of individual rights. For example, the emphasis on data quality and lawful sourcing addresses concerns about AI models producing inaccurate or harmful content

due to poor training data. Providers are encouraged to innovate responsibly while being held accountable for the societal impacts of their technologies.

Measures for Labeling of AI-Generated Synthetic Content (Official Draft for Suggestion)

As a draft measure not yet in effect, the Measures for Labeling of AI-Generated Synthetic Content aim to standardize the labeling practices for AI-generated content, enhancing transparency and protecting users from potential deception or misinformation.

Key AI-Relevant Provisions:

Scope and Definitions: The measures apply to online information service providers offering services that generate synthetic content using AI technologies. AI-generated synthetic content includes text, images, audio, video, or other information produced using AI.

Labeling Requirements: Service providers must add both explicit and implicit labels to AI-generated content. Explicit labels are perceivable by users and must be added in appropriate positions depending on the content type (e.g., text notifications for text content, conspicuous symbols on images). Implicit labels are embedded in the metadata of content files, including information such as the provider's name and content reference number.

Verification and Regulation by Platforms: Online content transmission platforms must verify whether uploaded content contains implicit labels and add conspicuous notification labels when necessary. They are responsible for alerting users if content is AI-generated or suspected to be so.

User Obligations: Service providers must explain labeling methods and specifications in user agreements, ensuring users are aware of their obligations. Users are prohibited from maliciously deleting, altering, or concealing labels and must not harm others' rights through improper labeling.

Compliance with Standards: Service providers must conduct labeling in accordance with another mandatory national standards, which is also soliciting public suggestion, therefore ensuring consistency and compliance across the industry.

Implications for AI:

By mandating clear labeling of AI-generated content, the draft measures aim to enhance transparency and allow users to identify and assess content appropriately. This is particularly important in combating misinformation and

rumor and ensuring that users are not misled by synthetic content that appears authentic. The measures encourage responsible use of AI technologies and hold both service providers and users accountable for maintaining transparency.

Combined Implications for AI

The four measures collectively reinforce China's commitment to regulating AI technologies in a manner that promotes innovation while safeguarding national security, social order, and individual rights. Key themes across these measures include:

Transparency and Accountability. Providers of AI services are required to be transparent about their algorithms, data sources, and the nature of the content they generate or recommend. Mandatory labeling of AI-generated content ensures that users are aware when they are interacting with synthetic media.

Content Management and Security: Providers are responsible for preventing the dissemination of illegal or harmful content, including misinformation, extremism, and content that could negatively impact vulnerable groups such as minors and seniors. Robust content review mechanisms and security measures are mandated.

Ethical Use of AI Technologies: The measures stress adherence to ethical standards, including respect for intellectual property rights, prevention of unfair competition, and avoidance of practices that could lead to addiction or exploitation of users.

Protection of Individual Rights. There is a strong emphasis on protecting personal information, preventing discrimination, and safeguarding the lawful rights and interests of individuals. AI systems must be designed and operated to respect privacy, image rights, and prevent harm.

Regulatory Oversight and Compliance: Providers must comply with filing requirements, security assessments, and cooperate with regulatory authorities. Non-compliance can result in penalties, including fines and suspension of services.

V. OTHER SECTORAL AI REGULATIONS AND TECHNICAL STANDARDS IN CHINA

In addition to the overarching laws and specific measures previously discussed, China has implemented, and is enacting, various sector-specific legal instruments and technical standards to further guide the development and application of AI across different industries. These sectoral regulations complement the broader legal framework established by the laws and measures mentioned above. A surely incomplete list of the most important contexts include:

1. Financial Sector Regulations. Banks, insurance companies, and other financial institutions are encouraged to use AI to improve service efficiency while ensuring compliance with cybersecurity and data protection laws. Strict security measures are required when using AI technologies for activities like credit scoring, fraud detection, and personalized financial services. Compliance with data localization and cross-border data transfer regulations is crucial.
2. Healthcare Sector Regulations. AI systems used for medical and diagnostic support must meet stringent safety standards, and patient data must be protected in accordance with personal information protection laws. It is worthwhile to mention that the National Medical Products Administration provides guidance on the registration and evaluation of AI-based medical devices, ensuring they meet quality and safety requirements before entering the market.
3. Autonomous Vehicles and Transportation. The Ministry of Industry and Information Technology has issued policies to promote the development of autonomous driving technologies. Regulations focus on safety testing, data security, and the protection of passenger information. Companies testing autonomous vehicles must obtain necessary permits and ensure compliance with traffic laws. Besides, several local governments, such as Beijing, have implemented specific guidelines for the testing and deployment of autonomous vehicles on public roads, including requirements for safety drivers and reporting of test results.

Technical standards also play a crucial role in guiding the development and deployment of AI technologies. China has been actively developing national standards to ensure consistency, safety, and interoperability of AI systems. China also actively participates in international standardization organizations like the International Organization for Standardization (ISO) and the International

Electrotechnical Commission (IEC), contributing to global AI standards and aligning domestic standards with international practices.

All these sectoral regulations and technical standards complement the general laws and measures previously discussed. They provide detailed requirements tailored to specific applications of AI, ensuring that industry-specific risks are managed effectively. Together, they form a comprehensive, well-balanced, and indeed workable regulatory ecosystem. Entities operating within these sectors must not only comply with the general laws and measures but also adhere to the specific regulations and technical standards relevant to their industry. This multi-layered approach ensures that AI technologies are developed and applied in a manner that is consistent with China's policy objectives and legal requirements. As AI continues to evolve, these regulations and standards will play a crucial role in guiding stakeholders toward sustainable and socially beneficial AI innovation.

VI. AI LEGISLATION PROPOSALS BY LEADING EXPERTS

Chinese legal experts have proposed comprehensive legislative frameworks to guide the responsible development and application of AI. Two notable expert proposals are the MAIR and AIR. While neither is a binding legal instrument, they offer valuable insights into the potential future direction of AI regulation in China. This section provides a summary of these two proposals, and a comparative analysis of their key features.

Overview of the MAIL

The MAIL is the earliest expert draft proposal aiming to suggest for a comprehensive legal framework for AI governance in China.

Key Provisions:

Scope and Application: The law applies to AI-related activities within China and extends to activities outside China that affect its national security or the rights of its citizens.

Principles of AI Governance: It outlines several guiding principles, including: Human-Centric Approach, AI should promote human welfare and remain under human supervision and control; Safety and Security, necessary measures must be taken to ensure the safety and security of AI systems and related data; Openness and Transparency, AI providers should appropriately label AI systems and clarify their purposes, principles, and effects; Accountability, AI developers,

providers, and users are responsible for their respective activities; Fairness and Equality, AI should avoid unreasonable differential treatment and consider the needs of special groups like minors and the elderly; and, Green Development, encouraging the use of energy-saving technologies in AI development.

Promotion of AI Development: Infrastructure Support, the state supports building AI infrastructure, including public computing resources and data centers; Algorithm and Model Innovation, encouraging open-source platforms and the protection of foundational models; Data Supply, supporting constructing foundational and specialized databases, promoting data sharing and ensuring data quality; Industrial Development, accelerating the application of key AI technologies in various industries; and, Talent Cultivation, supporting educational institutions in improving AI-related disciplines and research.

Regulatory Framework: Negative List Oversight System, establishing an "AI negative list" for AI applications that require licensing due to potential risks; Licensing Requirements, outlining conditions for obtaining licenses, including legal entity status, security measures, and compliance systems; Obligations of AI Developers and Providers, specifying safety obligations, risk management, transparency, explainability, fairness, and ethical review requirements; and, Special Obligations for Foundation Models, developers of foundational AI models have additional responsibilities, such as ensuring security, openness, and assisting others in compliance.

Comprehensive Governance Mechanism: Regulatory Authorities, establishing a China Administration of Artificial Intelligence as the main authority for AI governance; Collaborative Governance, promoting a governance mechanism involving government, corporations, industry self-regulation, and user self-discipline; and, Expert Committees, establishing national and local committees of experts on AI ethics.

Liability and Penalties: Administrative Penalties, specifying fines and corrective measures for violations; Civil Liability, outlining tort liability for damages caused by AI systems; Exemptions, providing liability exemptions for open-source AI providers under certain conditions; and, International Cooperation, encouraging participation in international AI rule-making and cooperation, aiming to develop global governance frameworks.

Overview of the AIL

The AIL is another influential proposal by Chinese academics that seeks to establish a legislative framework for AI governance.

Key Provisions:

Scope and Application: Applies to AI development, provision, and use within China, and to activities outside China that affect its national security or citizens' rights.

AI Regulation Principles: Scientific and Technological Ethics, AI development should be people-centered, respect personal freedom and dignity, and promote public welfare; Innovative Development, implementing inclusive and prudent supervision, encouraging AI industry innovation while ensuring safety; Fairness and Impartiality, Protecting legitimate rights and interests, avoiding unreasonable differential treatment; Transparency and Explainability, requiring AI to be transparent and explainable, providing users with necessary information; Safety and Accountability, AI entities must adopt measures to ensure safety and bear corresponding legal liabilities; and, Proper Use, AI must comply with laws, respect ethics, and not harm national security or public interests.

Promotion of AI Development: Infrastructure Development, strengthening AI infrastructure, including computing resources and public data platforms; Intellectual Property Protection, protecting AI-related intellectual property rights and encourages innovation; Open Source Ecosystem, supporting building open-source platforms and communities; Data Utilization, encouraging data sharing and reasonable use of data for AI development; and, Industry-Academia Integration, promoting collaboration between enterprises, educational institutions, and research bodies.

Protection of Users' Rights and Interests: Equal Rights, ensuring universality and fairness in AI products and services; Right to Know, providing users with clear information about AI services, including functions, purposes, and risks; Privacy Protection, safeguarding personal information and prohibits unlawful data retention and provision to others; Right to Explanation and Refusal, users can request explanations of AI decisions and have the right to refuse or withdraw from AI services; and, Protection of Vulnerable Groups, considering the needs of minors, the elderly, and people with disabilities.

Regulatory Framework: Main Oversight Departments, assigning responsibilities to relevant government departments for AI supervision; Grading and Categorization, establishing a system to grade and categorize AI based on importance and potential harm; Special Supervision of Critical AI, implementing special management for critical AI applications, such as those affecting personal rights or used in critical infrastructure; Risk Monitoring and Emergency Response, developing mechanisms for AI security monitoring and incident response.

Obligations of AI Entities: Security Obligations, AI developers and providers must ensure safety, conduct risk assessments, and report major incidents; Content Security, must prevent the generation of illegal or harmful content and add identifiers to AI-generated content. Compliance Guidelines: Ensure certain products and services comply with laws and regulations.

Special Application Contexts, including but not limited to: Use by State Organs, AI decisions can only be references for administrative actions; Judicial AI: AI assists in judicial work but cannot replace human judgment; Medical AI: AI decisions serve as references in medical activities; Autonomous Driving: Requires appropriate licenses and prioritizes safety.

Liability and Penalties: Administrative Penalties, specifying fines and corrective measures for violations; Civil Liability, outlining tort liability for damages, including punitive damages in severe cases; and Safe Harbor Rules, providing conditions under which AI service providers are exempt from liability.

Preliminary Comparative Analysis

While both proposals aim to establish comprehensive AI legislation in China, there are a few notable consensuses and differences in their approaches and emphases.

Similarities:

- Comprehensive Scope: Both laws cover AI development, provision, and use within China and extend to activities affecting China from abroad.
- Guiding Principles: Both emphasize ethical considerations, including human-centric development, safety, transparency, fairness, and accountability.
- Promotion of Innovation: Both support AI innovation, infrastructure development, data sharing, and open-source ecosystems.
- Protection of Rights: Both prioritize the protection of individual rights, including privacy, personal information, and the rights of vulnerable groups.
- Regulatory Oversight: Both establish regulatory frameworks involving government authorities responsible for AI governance.
- Liability Provisions: Both outline legal liabilities for AI developers, providers, and users, including administrative penalties and civil liabilities.
- International Cooperation: Both recognize the importance of participating in global AI governance and promoting international cooperation.

Differences:

Regulatory Structure:

- MAIL: Establishes a specific authority, the China Administration of Artificial Intelligence (CAAI), as the main regulatory body, and introduces a Negative List oversight system requiring licenses for certain AI applications.
- AIL: Distributes regulatory responsibilities among relevant government departments without specifying a central authority and emphasizes a grading and categorization system for AI oversight.

Licensing and Oversight Mechanisms:

- MAIL: Requires licenses for AI applications on the Negative List and details the licensing process, conditions, and obligations.
- AIL: Focuses on registration and special supervision for critical AI but does not specify a licensing system like the Negative List.

Obligations for Foundation Models:

- MAIL: Imposes specific obligations on developers of foundational AI models, including security measures, openness, and assisting others in compliance.
- AIL: Mentions foundation models in the context of critical AI but provides less detail on specific obligations.

Promotion of Open Source:

- MAIL: Encourages open-source development and the establishment of open-source AI foundations.
- AIL: Supports building open-source platforms but places more emphasis on intellectual property protection and reasonable use of data.

Special Application Scenarios:

- MAIL: Does not elaborate extensively on specific application scenarios but includes general provisions for AI use by state organs.
- AIL: Provides detailed provisions for specific AI applications, such as judicial AI, medical AI, social bots, biometric recognition, and autonomous driving.

Liability and Exemptions:

- MAIL: Offers liability exemptions for open-source AI providers under certain conditions and details about administrative penalties.
- AIL: Introduces safe harbor rules for service providers and includes punitive damages for severe violations.

Both the MAIL and AIL reflect China's proactive approach to AI governance, aiming to balance innovation with regulation and ethical considerations. While sharing common goals, they differ in regulatory structures, emphasis on licensing mechanisms, obligations for foundational models, and approaches to liability and international measures. Both proposals contribute valuable perspectives to the ongoing discourse on AI legislation in China. As China continues to refine its AI governance strategies, these expert contributions will possibly contribute to future legislation, shaping the country's role in the global AI landscape.

VII. CONCLUSION

China's regulatory approach to AI reflects its commitment to both advancing innovation and upholding rights and interests of state, society, and individuals in an increasingly complex digital landscape. Through a multifaceted and sector-specific framework of foundational laws, specific regulatory measures, sectoral guidelines, and technical standards, China seeks to balance the benefits of AI with the responsibilities it entails. In sum, China's AI regulatory landscape provides a robust, open, and very promising regulatory framework. As the global AI landscape evolves, China's approach will probably influence international norms and inspire continued discourse on achieving responsible, ethical, and sound AI development worldwide.

ARTIFICIAL INTELLIGENCE: WHAT CHALLENGES SHOULD AI REGULATION IN MEXICO ADDRESS?

SUMARIO:

I. Introduction **II**. Algorithmic harms and regulation based on human rights frameworks **III.** Bias and the representation of minoritized groups in artificial intelligence **IV.** Data custody and a possible sovereignty policy for artificial intelligence **V.** Conclusion

> MARIEL GARCÍA MONTES

Researcher and PhD candidate in Science, Technology, and Society, MIT

> SAIPH SAVAGE

Assistant Professor at Northeastern University at the Khoury College of Computer Sciences, where she directs the Northeastern Civic A.I. Lab.(Northeastern University & Universidad Nacional Autónoma de México (UNAM)

I. INTRODUTION

In 2002, the National Institute of Standards and Technology (NIST) of the United States used the photographs of 120,000 Mexicans to train its Facial Recognition Vendor Test (FRVT). These photographs of individuals who had requested an entry visa to the United States and went to a consular appointment to be photographed. When the algorithm was developed, the Department of State gave NIST access to a database that contained the photographs of over six million Mexicans, and NIST took a sample of individuals who had been photographed twice[1].

Two and a half decades after the start of these periodical evaluation exercises, NIST has consolidated as an authority worldwide in the evaluation of facial recognition and analysis. The importance of these evaluations is such that different stakeholders in the computational verification of identity treat NIST evaluation results as a quality stamp, a guarantee of the good

functioning of any given facial recognition system. One of the companies that benefits from their high position in NIST's annual ranking is Incode Technologies. This biometric verification company was founded by Ricardo Amper, and it is regarded as a success story for Mexican entrepreneurship in Silicon Valley. Incode Technologies, which has contracted with different public and private organizations in Mexico, states in their marketing materials that the enterprise is ranked in the first place in their category in NIST's evaluation[2].

1 NIST. Summary of NIST Standards for Biometric Accuracy, Tamper Resistance, and Interoperability. 13 nov 2002. https://www.nist.gov/system/files/documents/2021/10/25/nist_ appendix _pact _ nov02.pdf

2 Incode. Incode's Stellar Performance in NIST FRTE 1:N Identification Testing: A Testament to Accuracy and Leadership in Facial Recognition. 17 sept 2024. https://incode.com/blog/incodes-stellar-performance-nist-frtr-1n-identification-testing/

Facial recognition is a form of computer vision that abstracts human faces into numerical objects, enabling their analysis in big databases, and it is one of the forms of artificial intelligence that has seen more technological advancements in the last few years. It is also one of the most controversial ones: its use in predictive policing and court proceedings has been problematized by racial bias and other discriminatory practices[3] that emerge from the different layers of the artificial intelligence process. These layers include the creation of the data corpus for algorithmic training, the construction and structuring of the databases, the development of machine learning algorithms, and their application in different spheres of public life.

One of the applications of computerized facial recognition and analysis in Mexico is public safety. In April 2019, after an all-expenses-paid trip to China to tour Dahua Technology's factories, the then-governor of the state of Coahuila, Miguel Ángel Riquelme, announced the purchase of 1,281 video surveillance cameras enabled with facial recognition capabilities, to be used in nine different cities. The purchase was made through a direct award, and the global CEO of Dahua Technology, Zhijie Li, attended the press conference for the announcement in Saltillo.4 Since then, other cities have made similar purchases, such as Ciudad Juárez in 2022[4]. Different civil society organizations and journalists have denounced the lack of transparency over the deployment of the facial recognition technology, particularly its effectiveness, and the data used for and captured by its operation[5].

Dahua Technology is not the only international vendor of artificial intelligence for public safety in Mexico. Another notable case is Thales Group, the French information technology company that has operated the Ciudad Segura, "Safe City" project since 2009. This project built cameras, and, more broadly, an infrastructure network for the Command and Control Center, known as C5, in Mexico City[6] and in other cities across the country. In the words of

3 Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". Conference on Fairness, Accountability and Transparency. PMLR, 2018.

4 Elizabeth Rosales, "Ciudad Juárez compra equipos de videovigilancia a firmas chinas vetadas por EE.UU", *Empower LLC*, 7 june 2023 https://empowerllc.net/2023/06/07/chihuahua-coahuila-videovigilancia

5 Federico Pérez Villoro y Paloma Robles, "Vigilancia biométrica: el tortuoso camino de Coahuila hacia el reconocimiento facial", *Quinto Elemento Lab*, 11 de noviembre de 2020, https://quintoelab.org/project/vigilancia-biometrica-reconocimiento-facial-coahuila.

6 Madeleine Watterberger, "C5, la millonaria promesa de videovigilancia de CDMX que no

Jean Marie Letort, then vice-president of the cybersecurity advisory board of Thales, their product enables "not just the surveillance and footprint analysis of a criminal, but the prevention of crime through video surveillance systems that use artificial intelligence to prevent a delinquent's culmination of a crime"[7]. To date, Thales Group and the Command and Control Centers have not made any documents public to allow for independent evaluations of the algorithms, which would analyze the video captured by their cameras. There is no transparency around the origin of the data used as a corpus for algorithmic training tasks. We do not know, for instance, if there are automated decision-making processes, or if human verification is required at any step; nor do we know if there are any oversight or review resources for cases where computer analysis gives incorrect results.

Beyond the origin of data for artificial intelligence, the creation of algorithms and applications, and the import of algorithms made in other countries, Mexico, in the present day, is one of the countries that host big data centers where data are stored, processed, and distributed. These centers are a key site for artificial intelligence. In a way, one could say that the cloud is, among other places, in Mexico. In May 2024, Microsoft announced the launch of their first hyperscale data center in the country, with a financial projection of earnings over 70 billion dollars[8]. Data centers require an ample amount of material resources, especially electricity, as well as water to cool their infrastructure. Microsoft built this new data center in Colón, a district in the state of Querétaro that, at the start of 2024, had experienced exceptional and extreme drought[9].

These prior examples show that Mexico has served as a space of creation for the original data behind facial recognition technology, a role that it continues play through its contracts with international companies and local governments. The country has also been the site of implementation for different computer

previene ni resuelve delitos". *Animal Político, 14 February 2021*, https://www.animalpolitico.com/2021/02/c5-millonaria-promesa-videovigilancia-cdmx-no-resuelve-delitos.

7 Eréndira Reyes, "Cámaras con IA de la CDMX ya podrán evitar delitos", Expansión, 5 June 2019, https://expansion.mx/tecnologia/2019/06/05/camaras-con-ia-de-la-cdmx-ya-podran-evitar-delitos.

8 Microsoft, "Microsoft launches its first hyper-scale cloud datacenter region in Mexico", *Microsoft, 7 May 2024*, https://news.microsoft.com/es-xl/microsoft-launches-its-first-hyper-scale-cloud-datacenter-region-in-mexico/

9 Viviana Estrella, "Querétaro, con 14 municipios en sequía extrema y excepcional", *El Economista*, 1 January, 2024 https://www.eleconomista.com.mx/estados/Queretaro-con-14-municipios-en-sequia-extrema-y-excepcion al-20240101-0050.html.

vision algorithms; both made in Mexico and abroad. Finally, the Mexican territory has also provided material sustenance, in terms of energy and natural resources, to the infrastructures needed for data processing and other artificial intelligence tasks by transnational companies.

In other words, these prior examples show the different ways and dimensions in which artificial intelligence has already made an impact in Mexico, as well as the ways that Mexico has made an impact on artificial intelligence worldwide. From these examples, questions emerge to guide national discussions on facial recognition and other forms of artificial intelligence away from potential problems, and into the actual, already existing challenges. We pose these questions as a guide to the discussion of regulatory frameworks and general guidelines for implementation:

1. What are the institutional configurations that allowed for the use of Mexican citizens' personal photographs in the United States for the early training of facial recognition algorithms in 2002, and what safeguards should Mexico build to protect the biometric data of its citizens, both in Mexico and abroad?
2. How can we extend the protections we lacked in 2002 to the present day uses of biometric data in Mexico, especially amidst the rise of biometric technology contracts between local governments and companies abroad? What safeguards should there exist when Mexican governments enter commercial agreements with companies originally based in countries such as China, the United States, France, and Israel?
3. Regarding the use of artificial intelligence for public security, which aspects of algorithmic transparency should never remain in the opacity of alleged commercial secrets or national security? What regulations and transparency provisions are necessary so that audits on algorithmic bias and alignment can also be undertaken in Mexico?
4. Regarding the data centers in Mexico, which forms of environmental impact analysis and water management should be part of accountability requirements to construct data centers in Mexico? How can we ensure that the communities at the sites where these centers are built receive the benefits, not the harms, of artificial intelligence?

In this chapter, we analyze some of the needs that arise from the current state of artificial intelligence in Mexico, as well as some of the responses that have arisen in computer science and critical data studies research.

II. ALGORITHMIC HARMS AND REGULATION BASED ON HUMAN RIGHTS FRAMEWORKS

CASE STUDY: FANID (FEDERACIÓN MEXICANA DE FÚTBOL E INCODE TECHNOLOGIES)

The social studies of artificial intelligence have generated analysis and evidence on the social costs and harms perpetrated by the use of algorithms in everyday decision-making. For example, in 2016, news outlet ProPublica published an investigation into COMPAS, a predictive algorithm of recividism widely used in the United States justice system, which aims to assist judges' decisions by providing scores on individuals' likelihood to commit a crime. ProPublica analyzed the cases of 10,000 individuals accused of crime, and it compared the recividism rates predicted by COMPAS against the real recividism rates two years later. The journalists concluded that the algorithm led to twice as many errors among people of color as it did among white people, among other algorithmic racial biases[10]. This algorithmic bias was reflected in court decisions on bail and release dates. ProPublica's research was one of the first to demonstrate the real harms of predictive algorithms in legal decision-making, and it inspired the analysis of harms caused by the use of algorithms in other industries.

In Mexico, despite the existence of a data science and predictive algorithms industry, we do not have evidence of the effectiveness of the algorithms in use yet, and we have even less evidence of their social impact. These are algorithms designed and employed in private contexts that have no algorithmic transparency requirements. On the face of this commercial opacity, social research has focused on the side of the offer: the private industry that sells data science and artificial intelligence as a service, generally through consultancies or through software development. Their main clients are in industries such as banking, telematics and transport protection, and commerce. There have also been Mexican developments in artificial intelligence applications for security, workplace and human resources management, and identity verification. These services have focused on specific applications of computer vision, such as facial recognition and biometric data analysis.

10 Jeff Larson *et al.*, "How We Analyzed the COMPAS Recidivism Algorithm", *ProPublica*, 23 may 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

One of the most paradigmatic cases of artificial intelligence uses for public security in Mexico is FanID, a project for the biometric verification of soccer spectators in stadiums. FanID arose as a technological solution in response to an episode of extreme violence in a soccer match in Querétaro. Facing criticism from soccer fans due to the security crisis that led to a tragedy during a match on March 5th, 2022, the Mexican Soccer Federation announced an agreement with Incode Technologies to use artificial intelligence and biometric verification to increase security in different stadiums across the country[11] The Security and Citizen Protection Secretary announced a series of meetings with the Mexican Soccer Federation where the design of FanID was discussed. However, they refused to release meeting documents when a citizen requested them via a transparency request, Mexico's equivalent to a Freedom of Information Act request[12].

The social and technical promise that FanID makes is the elimination of anonymity in soccer stadiums, giving security teams a tool to identify and vet any person who engages in delinquent behavior. In practice, FanID was built as a digital platform that includes a mobile application where users upload their photograph, accompanied by their government-issued ID, and by other personal data such as their phone number and an email address. In the platform, this application integrates with the database that holds all of these personal identifiable data. Incode's facial recognition algorithms would allow security teams to identify offending individuals through photographic comparison, and to flag them in the system to ban their entry to future matches or stadium events.

With the implementation of FanID, different questions arise regarding the harms that such an ambitious biometric project could cause. In December 2022, different stadiums across the country announced FanID registration as an entry requirement for certain soccer matches. On the one hand, to ensure its functioning, FanID would need the universal participation of soccer match spectators; however, is it reasonable to condition a person's entry to owning a smartphone with adequate specifications for the FanID platform? The idea is that this would become compulsory for every spectator over 18.

11 Incode Technologies. "How Fan ID Transformed Mexican League Football". Incode Technologies, 2 march 2023 https://incode.com/blog/how-fan-id-transformed-mexican-league-football/.

12 INAI. "Stenographic Version of the Ordinary Plenary Session on the Federal Law of Transparency and Access to Public Information (INAI)". INAI, September 6, 2023.https://home.inai.org.mx/wp-content/documentos/AudiosyVersionesEstenograficas/06-09-2023SesionPl_eno_VE.pdf.

Another question that arises in an all-ages space is the management of the personal identifiable data of spectators under 18. Is parental consent enough to give a minor's biometric data, especially in early childhood, when they attend a soccer match with their parents? What effects could they experience as a result of their early inclusion in FanID, and how could it impact their right to privacy in the future? For the time being, in November 2024, the FanID website says that underage verification has not yet begun, but it will in the future; it currently is under review at the Instituto Nacional de Acceso a la Información y Protección de Datos Personales (INAI), Mexico's data privacy watchdog.

Given multiple major leaks of personal data in Mexico in recent years, a platform such as FanID raises questions about the harms of potential security breaches. Soccer fans are sharing their photographs, scans of their government-issued identification cards, and their phone numbers. How should the Mexican Soccer Federation and Incode Technologies protect user data? Which protocols must exist to mitigate the possible leaks of information, as well as the individual violations that could arise in a stadium? Further, if FanID compiles personal and sensitive data such as biometrics, could they eventually be used to cross-reference with other databases? In 2023, the INAI plenary discussed a transparency request of someone who wanted to know if FanID was cross-referencing with Plataforma México, a data platform for public security that uses computer vision to find information, for instance, on individuals with criminal backgrounds. This request points to well-founded concerns over the cross-referencing of data enabled by FanID, highlighting the need for data protection guidelines for the private use of biometrics in a wider context of data for public security.

Some of these questions were reflected in INAI's final documents on FanID. After the Mexican Soccer Federation announced an agreement with Incode Technologies, also involving the Security and Citizen Protection Secretary, INAI asked to review an impact evaluation that had been undertaken by Incode Technologies and the Mexican Soccer Federation in June 2022. Even if INAI primarily works with public, not private, entities, the scale of this project across the country made it necessary to analyze the impact FanID would have on personal data management at the national level, all in the name of public security.

In August 2022, INAI announced "a normative-operative analysis regarding the treatment of personal data in the application of Fan ID, with the objective of upholding the principles, duties, and obligations in the Federal Law for the

Protection of Personal Data under the Posession of Private Entities[13]. One of its recommendations was to perform an analysis before building the platform in order to confirm the inexistence of "less invasive mechanisms to promote safety in stadiums" given the national and international antecedents of harms from large scale biometric management. Later, in November 2022, INAI warned that FanID had "important gaps regarding the safeguard of the personal data that would be collected"[14].

Considering that stadiums would eventually announce the deployment of FanID in December 2022, one has to wonder how much importance was really given to INAI's technical analysis of the project. On the one hand, the early involvement of a biometrics-based identity verification company had the most significant consequence. It meant that INAI's main recommendation, the genuine search for less invasive mechanisms to support security tasks in soccer stadiums, would hardly be taken into consideration. On the other hand, with Mexico's current data protection framework, private entities are not required by law to undertake any privacy impact evaluations, nor to follow INAI's recommendations on any evaluations that do take place.

Under other legal frameworks, or in a context with different institutional configurations, the FanID platform would never have been able go forward in the way it was designed and implemented. There would also be public documentation on the normative-operative analysis undertaken by INAI; in Mexico, the Institute refused the transparency requests for these documents, alleging that their analysis contained industrial secrets[15]. In other words, the current personal data protection framework does not protect citizens from the mass collection of biometric data when private entities are involved. Due to industrial secrets and the Industrial Property Law, INAI does not require the publication of the documents for further privacy impact analysis, or the audit of artificial intelligence systems at play. It is the perfect legal and institutional configuration to promote algorithmic opacity.

13 INAI. "Inai emite opinión técnica sobre el documento que presentaron la FMF y la LigaMX ante uso de sistema FanID en estadios". *INAI*, AUGUST 2022 https://home.inai.org.mx/wp-content/documentos/SalaDePrensa/Comunicados/Comunicado%20INAI-248-22.pdf

14 INAI. "INAI concluye evaluación de Fan ID: FMF reforzará medidas para proteger datos personales y privacidad de aficionadas y aficionados". INAI, 6 november 2022. https://home.inai.org.mx/wp-content/documentos/SalaDePrensa/Comunicados/Comunicado%20INAI-341-22.pdf

15 Red en Defensa de los Derechos Digitales. "INAI y LigaMX ocultan funcionamiento del FanID". [INAI and LigaMX hide the FanID operations]. https://r3d.mx/2023/01/10/inai-y-ligamx-ocultan-funcionamiento-del-fan-id/

Cases such as FanID are, therefore, at the center of civil society discussions on artificial intelligence in Mexico and other countries in Latin America. This case exemplifies the way that technological advancements reach contexts with frameworks and values coded in times that had very different technological possibilities from the ones of today. With the drive to seize new possibilities to solve old problems, product design is rushed, leaving no room to examine the sociotechnical legacy that prior solution and governance efforts left behind. Computer vision technologies for biometric identification land in a context that grants more protections to industrial property and private interests than to the individual and collective right to privacy.

This is why different professional associations, civil society organizations, and researchers on responsible artificial intelligence call for the development of an AI design and regulatory framework that centers human rights.

> "Any attempts to regulate technology must focus on the protection of human rights as the main and indisputable objective, with consideration for the current social reality in the country where the regulation will have an impact, and at the same time, and equally importantly, consider the mechanisms required to promote innovation and the development of technical competencies in the country".

Hidvegi and Leufer propose to center artificial intelligence discussions on human rights, reinforcing the basic rights of those who would be most harmed by the applications of artificial intelligence[16]. In other words, every application of artificial intelligence must examine its impact on the basic rights of all people, and then society must abandon the projects that sacrifice the rights of some for the benefits of others. It is the sociotechnical application of a legal maxim: human rights are non-negotiable.

Hidvegi and Leufer, as well as civil society organizatons, propose the centering of human rights in artificial intelligence discussions as a counterpart to the mode of analysis that sacrifices the basic rights of minoritized groups for the sake of the protection of broader groups. The utilitarian spirit underlying this type of decision-making has led to technologies and practices that society would hardly accept if their impact on the rights of complete populations were discussed.

16 Fanny Hidvegi y Daniel Leufer, "The EU should regulate AI on the basis of rights, not risks". EDRI, 24 de marzo de 2021, https://edri.org/our-work/eu-should-regulate-ai-on-the-basis-of-rights-not-risks/.

Unfortunately, AI regulation in Mexico has followed guidelines that are incompatible with a human-rights-centric approach. Using the regulatory framework of the European Union as a reference has enabled the adoption of AI applications that align with principles of public and national security, despite the violations to human rights they might perpetrate. A recent study of the regulatory roads for AI in Latin America concludes that regulation in Mexico "is centered on risks that have not arrived yet, and it overlooks specific dangers of biometric mass surveillance under the premise of 'national security'"[17]. The FanID platform is an example of the mass biometric surveillance this report denounces. The European framework allows for the use of similar platforms for migration control in the region, and particularly in refugee camps.

> Risk-based models are not sufficient to protect people from the harms generated by artificial intelligence. A model centered on human rights will help delineate each of the obligations for human rights that must be met at each step in the artificial intelligence lifecycle[18].

What elements should a human-rights-centric regulatory framework for AI include? Access Now proposes the adoption of transparency principles, efficient oversight instruments that include rights impact studies, localized analysis, and the upholding of the law (and the ability to sanction). The FanID case only partially meets only one of these recommendations: on paper, they did undertake a privacy impacts evaluation.

III. BIAS AND THE REPRESENTATION OF MINORITIZED GROUPS IN ARTIFICIAL INTELLIGENCE

CASE STUDY: PUBLIC SECURITY CAMERAS ENABLED WITH FACIAL RECOGNITION IN COAHUILA AND CIUDAD JUÁREZ

In 2017, Joy Buolamwini, computer scientist at the Massachusetts Institute of Technology, began to study the workings behind facial recognition algorithms.

17 Consorcio Al Sur, df"Los caminos regulatorios para la IA en América Latina". 2024 https://www.alsur.lat/sites/default/files/2024-09/ALSUR%20-%20IA%20en%20Latam%20%5BESP%5D.pdf

18 *Ibid*.

Buolamwini, as a Black researcher, soon realized that the systems constantly failed when she tried to use algorithms to recognize her own face[19]. To render her face legible to the algorithms and the cameras in her tests, sometimes she would have to wear a white mask. After this remarkable finding, she has since studied the ways artificial intelligence excludes and perpetrates harms against minoritized groups such as Black people in the United States.

Buolamwini's research catalyzed a wave of audits and investigations into the ways artificial intelligence incurs in racial discrimination. Tarleton Gillespie describes these consequences as "representational harms" of artificial intelligence, which articulates the harms perpetrated by systems when they include few or bad data of a population or minoritized group[20]. In other words: the facial recognition algorithms did not read Buolamwini's face because they had not been trained on Black faces. The resulting discussion from this line of research is known more popularly as artificial intelligence bias.

In Mexico, we do not yet have ample research projects that study, demonstrate, or elucidate our understanding on the ways that algorithmic bias plays out in the country. However, some AI developments that have taken place in the last few years leave us with questions that point in this direction. We are dealing with the same technologies and algorithms that did not recognize Buolamwini's face in her testing, and now their use is widespread in Mexico in the context of public security.

In April 2019, after an all-expenses-paid trip to China to tour Dahua Technology's factories, the then-governor of the state of Coahuila, Miguel Ángel Riquelme, announced the purchase of 1,281 video surveillance cameras enabled with facial recognition capabilities, to be used in nine different cities[21]. Other cities in the country have followed the same trajectory. For example, also in 2019, the Mexico City government announced its plans to renew and augment 18,500 video surveillance cameras in the city to opt for cameras with

19 Joy Buolamwini y Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification. " *Conference on Fairness, Accountability and Transparency*. PMLR, 2018.

20 Tarleton Gillespie, "Generative AI and the Politics of Visibility", Big Data & Society 11, no. 2 (2024): 20539517241252131, https://doi.org/10.1177/20539517241252131

21 Segundo informe de gobierno, Miguel Ángel Riquelme Solís, Gobernador Constitucional del Estado de Coahuila de Zaragoza, 2017-2023" [Second mandate report, Miguel Ángel Riquelme Solís, Constitutional Governor of the State of Coahuila of Zaragoza], Gobierno de Coahuila, accessed 5 February 2023, https://coahuila.gob.mx/archivos/pdf/micrositio/Documentos%20del%20Segundo%20Informe%20de%20Gobierno/2o-Informe.pdf

high resolution, 360° vision, free wifi, increased storage (from 7 to 30 days), as well as facial and car plate recognition[22]. Cities in Chihuahua and Jalisco also made similar acquisitions.

The arrival in Mexico of facial recognition technology made in China shows parallels with the entry of the same products into the African continent. Also in 2019, Zimbabwe entered a commercial agreement with Chinese companies, such as CloudWalk and Hikvision, to become the first country to introduce Chinese cameras into Africa[23]. This program turned out to be a pilot to expand the technical capabilities of these companies so that their products could recognize African faces; they donated equipment to Zimbabwe, and Zimbabwe, in turn, gave back its citizens' data[24]. In the case of Coahuila, two circumstances obscured our knowledge on whether a similar agreement to the one in Zimbabwe happened in Mexico: the government's decision to reserve all contracts and acquisition documents, and the industrial secret opacity that protects Dahua Technology. There are no public documents that indicate how Mexicans' faces are captured, stored, or transferred, nor the uses that Dahua Technology may make of them when they are under the company's custody.

In the context of institutional opacity, one of the confounding aspects about the workings of these technologies relates to the data sources for the training of their underlying algorithms, as well as for cross-referencing. What is the corpus they use to work in the Mexican market, and what is the data provenance at play? We know that, under Mexican law, the country's largest biometrics database, which is the Federal Voter Registry, cannot be used for algorithmic training purposes, nor for cross-referencing with public security cameras. However, there are other databases that may be in use. Do Coahuila's security cameras cross reference with, for instance, Plataforma México? Or do they cross reference with the Ministry of Foreign Affairs' passport and visa databases?

Beyond data provenance and algorithmic training, there is no clarity on how the information that results from these artificial intelligence deployments will

22 Gobierno de la Ciudad de México. "Visión 360°" [360° vision], Centro de Comando, Control, Cómputo, Comunicaciones y Contacto Ciudadano de la Ciudad de México, June 3, 2019, https://c5.cdmx.gob.mx/storage/app/media/Boletin/Vision-360.pdf.

23 The Herald, "Chinese tech revolution comes to Zim". The Herald, 9 october 2019. https://www.herald.co.zw/chinese-tech-revolution-comes-to-zim/

24 Jili Bulelani, "The Rise of Chinese Surveillance Technology in Africa". Electronic Privacy Information Center. 31 may 2022. https://epic.org/the-rise-of-chinese-surveillance-technology-in-africa/

be used. Are there human validation steps each time the technology gives a match or flags an individual? Will there be oversight requirements for erroneous identifications? If these technologies are integrated into public security systems, will their data also be used in court proceedings?

Generally, this deployment of AI-based surveillance cameras in public spaces emerges in parallel to other uses of AI in public life. For instance, in the past decade, private companies

have sold predictive algorithms to industries such as banking and retail in Mexico. These uses of artificial intelligence in public life raise questions: What data biases exist in the AI applications already in use today? What would we need in order to redress the most harmful social biases in their operations? Who is most disproportionately harmed by these practices?

Algorithmic biases have caused the need for initiatives that augment the representation of minoritized groups, contents, and values in artificial intelligence's different layers. Some efforts have promoted this representation through the design of tools that incorporate underrepresented data. These efforts require, in some instances, the creation of training data for machine learning. For instance, the Data Against Feminicide project at the Data + Feminism Lab at the Massachusetts Institute of Technology had to generate its own datasets in order to develop a news detection tool for gender-based violence, given its importance for regional efforts for building national feminicide registries[25].

We can find more examples of these kinds of initiatives among the projects supported by Red de Investigación Feminista de Inteligencia Artificial, Feminist Artificial Intelligence Research Network, which seeks to identify and correct bias in artificial intelligence and other digital spaces, especially regarding gender bias. In Argentina, AymurAI, an initiative by the civil society organization Data Género and Criminal Court 10, supports data collection on gender-based violence in court sentences. The use of artificial intelligence in this context allows for faster data collection for gender equality research as, without it, data would have to be manually consulted, sentence by sentence, resulting in high costs of time, labor, and emotional impact, as is typical of in gender-based violence data creation work[26]. Data on gender-based violence in court systems have been central to discussions on topics such as feminicide.

25 Catherine D'Ignazio, I. Cruxên, H. Suárez Val, A. Martínez Cuba, M. García-Montes, S. Fumega y W So. "Feminicide and counterdata production: Activist efforts to monitor and challenge gender-related violence". Patterns 3(7), 2022.

26 *Ibidem*.

Another example of artificial intelligence tools against gender bias, also from Argentina, is E.D.I.A. This project is a collaboration between civil society organization Vía Libre and two computer science professors who developed a natural language processing tool to find gender bias in text. In their pilot, they tested the tool with 500 middle school workers in Argentina to incorporate their tool into their classrooms, using AI to bring evidence on the material ways that gender bias takes place in everyday life. Finally, it is worth highlighting the work that creates artificial intelligence that serves workers and their rights. In the current gig and platform economy, where independent workers register to provide different services, how platforms are designed makes it impossible for workers to communicate amongst themselves[27]. La Independiente is another project supported by Red de Investigación Feminista that seeks to use artificial intelligence to promote communication and workers' support networks among that may be hindered by the platforms where they work[28].

These artificial intelligence projects, as well as others supported by Red de Investigación Feminista, exemplify the way that we can develop artificial intelligence that incorporates and centers groups historically marginalized in computational development. To reduce gender bias, we must work with women and individuals across gender diversity, and find applications of technology that can support the road to equality.

IV. DATA CUSTODY AND A POSSIBLE SOVEREIGNTY POLICY FOR ARTIFICIAL INTELLIGENCE

CASE STUDY: IMMIGRATION MANAGEMENT TECHNOLOGIES IN LATIN AMERICA

At the start of this chapter, we highlighted the case of the consular database of Mexican faces that served as a training corpus for facial recognition algorithms.

27 Mary Gray, Mary y Siddharth Suri. Ghost work: How to stop Silicon Valley from building a new global underclass. Eamon Dolan Books, 2019.

28 Maya De Los Santos, Norma Elva Chávez, Alberto Navarrete, Cristina Martínez Pinto, Luz Elena González, Tatiana Telles-Calderon, and Saiph Savage. 2023. La Independiente: Designing Ubiquitou Systems for Latin American and Caribbean Women Crowdworkers. In Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct), October 08--12, 2023, Cancún, Quintana Roo, Mexico.

In 2002, there were no counterweights or safeguards that could impede the United States Department of State from giving NIST access to a database with the data of over six million Mexicans. At the same time, it is difficult to imagine that individuals who took their photo back then to request an entry visa would imagine that their facial data would be used to train technology for the computerized detection of faces, especially long before facial recognition became a prominent topic in public discourse.

This same dynamic, personal data reappropriation with objectives that differ from the initial ones, can be seen in Incode Technologies and the Mexican Soccer Federation's Fan ID. It is hard to imagine that someone who wants to attend a soccer match is also interested in consenting to the use of their personal data for the automated analyses of their face, especially for purposes other than soccer spectatorship. Today's regulatory framework does not allow such soccer spectators to know what are the analyses that Incode Technologies and the contracting Federation agreed upon for the FanID program.

Finally, this dynamic becomes evident as well as imported public security cameras enabled with facial recognition get deployed in public space. International experiences show evidence of the data extraction enabled by these commercial agreements, and it is hard to imagine, in the context of contracting opacity, that local governments in Mexico would have advocated for the privacy rights of the people to be recorded by these cameras. How can we guarantee Mexicans' right to privacy when their personal data cross borders, given away by local administrations?

The harms of this trajectory of extraction become even more evident with technologies that use Mexican data to then turn them against the interests of Mexicans. A paradigmatic case of this relationship is immigrations management technology, including a suite of mobile applications launched in the past decade, during the COVID-19 pandemic. As part of the immigrant applications across the region, Mexican migrants, as well as migrants from from other countries in Latin America, are expected to give their biometric data. Despite the fact that these applications are managed in different countreis, migrants sometimes use them while they are in Mexico, and many of the users are Mexican themselves. Given the impact of these technologies on the rights of migrants, particularly their rights to privacy and personal data protection, international organizations have analyzed the use of technology in the immigration context. Amnesty International questions the technological promise made by immigration authorities, which guarantee individuals that it would be easier to request asylum through these applications, showing that the reality was more complicated.

In a study with hundreds of migrants and their accompanying organizations, Amnesty International found that there were significant challenges to use the application[29].

Different research groups have studied the technical characteristics of these applications, documenting the technological decisions with the highest impact on the privacy of migrants.

Some examples are the requirement to activate a mobile phone's geolocation sharing, and the obligation to take user photographs through the application. Another controversial technological decision has been the 'liveness' test requirements, as migrant rights organizations denounce that migrants from Haiti and Africa have found that the application cannot recognize individuals with darker skin tones. The former UN Special Rapporteur on the Contemporary Forms of Racism, Racial Discrimination, Xenophobia, and Related Forms of Intolerance, E. Tendayi Achiume, expressed her concern for the potential to create problems around privacy and non-discrimination[30].

She proposes the understanding of these technologies as a mechanism of digital racial borders, used to impede the movement and migration of individuals based on their race[31]. The discrimination against Black faces perpetrated by the application is the most evident example of the ways this technology works for exclusion; however, the data that gets collected may cause different types of harms especially in the emergency context that migrants in Latin America inhabit, and it becomes urgent to evaluate the impact of such technologies.

> Although border digitization may bear the veneer of a more humane approach to traditional border technology, a study of the southern border found that the introduction of"smart" borders resulted in a more than doubling of border deaths. Those dying are almost exclusively Black and Brown migrants and refugees[32].

29 Amnistía Internacional. "CBP One: ¿Una bendición o una trampa?" Amnistía Internacional, 8 de mayo de 2024. https://www.amnesty.org/es/documents/amr51/7985/2024/es/

30 Naciones Unidas. "Discriminación racial y xenófoba y uso de tecnologías digitales en el control de fronteras e inmigración. Informe de la Relatora Especial sobre las formas contemporáneas de racismo, discriminación racial, xenofobia y formas conexas de intolerancia, E. Tendayi Achiume". A/HRC/48/76, 17 December 2021. https://documents.un.org/doc/undoc/gen/g21/379/64/pdf/g2137964.pdf

31 E. Tendayi Achiume. "Digital Racial Borders". AJIL Unbound. 115:333-338. doi:10.1017/aju.2021.52

32 Ibid.

Similarly to all the cases mentioned in this chapter, the documentation on the technical development of the platform, its artificial intelligence algorithms, and its impact assessments are not available to the public. Particularly, in the cases where a foreign government collects.

Mexicans and migrants' personal data while they are in Mexican territory, data custody questions arise. What regulatory framework would have to exist to protect the right to privacy of Mexicans on the face of these uses of artificial intelligence? What measures would be needed to guarantee that data collected in Mexico will not be used against Mexicans by other governments or enterprises abroad? How can we extend the same guarantees to migrants and refugees from other countries crossing Mexican territory?

These questions point to the need of a data sovereignty policy in Mexico. Data sovereignty becomes indispensable in a context where the value of analog data changes with the advent of computational analysis. The photographs that individuals gave on their visa applications in the 1990s changed value when their use for facial recognition algorithmic training became possible. The photocopies and back up photographs of government-issued identification cards, such as Mexico's voter card, changed value when private companies began to store them in soccer spectator databases. The selfie taken with a mobile phone changed value when it was subjected to a liveness test, intentional or accidentally designed to exclude Black faces, on a border database.

To ideate a data sovereignty policy in the country, we must recognize the importance that data have in the development of public policy and the advancement of social welfare[33]. If we will create data for the functioning of new tools to solve the problems we face, we must accept the importance and fragility of these data. Data creation and use processes must be accompanied by clear regulation that guarantee the uses for welfare, and protect Mexicans from private and public extraction in Mexico and abroad.

> "We conceive a sovereignty policy as one defined from the standpoint of the interests of the society where it emerges, without influence from foreign interests, and which responds to the diverse priorities and needs in Mexico. Moreover, a policy anchored in a human rights perspective proactively works to reduce inequality, and for all people, no matter their condition, to see their rights protected on the face of the risks of AI-based systems, and

33 Paola Ricaurte (ed). *Hacia una política de Inteligencia Artificial centrada en los derechos humanos: Oprtunidades y Recomendaciones.* 7 de octubre de 2024, https://zenodo.org/records/13899856.

to have, if they desire, access to the benefits of these technologies for a life with dignity".

–Paola Ricaurte *et al.,* "Hacia una política de Inteligencia Artificial centrada en los derechos humanos: Oportunidades y Recomendaciones"

V. CONCLUSION

In this chapter, the authors analyzed some of the developments and applications of artificial intelligence that have already taken place in Mexico, as well as the tensions that they reveal between computational analysis and social welfare. We focus primarily on facial recognition technologies, a type of computer vision that is proliferating rapidly in Mexico for the purposes of identification, public and private security, and immigration control. Our analysis explains some of the technical requirements that affect the rights of Mexicans, as well as the current challenges for impact evaluations and algorithmic responsibility audits. We particularly describe the difficulties for the research and auditing of artificial intelligence in the context of opacity enabled by the configurations of public and private stakeholders, both domestic and foreign, in Mexico. We show how these decisions bring about algorithmic harms not just in potential, but in ways that are already actualized in Mexican public life.

Based on this analysis, the authors see the following opportunities for discussing regulatory frameworks for artificial intelligence in Mexico:

1. To prevent and redress the algorithmic harms of present and future artificial intelligence applications in Mexico, the regulatory framework and general implementation guidelines must center human rights. Before every development or application, we must ask: Who would be most disproportionately affected by this endeavor? How can we guarantee that the system will not violate anyone's basic rights?
2. To avoid the algorithmic harms that arise specifically from bias, Mexico must invest in developments by minoritized groups and populations who have been absent from mainstream discussions on artificial intelligence. Feminist data science is an example of the conceptual and practical frameworks that can guide such efforts.
3. As the value of personal data expands in the context of artificial intelligence technologies, one of the biggest challenges to the right to privacy in Mexico is the extraction of Mexicans' personal data for profit abroad. AI regulation in Mexico must contemplate the sovereignty of Mexican data.

GENERATIVE ARTIFICIAL INTELLIGENCE AND ITS IMPACT ON CREATIVITY AND INTELLECTUAL PROPERTY LAW

CONTENT:

I. Law as a Regulatory Science **II.** Artificial Intelligence **III.** International Landscape **IV.** Landscape in Mexico **V.** Proposals

> KIYOSHI TSURU

Founding Partner and Director of TMI Law. Kiyoshi is a leading attorney specializing in intellectual property, particularly in the fields of innovation and cyberspace, domain name dispute resolution, copyright protection in the digital environment, software protection, patents, geographical indications, electronic transactions related to intellectual property rights, and arbitration.

Country Manager of BSA | The Software Alliance. Representative in Mexico for the Entertainment Software Association (ESA).

I. LAW AS A REGULATORY SCIENCE

According to Dr. José Ramón Cossío Díaz, Law is not an absolute, independent, or self- sufficient science. There is a symbiotic relationship between Law and Science, as the former needs to regulate the latter, but due to the exponential scientific evolution, the latter needs to be nourished by the former and understand that, without seeking the help of the scientific sphere, Law is in a limited state, unable to understand and resolve the controversies that arise in our complex contemporary reality[1].

For his part, Dr. Villoro Toranzo re-examines the notion of Saint Thomas Aquinas that "The Law is nothing more than a reasonable ordering, for the common good, promulgated by one who has care of the community" to explain the normative essence of Law and the contemplative function of the jurist. Specifically, this scholar refers to the work of the jurist with respect to the creation of a norm, which lies in drawing upon knowledge from other sciences (such as psychology and economics) to understand the legal reality. Once the legal reality is known, the work of the jurist is to order it through norms, for which the first thing to do is to value whether the legal data (that compose the legal reality) are just or unjust, and then discover the instrument that will regulate this reality and address the legal data for finally applying the legal instrument[2].

The argument of Villoro Toranzo and other authors has been raised again as to when the process of copyright creation should be understood as a norm, a simple exercise of contemplation of reality to understand and then regulate it in

1 Cossío Díaz, José Ramon, Law and Science. Tirant lo Blanch, México City, México, 2015, Chapter 1.

2 M. Villoro Toranzo, Introduction to the study of Law. Porrúa, México City, México, 1972.

accordance with the sanctions of the just and unjust according to a certain social, political, and temporal context, which at times makes us forget the notion that law is inherently reactive and, in this sense, that first it must exist in order to be taken as a norm to then be understood and then to try to regulate it. The above takes on much greater relevance when we think about how the cultural and social contexts that transformed us in ancient times and today have been reversed. This principle, unpredictable for the authors who have referenced it, has been reversed by the great impetus and speed in the changes brought about by technology in the ecosystem. As has been established before, we can conclude that, by visualizing the reality of law, it is clear that the second, technological and natural, comes after the first.

Technological development has had the consequence that the legal reality referred to by Cossio Díaz and Villoro Toranzo is now more difficult to understand and contemplate later as a norm. The legal reality is accelerated by technological development, which is more rapid than the ability that law has to digest the legal facts and principles and to create assumptions and consequences. This explains why with such speed, the legal systems become anachronistically obsolete or, more recently, the legal systems that will deal with a new specification and "fall short" in the face of the acceleration of technology.

Following this schematic of factual reality and law as a regulating activity, it is evident that while society was creating new social contexts, law responded with norms that intended to prevent and resolve disputes in its contemporary environment. For example, in the early stages of our civilization, the protection and preservation of life and the right to private property were paramount. With the passage of time, society required principles that guaranteed the creative freedom of the human soul and safeguarded to the greatest extent the freedom of expression, a right that is recognized as a personal and inalienable creative right to be an author, and its power to authorize or prohibit its exploitation and to enjoy the fruits of it is guaranteed.

One of the first antecedents registered in our historical records is the protection system for the recipes of the Sybaris, in ancient Greece, documented in an initial instance by Atheneus of Naucratis and later retrieved by the German historian Conrad Cichorius in 1922. This system consisted of the cooks who created a new dish, which would be a favorite among diners, receiving the exclusive right to prepare it, with the exception of the rest of the population, for a period of one year. This protection was limited to the territory of Sybaris.[3]

3 M. Witty, *Athenaeus: describes the most ancient intellectual property*, *Prometheus*, Pluto Journals, Leicester, England, 2018.

Another precedent, considered for the birth of Intellectual Property Law[4], is the Statute of Queen Anne of 1710, which replaced the "Licensing Act", and which established for the first time that the right to print a copy (*copyright*) belonged only to persons in addition to the publishers (stationers), granting rights over the printing of a specific work, for 14 years that could be renewed for another 14 years if the author was still alive, or for the duration of 21 years non-renewable for those works already in force in the Statute. Authors such as Deborah Halbert, with whom we agree, have ruled that the true intention underlying the Statute of Queen Anne of 1710, rather than constituting a legal framework for the protection of authors' rights, was really an instrument designed to allow the English crown to regulate the incipient printing industry, and thereby exercise censorship over this powerful and revolutionary technology that allowed for mass dissemination, which was dangerous for the monarchy[5]. The result is fascinating as this topic is still relevant today, in which we observe fascinating tensions between freedom of thought, neuro-rights, freedom of expression, content moderation, censorship, and that which Madison called the capture of democracy by a "factional ego"[6] which has been well portrayed by Dr Pedro Salazar when he has referred to the contemporary kakistocracy that simultaneously detains political power, economic and ideological. "The Trump-Musk duo and the gang that surrounds them are the greatest danger to our freedoms that has ever existed in the modern history of humanity. The amalgamation of powers, which they hold, with its dissociative drift, threatens the bonds of community and without defense"[7].

Regardless of the above, Halbert introduces a notion that we will return to later, in the sense that "the battles in copyright law reflect the mutually exclusive intentions of maintaining authority and control and the need to relax control to

4 This term is used in Mexico by Dr. David Rangel Medina, and it is the one I prefer instead of "intellectual property", because it reflects the true nature of this disciplinary field, whose first principle and ultimate purpose is social. Rangel Medina, David, Intellectual Law, National Autonomous University of Mexico, Mexico, 1999.

5 Probably it is not a coincidence that the English Monarchy survived while the French Monarchy did not, probably because the French Monarchy failed to impose a powerful censorship on the illustration ideas. Hence it was not until the creation of internet (the niece of the print machine) that humanity understood the real dimension of monarchies around the world, as a decorative institution, instead of being a divine institution.

6 Hamilton, Alexander, et al, The Federalist Papers. Signet Classics, an Imprint of New American Library, a Division of Penguin Group, USA, 2005.

7 Salazar, Pedro, Kakistocracia 'reloaded', column in El Financiero, January 29, 2023, available at https://www.elfinanciero.com.mx/opinion/pedro-salazar/2023/01/29/kakistocracia- reloaded.

facilitate the circulation of a work). Each person's stance in this debate depends on their economic incentive to either share or control information". In addition, he takes up the notion of Bred Sherman and Alan Stowel that "copyright is a normative body that must be 'put in context,' which means that it is currently being subjected to the historical scrutiny to which other aspects of Law have already been subjected", highlighting the importance of considering the legal reality of a given time and place in history to understand the normative system proposed by Law to regulate it[8].

Now, creations of the mind, specifically artistic works (regardless of their nature), like Law itself, have been the subject of a transformation driven by technological development and the tools and media made available to creators. Randall Parker and Ken Jordan refer to the relationship between artists and scientists, as the former, like judges, use the latter's advances to implement improvements in their processes and results: "subsequent generations of artists sought and found integrated forms and interdisciplinary strategies to express their concern for the individual and social condition and the extreme states of subjective experience. In the years since World War II, scientists have pursued personal computing and human-computer interactivity as vehicles to transform consciousness, expand memory, increase knowledge, intensify intellect and empower creativity. The idealistic and ideological aspirations of both groups have given rise to a new medium that emphasizes individual choice, free association, and personal expression". Specifically, Packer and Jordan Mention as relevant advances for the creation of multimedia to internet, the CD-ROM, the digital camera and virtual reality, understood as the multimedia as a way of concatenated art, to which artists such as Wassily Kandinsky referred as the corset of arts that, when used properly, will give rise to truly monumental art[9].

Intellectual Property Law has known how to adapt to changing standards that impose a new legal reality reformed by technology. While Copyright Law emerged in the modern Western world, without being expressly planned, from an editorial control mechanism over printing presses, it is true that it is the most noble of legal disciplines, having had an impressive capacity for evolutionary adaptation to the reality that has historically surrounded it.

In response to the technological revolutions of our civilization, law has had to react, sometimes with equanimity and other times convulsively, to the advent

8 D.J. Halbert, Intellectual Property in the Information Age, Quorum Books, Connecticut. USA, 1999.

9 R. Packer y K. Jordan, Multimedia From Wagner to Virtual Reality, Norton C Company, New York, USA, 2001.

of new technologies. It has had to find forced solutions due to the fact that, while law provides for factual situations and grants legal consequences, it is also true that many controversies are not provided for in the normative assumption, and therefore there is no applicable legal body for that particular situation. These controversies are usually resolved through jurisprudence, in those cases where judges venture to assume them as such and make an interpretive and integrating effort that allows them to solve new problems with existing legal tools.

I do not believe that it is necessary to create a new right for Artificial Intelligence (AI), in the same way that Judge Frank Easterbrook did not believe it was necessary to create a new right for the horse[10]. When he referred to the supposed right of Cyberspace[11], AI is permeating, transversally, countless aspects of our civilization as we know it today. We consider that a good part of the legal questions arising around the Generative Artificial Intelligence (GAI) can be resolved with existing legal tools. Again, judges will have to venture to assume it, and not as legal automatons. In some areas, such as the criminal or corresponding to Administrative Sanctioning Law[12], where strict normativity is needed, because it deals with strictly defined environments, where interpretation is not permitted, as in the case of crimes or administrative infringements.

Of course, in the medium and long term, legislative, public policy, and governance developments surrounding AI will be welcomed. Similarly to what was expressed by Judge Easterbrook, Dr. Cossio has used the analogy of Railroad Law[13]; when this new means of transport was created, there were no legal precedents that regulated the legal and illicit acts that arose from its adoption. This did not result in the creation of a new railroad regime. Judges simply began to resolve the new controversies that arose with this technological innovation. Over time, jurisprudential criteria were formed and specific laws were issued for that field (for example, in the labor, criminal, contractual, and regulatory fields).

10 Easterbrook, Frank H., Cyberspace and the Law of the Horse, University of Chicago Legal Forum: Vol. 1996: Artícle 7, https://chicagounbound.uchicago.edu/uclf/vol1996/iss1/7/

11 Recently in Mexico, we have seen a proliferation of legislative initiatives that define "cyberspace" as a legal term, with very different and even contradictory natures, the term comes from the novel of the "Neuro-Romancer" a science fiction trilogy from Gibson: Gibson, William, Neuromancer, Ace Science Fiction Books, New York, USA, 1984

12 See jurisprudence 2a./J. 124/2011 (10a.), NORMS OF ADMINISTRATIVE LAW. FOR THEM TO RESULT IN THE PRINCIPLES THAT GOVERN THE SANCTIONING ADMINISTRATIVE LAW, IT IS NECESSARY TO HAVE THE QUALITY OF BELONGING TO THE ADMINISTRATIVE LAW., available at https://bjs-scjn.gob.mx/doc/Tesis/4/124/2011a/10a/Tesis.

13 Cossío Diaz, José Ramón and Kiyoshi Tsuru, Round Table: Who is liable for of acts committed with artificial intelligence? Mexican Bar Association, México City, México, September 2023.

Also, I hope that someday we can reach an international consensus on a global and uniform regulation of AI. However, the global landscape does not seem encouraging. The polarization between the perspectives of the world's technological powers has its roots in the industrial competition that is predicted in the matter of AI. If there were a clear consensus on the criteria of the different technological powers, there would be a homogenization that would create a *lex mercatoria* that would allow a "field of play" under equal conditions for the most relevant agents in the AI industry, including international and national regulatory authorities.

For example, the European Union has made it clear with the law in force in various legal and regulatory systems (whose epitome is the AI Law published in 2024) that its focus on the development and use of AI must be anthropocentric, beginning as an initiation and giving respect to human rights, establishing guiding principles and highlighting possible risks and biases that can be derived from the use and/or development of Artificial Intelligence.

Other countries like China or the United States have also tried to mark their position regarding this, paradoxically coinciding that at the moment it is not necessary to over- regulate AI, to leave it free to innovate in the matter.

Mustafa Suleyman refers to this in his chapter "National pride, strategic need" in which he describes the importance that China gives to its agenda for 2030, a goal that points to the Asian giant becoming the global leader in all things related to AI[14]. The previous vision contrasts directly with the North American vision, but in essence it's the same, a launching point for innovation in the field. This new cold war was evident in the release of the Chinese model "DeepSeek", whose premiere resulted in an unheard-of increase in Nvidia's stock market value of $593 billion, which is a chip manufacturer based in Santa Clara, California[15], right before the exasperation of the American government (with all its last-minute announcements regarding its policy on Artificial Intelligence), followed by descriptions of the use of the DeepSeek model in various sectors, both public[16] and private[17], in the United States.

14 M Suleyman. "The Coming Wave" Norton C Company, Crown New York, USA, 2023.

15 https://www.Reuters.com/technology/chinas-deepseek-sets-off-ai-market-rout-2025-01-27/

16 https://www.cnbc.com/2025/01/28/us-navy-restricts-use-of-deepseek-ai-imperative-to- avoid-using.html

17 https://www.business-standard.com/technology/tech-news/deepseek-s-ai-restricted-by-hundreds-of-companies-amid-concerns-over-leak-125013100085_1.html

Let's examine an example that comparative law provides: the Supreme Court of Justice of the United States of America's ruling in the case of Burrow-Giles Lithographic Company against Napoleon Sarony in 1884. In the context of that era, the daguerreotype, as the photographic apparatus and procedure was called, had been invented and presented approximately 45 years before the sentence was issued. The daguerreotype, as a creative tool, faced all kinds of disapproval and criticism, specifically from painters, who conceived of the daguerreotype as a threat to the "true" canvas and brush. How could a mere act of pressing a button be a creative process, comparable to the act of drawing sketches or Painting oils? One of his specific critics, Paul Delaroche, a French painter from the 19th century, coined the phrase "Today, painting is dead! Photography killed it!" at the first daguerreotype. One thing is certain: photography certainly did not extinguish painting, but in some way transformed the way the world "looks, understands and explores the world"[18].

The relevance of the case for the authorial analysis of AIG is quite obvious: the core of the question is whether the product resulting from an operation carried out with AIG tools can be considered an authorial work, and the person who used said tool an author. The case has been cited and discussed in forums and classrooms around the world, both by enthusiasts and detractors of Artificial Intelligence.

The facts of the case consist of the photographer Napoleón Sarony taking a portrait of the writer Oscar Wilde, which he titled "Oscar Wilde No. 18", in which the writer appears seated, with his suit and tie, and covered with a coat, his left arm resting on his leg, supporting his chin, with his face slightly inclined. The lithographic company Burrow-Giles Lithographic reproduced Sarony's photograph without his authorization and sold it for commercial purposes, to which the latter sued the lithographic company claiming infringement of copyright. At that time, photography was not considered authorial work in US legislation.



Minister Miller decided the case, establishing that photographs with considerable

18 C. Bellinetti, "Today, Painting is Dead: Photography´s Revolutionary Effect", "Art C Object" North Carolina, USA, 2019.

works in terms of authorial legislation, since they are not limited to written texts, and especially due to the human activity that underlies the taking of the photograph in particular, the posture and clothing of the model portrayed, the lighting, the staging, the textile arrangements in the portrait, and in general, the composition of the photograph itself, elements that constitute creative decisions of the author that are translated into originality.

It is therefore evident that we are at a moment in history (both from a legal perspective and a technological-cultural perspective) in which paradigm shifts are on the horizon. As previously mentioned, the rapid pace of technological evolution, and the fear or apprehension of the legal sector about regulating something it does not yet fully understand, and which it knows it cannot delay its decision on, is anticipating the future development of these technologies and their possible scope and risks; legislators around the world have taken on the task of regulating artificial intelligence.

Clarifying the paradigm in which we find ourselves, authors like Ian Ayres and Jack Balkin[19] have attempted to rethink the regulation of artificial intelligence by making use of other existing regulations in most jurisdictions, such as the duty of agents or fiduciaries, based on the understanding that the tools, programs, or application actors of artificial intelligence cannot be considered as imputed or responsible legal actors, but rather that they are tools and users who should be responsible for the harms they cause.

For the above, two main strategies are established for the regulation of artificial intelligence: "the first strategy is that the law attributes intentions to an entity. For example, in the case of a voluntary unlawful act, the law attributes intentions to the actors. It is presumed that the actors who injure and others who tend to intend the consequences of their actions will have them. The second strategy obliges the actors to comply with a standard of behavior, normally of reasonableness, independently of their real intentions. This norm of behavior is external to the real intentions of the actors". (for example, due care). Through these two perspectives, the first, the behaviors to which the intention of the actor is not important, the law regulates and imposes a sanction according to the behaviors without importance of the actor's intention, the law catalogues the behavior based on the objective criterion (with a basis in hypothetical circumstances independent of the actor) and the law regulates and sanctions in case of not conforming to the criterion.

19 I. Ayres and J.M. Balkin, "The Law of AI is the Law of Risky Agents without Intentions", 2024.

The main point behind the work of Ayres and Balkin is to highlight how the law has faced this type of actors before, and how there are already mechanisms and concepts in the majority of legal systems that could serve to understand the phenomenon.

II. ARTIFICIAL INTELLIGENCE

The term "artificial intelligence" was coined at a 1956 conference on informatics at Dartmouth College, when a clinical approach was proposed to study this phenomenon, understood as that related to human activities that can be sufficiently specific to be simulated by a machine[20]. In principle, we must demystify Artificial Intelligence (AI) to dimension it in its proper context, clearing up its characteristics and qualities with the aim of effectively knowing the scope of this technological tool. AI has been defined as a field of science and engineering that deals with understanding, from a computer science point of view, what is commonly known as intelligent behavior. It also deals with the creation of artifacts that exhibit this behavior[21].

More than 60 years after the field of computational science conceived of artificial intelligence, this technology is radically changing our civilization to the point where some nations consider this technology as potentially risky or subject to regulation and oversight.

Currently, the various technological changes have ceased to be a means of dissemination and reproduction, since from the invention of the computer and information technology, specifically artificial intelligence, developments have focused specifically on collaboration for the creation of various materials, and even for their creation and automation, a situation that has generated various questions related to Copyright, which we will address in this document.

As we have seen, the latest technological contexts have generated vertiginous changes in the dissemination, distribution, exploitation, and even inclusion of various works.

However, technological changes and their impact on the exploitation and protection of works and materials are not new to copyright and related rights, as we have illustrated through the example of the Statute of Queen Anne of 1710 and

20 Dartmouth College. "Artificial Intelligence (AI) Coined at Dartmouth" https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth

21 R. Pino Diez, R. Gómez Gómez, N. Abajo Martínez, Introduction to AI: expert systems, artificial neuronal links and comparative evolution, University of Oviedo, Publications Services, Oviedo Spain, 2001.

the verdict of the Burrow-Giles Lithographic Company against Napoleon Sarony of 1884.

Another example of this is the search for international regulatory frameworks that would promote respect for the rights of creators, leading to the Berne Union[22] by virtue of the development and modernity in the means of communication and transport of works in the world.

On the other hand, there have been technological elements that have revolutionized the way business is conducted; an example of this is the creation of mechanical machines that reproduced music, for which authors made claims for the use of their works, a situation that was not reflected until the Berlin Act of 1908 of the Berne Convention for the Protection of Artistic and Literary Works, a situation that led to the creation of figures such as compulsory licenses[23] or rights of simple remuneration for authors, performers, and executants.

Because the functionality of AI is related to intellectual activities, the technology is potentially applicable in a variety of human activity areas. Since the 90s, there have been commercial applications of AI in areas such as production, process control, and accounting[24]; however, generative artificial intelligence gained momentum in human activities related to creativity since 2014.

Various technological and computational advances have occurred that could generate materials assimilable to works by means of computers, which will be discussed briefly and superficially in the following paragraphs.

During the 80´s the cyber-thinkers initiated a movement called connectionism, a movement of cybernetics that addressed problems from the perspective of cognitive science. It was applied to neural networks, allowing for the purification of deep learning through distributed representation techniques and retro-propagation

22 A meeting of interested countries to agree on a regulatory framework that would protect the authorial creation of original works, which culminated in the drafting of the Berne Convention of 1886, an international treaty currently signed by 181 countries.

23 Defined by the World Intellectual Property Organization (WIPO) as a "Limitation on the rights of copyright holders or related rights holders, the essence of which is that it does not depend on the holder of a particular right whether authorization to carry out acts protected by the right in question is granted or not. (...) They must be granted mandatorily by the rights holders or, alternatively, by a collective management organization or by the competent authority. It differs from a statutory license in that, in the latter, the user does not even have to request a license, as the legislation directly authorizes the use in question and only remuneration must be paid". There are legal frameworks such as the Berne Convention and the Universal Copyright Convention, which provide for compulsory licenses.

24 S. Navas Navarro, Artificial intelligence, technology and law, Tirant lo Blanch, Spain, 2017.

algorithms, a situation that has been described as a sophisticated retro-feeding loop, which has been referred to since 2006 by researcher Geoffrey Hinton as "*deep learning*"[25]. While *deep learning* was primarily used to recognize images, this technology gave rise to what Ian Goodfellow now calls generative adversarial networks[26], which allow for the generation of images using a system that uses two competing networks: a discriminator and a generator. The discriminator network constantly analyzes the images created by the generator and determines if said images are "false" or are based on information provided to the system. This allows the results to improve to the point of creating diverse images that are similar to reality even though they are fake.

The above has allowed for the development of computational creations that are good enough to pass as human-made works, even to the point of imitating photographs of landscapes or artificially creating people.

Since then, the question has been raised: do materials created by means of artificial intelligence generate copyright? If so, who is the author of these materials?

In this regard, and before delving into an analysis of the international panorama, two important details must be considered regarding copyright. On one hand, Isidro Satanowsky stated that "an intellectual work is a creation of the mind, of a complete, integral, and unitary character. It must also have a specific existence, be complete in itself, and present to the spirit a unified meaning that has nothing to do with what is external". This means it's necessary to possess an individuality and requires not only an active and decisive task, but a creative and autonomous labor, which gives form and spiritual life to an idea[27]. On the other hand, Delia Lipszyc focuses on the concept of individuality as a requirement for the existence of a work, "in matters of copyright, originality resides in the creative and individualized expressive or representative form of the work, no matter how minimal that creation and individuality may be". Some authors prefer to use the term individuality, instead of originality, considering that it more adequately expresses the condition imposed by law for a work to enjoy protection: that it has something individual and proper to its author[28].

25 B. Caldas Vianna, "Generative Art: Between the Nodes of Neuron Networks" AI, Arts C Design: Questioning Learning Machines. Number. 26.

26 I. Goodfellow, "Deep Learning", The MIT Press, Massachusetts, USA, 2016.

27 (conf. Isidro Satanowsky, "Intellectual Law", Argentina Typographic Editorial (TEA), Bs. As.: 1954, p. 154)

28 LIPSZYC, Delia. Copyright and Related Rights. UNESCO/CERLALC/ZAVALIA Buenos Aires: 1993. Pg. 65.

III. INTERNATIONAL LANDSCAPE

Beginning with the United States of America, its Copyright Office (USCO) has stated on multiple occasions that material generated with the absolute intervention of artificial intelligence lacks the essential[29] requirement to be registrable as a work protected by copyright. First, in the case[30] "*A Recent Entrance to Paradise*", a drawing created by the AI system Creativity Machine, itself created by the controversial Stephen Thaler, the USCO Review Board determined in February 2022 that the registration application did not contain evidence of sufficient intervention by a human being in the creation of the work, which is why the USCO determined it was impossible to register a work that had not been created by a human author.

Derived from the aforementioned case, the USCO stated that works protectable under its jurisdiction are those that represent the fruit of intellectual labor and are based on the creative power of the human mind. In this same response, the USCO stated that, consequently, the products of a mere mechanical process, without any intervention of a human author, are not protectable in principle.

The USCO decided, in another registration attempt from February 2023, regarding the refusal to protect certain images generated by the "Midjourney" artificial intelligence service in the case of the comic titled "*Zarya of the Dawn*", that the process used by tools like the aforementioned one does not constitute human creativity. This is because, although the commands entered by the human user into the AI system (also known as a prompt) can influence the result the AI will provide, it is not possible to know this result by virtue of human creativity. In other words, human creativity is not involved; rather, it is a deep learning system combined with an adversarial AI system that randomly produces results. (Unlike the use of tools like a photographic camera, where the author has creative control a "Creative Spark"[31] that generates a result), the use of AI eliminates any element of human creativity required for copyright protection.

The USCO, through its criteria and published guidelines, has maintained that while a work created by AI cannot be registered under the applicable legislation, the portions of a work where the author's imprint has intervened are protectable, regardless of whether these portions are combined with AI-created content to compose the work. For example, in *Zarya of the Dawn*, it was decided that while the

29 As established in "Chapter 306: The Human Authorship Requirement" provided in the "Compendium of U.S. Copyright Office Practices". Third Edition.

30 Thaler v. Perlmutter, No. 22-cv-1564 (D.C. Circ. 2023).

31 Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340 (1991)

comic's illustrations were generated by AI and therefore not protectable, other aspects such as the dialogues, the arrangement of the images, among others, are the fruit of the author's imprint, meaning those portions of the comic are protected by copyright.

In this way, the USCO establishes a gradual criterion through which it requires applicants for copyright registration to declare which parts of their work (if applicable) were created using AI tools. This ensures that elements of these works that meet the legal requirements to be considered works are not left unprotected and, consequently, are covered by copyright law.

In addition to the cases cited above, the USCO has undertaken an initiative dating back to 2023 through which it has received comments from around the world on Artificial Intelligence. As part of this initiative, the USCO set out to publish a three-part report. The first part specializes in the topic of "Digital Replicas" or deepfakes; the second part addresses the registrability of works generated by AI, and the third part will address the legal implications of training AI models with works protected by copyright[32].

In the second part of the Report, the USCO shares certain criteria, for example, that current U.S. copyright law is sufficient, adequate, and does not require any modifications to regulate works generated by (or using) Artificial Intelligence. Furthermore, it elaborates in greater depth on the grading system that had already been outlined in previous USCO decisions cited herein, specifically asserting that copyright protection necessarily requires the author's imprint. However, it also acknowledges that AI can be used as a tool (somewhat similar to what was decided regarding the use of the camera in the case Burrow-Giles Lithographic Company v. Napoleon Sarony) that interacts with a human, following the user's instructions detailed in the prompts entered to achieve the desired result.

The USCO, however, makes it very clear that the analyses of registrability must be examined and decided on a case-by-case basis. The USCO underscores the differences in their generation and distinguishes between the aforementioned works, conceptualizing them as "AI-assisted works", separating them from those works that are generated by AI without any human interaction, leaving out of the equation the imprint of a human author in their



32 https://www.copyright.gov/ai/

creation, and therefore, disqualifying them for registrability and defining them as "AI generated works".

The USCO's criterion for registrability is now even more defined with its second report, clearly establishing that "AI generated works" lack creative decisions or controls behind them, which typically constitute the creation processes of any type of work. These processes are carried out by computational algorithms that seek to respond to the prompt based on the information with which they were trained. It is evident that in this process there is no moment where the imprint of an author could be argued, so there is no creative spark, since the algorithm will assign a specific result to the received prompt, without any human intervention. To illustrate the above, the USCO used the following image in its report:

Considering this criterion, the USCO has already approved the registration of an AI-generated work, specifically the work "*A Single Piece of American Cheese*" by Invoke, a Generative Artificial Intelligence platform. Invoice noted that in its registration application to the USCO, it emphasized that the creation of its work was achieved through a multi-stage process involving iterative refinement and creative decision- making by humans. This was done to request protection for the selection, refinement, and arrangement of the material generated by Artificial Intelligence in the final product. As a note, this work had already been submitted for review by the USCO in September 2024, which resulted in a denial. This decision was reversed in January 2025, shortly after the publication of the USCO's second report.



As a conclusion to the U.S. landscape, it is worth noting that there are currently more than twenty different lawsuits whose rulings will likely finalize the criteria that courts and authorities in that country will follow regarding material generated by Artificial Intelligence.

Shifting jurisdictions and focusing on an equally important territory in the global race for technology development, China has built a legal system that follows an industrial policy aimed at protecting the economic investment of those who create new technologies.

The Chinese perspective is reflected in three different judicial decisions from courts in Shenzhen, Guangzhou, and Beijing.

In the first case, Shenzhen Tencent Co., Ltd. v. Shanghai Yinxun Technology Co., Ltd., decided in 2020 by the Civil Court of the Nanshan District in Shenzhen, the ruling favored Tencent. The court held that although the reports generated by its AI tool called *Dreamwriter* generated at the end of the stock market's trading day, Tencent's staff arranged and managed the information contained in these reports. This allowed the reports to be considered as works, and consequently, the unauthorized reproduction by the defendant was penalized.

The second case, Li v. Liu, decided by the Beijing Internet Court, established the criterion that AI-generated works resulting from specialized *prompts* (composed of constraints, prohibitions, guidelines, creative directives, among other elements) are protectable based on the imprint the author determined in said prompt. In this case, the ruling addressed whether illustrations of women produced through Stable Diffusion using *prompts* of nearly 300 words were subject to copyright protection, as the plaintiff claimed that the defendant had made unauthorized use of his works generated by *Stable Diffusion*.

In the third and final case, Shanghai Character License Administrative Co., Ltd. v. the AI tool called "Táb", the Guangzhou Internet Court decided that the operator of Táb, which provides a Generative AI model through its website, infringed upon Shanghai Character License Administrative Co.'s exclusive right to reproduce the character "Ultraman" (licensed by Tsuburaya Productions Co., the original rights holder of the character). Specifically, because when using the word "Ultraman" in the *prompt*, Táb's AI model generated an identical copy of the original character.

Furthermore, the Court found that the platform also infringed upon the exclusively licensed right to create derivative works based on the Ultraman character, as through *prompts*, any person could generate variations of the character using the AI tool.

IV. LANDSCAPE IN MEXICO

Having analyzed the international landscape, it is necessary to examine the national scenario.

In the legislative sphere, it should be mentioned that in the previous Legislature, over 60 bills were introduced in the Senate and the Chamber of Deputies to directly or indirectly regulate some aspect of AI. Some of these included provisions related to intellectual property (some even proposed reforms to the Federal

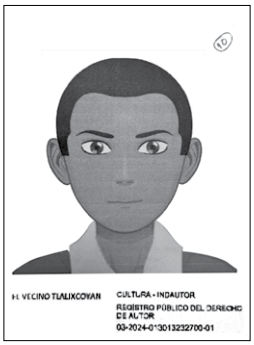
Copyright Law). None of these initiatives have been fruitful, so no modifications in this area have been implemented in our national legal framework.

It should also be emphasized that our Political Constitution establishes in its Article 28 that "In the United Mexican States, monopolies, monopolistic practices, state monopolies, tax condonations, and tax exemptions are prohibited under the terms and conditions set by law. The same treatment shall be given to prohibitions under the guise of industry protection. (…) Nor shall the privileges granted for a determined time to authors and artists for the production of their works, and those granted to inventors and improvers of any improvement for the exclusive use of their inventions, be considered monopolies. (…)". This is to highlight that the constitutional purpose of intellectual property protection is to incentivize the creation of works, a matter that is inapplicable to Artificial Intelligence, as it cannot be incentivized. If such incentives were applied to companies developing Generative Artificial Intelligence, the ultimate constitutional purpose previously mentioned would be undermined.

The Federal Court of Administrative Justice has set a precedent through the ruling of its Specialized Chamber in Intellectual Property Matters in Administrative Litigation Case Number 788/24-EPI-01-2. Through this ruling, it has established its criterion, specifying that "even if a natural person provides the guidelines and instructions to the artificial intelligence, the creation remains a product of the execution of the algorithms it employs, generating artificial content. Therefore, under the terms of the Federal Copyright Law, a work not created by a natural person cannot be registered, as it is expressly stated in the law that only such a person is capable of creating an original work (a requirement demanded by law) because human creativity is necessary for this purpose". This criterion allows our system to adapt to the gradation envisaged by the USCO and addressed earlier, since although this jurisprudence is applicable to AI- generated content, it does not specifically mention the case where a work is composed of both AI-generated material and works made by humans (authors exercising their creative activity through their imprint).

Based on the above criterion, registrations seeking protection for content generated by Artificial Intelligence have been rejected, such as the following:





V. PROPOSALS

It is necessary to establish clear criteria regarding when an author, a natural person, can be considered the author when creating a work with the assistance of a Generative Artificial Intelligence tool. I am not saying that a computer system should be considered an author. Clearly, that would be nonsense. Thaler's claims attempting to obtain copyright protection are absurd[33].

Intellectual Property Law has a primary principle, an ultimate purpose, and a social vocation. It is not an absolute right. It is not an end in itself. It is a means to achieve an objective aimed at the common good: to incentivize creators to create by granting them a temporary exploitation privilege, thereby increasing the volume of the nation's cultural heritage, but above all, to produce higher quality works by allowing the author to professionalize and dedicate all their time and attention to their work. Humans can be incentivized. Machines cannot.

It is clear that only a human being can be an author. This is the simple answer to the question. The more complicated issue is determining what degree of contribution from the user is necessary for their interaction with the AI tool to constitute the original creation of a work.

It is undeniable that authors are increasingly using (and successfully so) AI tools. To the extent that technology allows the user to have greater control over how the GAI (Generative Artificial Intelligence) generates results, and consequently, greater traceability of the author's imprint in the particular expression of their work that is, the immediate and direct relationship between the user's creative decisions and the generation of content from them.

It seems we are not yet at that point. According to the state of the technique, generating content from user promptings does not imply direct user control over the execution of the expression of the content in question. Copyright protects expression, not ideas. For now, the process of transforming a conceptual instruction into content (text, image, music, video) with the help of a GAI tool happens automatically, in accordance with the operation mechanism of Large Language Models (LLMs).

This analysis should not be taken as an immutable verdict. If technology evolves in a way that allows the user to have more direct and effective control over the execution of the creative process, so that their imprint can be clearly and identifiably impressed upon a work, and the traceability of creative decisions from

33 See Thaler v. Perlmutter, supra

conception to the creation of the expression is guaranteed, then we can move to the next level of discussion.

In an ideal scenario, criteria for originality should be uniformly standardized, for example, among all member countries of the Berne Convention or the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS).

Another alternative could be the establishment of a new sui generis regime for the protection of content created with the assistance of GAI tools. Our copyright system already incorporates, albeit very briefly, a sui generis model for the protection of databases, similar to that currently provided for databases in Chapter IV of Title IV of the Federal Copyright Law. Through this, the copyright of protectable works created by people and the protection of synthetic content generated by AI could coexist. This right would aim to protect not only the algorithms and computational mechanisms that make content generation possible through these systems but also the data used in the creation of such Artificial Intelligence models. In this way, authorship would not be granted to a set of computational algorithms, nor would the material generated by them be considered a work, but rather protecting the investment behind them, while also facilitating the identification of materials that could potentially infringe on works protected under current regulations[34].

This type of protection is not foreign to existing law. It is enough to look to our neighboring country and its jurisprudential precedents on the matter, specifically the case of Feist Publications, Inc. v. Rural Telephone Service Co., resolved by the Supreme Court of the United States. The context of the case is that Rural Telephone Service, a company providing telephone services, compiled a directory with the names and phone numbers of its users. Feist Publications copied part of the information collected and published in Rural's directories to create its own guides. The U.S. Supreme Court decided that although Rural's directories did not meet the requirements established in U.S. law to qualify as a work (and consequently, be protected by copyright), the effort behind the selection, compilation, and organization of information in Rural's directories deserved legal protection. In this ruling, the concept of "*sweat of the brow*" was outlined to conceptualize the work done by Rural and its employees that produced the telephone directories in question.

34 See the presentation of my analysis on a possible sui generis protection regime at the forum "Irruption of New Technologies: Paradigm Shifts in Copyright?" convened by the Inter-American Copyright Institute (IIDA) held on May 7, 2024, at the facilities of Ciudad Universitaria of the National Autonomous University of Mexico (UNAM), at the initiative of Maestra Carmen Arteaga.

By having such a *sui generis* protection, innovation would be promoted, investment would be protected in the Mexican jurisdiction, and an incentive would be created for creatives worldwide to use this jurisdiction as a base for their technological and cultural developments.

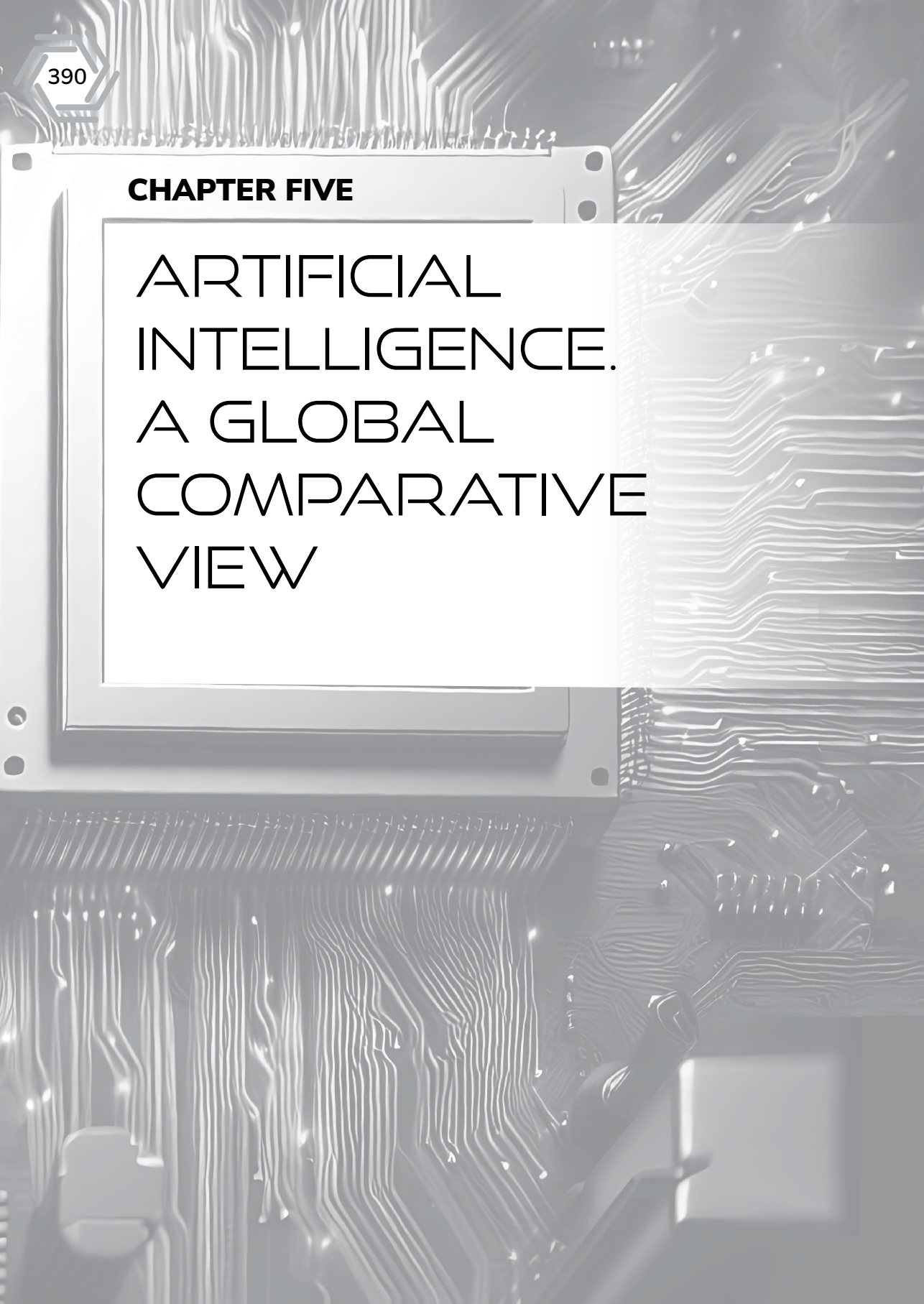
In the realm of administrative infringements provided for by our copyright legislation, specifically so-called commercial infringements, and in accordance with the principles of Administrative Sanction Law (for example, *nulla poena sine lege*), we believe that the unauthorized use or reproduction to copy or create derivative works through AI tools of works protected by copyright should be specifically added to the catalog of commercial infringements.

Lastly, but no less important, attention must be paid to text and data mining (TDM), which refers to the computational process of extracting information from public sources carried out by computational mechanisms (including *crawlers*, which automatically search for and extract text and information available on web pages). This technique, also known as scraping, aims to enrich the large datasets (datasets) used to train Artificial Intelligence models, on which the quality of the materials they generate depends. Various jurisdictions around the world are evaluating the possibility of including TDM processes in their respective copyright limitations. For example, in the United States, it is being argued (mostly by companies owning AI models) that TDM processes should fall under the U.S. doctrine of fair use. In our country, this exception could be provided for in Article 148 of the Federal Copyright Law, which outlines a *numerus clausus* list of limitations to economic rights.

Finally, we believe it is of vital importance to emphasize the need to begin building a legal framework in this area based on the analysis and evaluation of the registrability of works on a case-by-case basis. The most important aspect will be to define the degree of participation of a person using Artificial Intelligence as part of their creative process, such that their imprint and originality are implemented in their work.

CHAPTER FIVE

ARTIFICIAL INTELLIGENCE. A GLOBAL COMPARATIVE VIEW



THE AGE OF AI IN CHINA: A PARTIALLY CONSIDERATION ON SOME NEGLECTED ISSUES

SUMMARY

I. Introduction **II.** A methodology: The Daoist philosophy can guide us to interpret the AI in China **III.** A paradigm and its shift: The rise and fall of a big power or many big powers **IV.** Does the AI in China verify the theory "Dong Sheng Xi Jiang"? **V.** China's response to the industrial revolution **VI.** The "great divergence" rather than the "great convergence" in AI between the US and China **VII.** China's Trying to Govern the AI **VIII.** In the parallel worlds, the different AI worlds? **IX.** Which world today's and future's China belongs to? **X.** Concluding remarks

> PANG, ZHONGYING

Scholar in humanities and social sciences, teaches Global International Relations/International Political Economy and Global Economic History at the Sichuan University in Chengdu, China. He is currently a Visiting Senior Fellow at the Yusof Ishak Institute (Institute of Southeast Asian Studies - IEAS), Singapore. He is a former Visiting Fellow at the Brookings Institution in Washington DC, United States.

I. INTRODUCTION

In the field of artificial science and technology, I have been not literate, although I have been learning the Computer Science to catch up the change 21st century science and technology brings. However, as a senior student of Global International Relations, I was honored to accept a great invitation from Dr. Arturo Oropeza García for UNAM's global project entitled "Artificial Intelligence: A new era in the history of humankind". Arturo asked me to comment on "Artificial Intelligence in China, its Advances and Pending Challenges".

In doing this most difficult assignment, I was so encouraged by parts of Chinese publishing market, which is still keeping at least effective contacts with their global partners. China's largest publisher, the CITIC Press, introduced a series important AI books from the US market, although the current Chinese-American relations are with geopolitical tensions. The Chinese version of the Age of AI and Our Human Future written by Henry Kissinger with Eric Schmidt, Daniel Huttenlocher was introduced and published by the CITIC Press[1]. An AI visionary, Dr Fei-Fei Li, a Chinese-American from Chengdu at Stanford University in San Francisco[2], published her best selling book *The World as I See: Curiosity, Exploration, and Discovery a the Dawn of AI* (Princeton University Press, 2023) by the CITIC Press in 2024. The CITIC Press is responsible to publish the Chinese version of Yuval Noah Harari's NEXUS: A Brief History of Information Networks from the Stone Age to AI[3]. Such books help deepen my understanding on the AI world.

1 See, [美] 亨利·基辛格[美] 埃里克·施密特[美] 丹尼尔·胡滕洛赫尔: <人工智能时代与人类未来>, translators: 胡利平 and 风君, 中信出版社, 2023.

2 https://profiles.stanford.edu/fei-fei-li

3 https://www.ynharari.com/book/nexus/

It is a great idea to have Henry Kissinger's views to help start our thinking on the subject of AI. Henry Kissinger thought deeply about the digital era. His *World Order: Reflections on the Character of Nations and the Course of History* (2014) includes a chapter entitled Technology, Equilibrium, and Human Consciousness[4]. In 2019, Kissinger wrote about the increasing tendency to give control of nuclear weapons to computers operating with AI[5].

Kissinger often drew an analogy between nuclear weapons and AI. He travelled to Beijing on July 2023 for his final global visit in person. The People's Republic of China has been with nuclear deterrence since 1960s. Kissinger visited China more than 100 times since 1971. HK published his On China in 2010.

China's official news reports on HK's farewell journey included no any content on AI. However, according to The Economist, a British global journal, during the meeting, among the words Kissinger delivered to Chinese leader Xi Jinping, he focused in part on how to protect the world from the dangers of AI[6].

In writing this piece, by doing literature review, I know China's top leader worries the AI's danger to the humanity[7] and Chinese diplomats call for must controlling AI[8].

In the following, based on my specialty in International Studies in Economy, History, Philosophy and Politics, I try to give an extremely partially assessment on AI in China.

II. A METHODOLOGY: THE DAOIST PHILOSOPHY CAN GUIDE US TO INTERPRET THE AI IN CHINA

The legendary founder of Taoism, China's only indigenous philosophical cornerstone of local culture for over two thousand years, is reputed to have been Lao Tzu/Laozi (老子), who's representative work is entitled Dao De Jing[9].

The 1st Chapter of the Dao De Jing(道德经):

4 Henry Kissinger, *World Order*, Chapter 9, 2014

5 https://en.wikipedia.org/wiki/Henry_Kissinger

6 https://www.economist.com/china/2024/08/25/is-xi-jinping-an-ai-doomer

7 https://www.economist.com/china/2024/08/25/is-xi-jinping-an-ai-doomer

8 Chinese Ambassador Fu Cong's remarks at the United Nations Security Council, https://www.mfa.gov.cn/eng/xw/zwbd/202410/t20241024_11515508.html

9 See, Wayne Ng, Finding the Way: A NOVEL OF LAO TZU， "Historical Note", Hong Kong: Earnshaw Books Ltd， 2018.

> "A Way called Way isn't the perennial/curtain Way. A name that names isn't the perennial/curtain name: the named is mother to the ten thousand things, but the unnamed is origin to all heaven and earth. In perennial/curtain Absence you see mystery, and in perennial/curtain Presence you see appearance/manifestations. Though the two are one and the same, once they arise, they differ in name. One and the same they're called dark-enigma, dark-enigma deep within dark-enigma, gateway of all mystery.(道可道 , 非恒道 ; 名可名 , 非恒名。无名 , 天地之始 ; 有名 , 万物之母。 故 , 恒无 , 欲以观其妙 ; 恒有 , 欲以观其徼。 此两者 , 同出而异名 , 同谓之玄。玄之又玄 , 众妙之门。)[10].

The topic sentence of above chapter is"In perennial Absence you see mystery, and in perennial Presence you see appearance/manifestations". I strongly think the sentence can serve as a universal guiding principle to partially know the existing world order of AI as well.In this article, I use it as my philosophical methodology to answer the assignment of "AI in China".

III. A PARADIGM AND ITS SHIFT: THE RISE AND FALL OF A BIG POWER OR MANY BIG POWERS

Many historians wrote many the rise and fall (兴衰, Xing and Shuai) of a big power or many big powers in the world. The rise and fall paradigm has been popular in Western and even Global academia. Historians and International Relations experts still like the paradigm very much. Since the late 1990s, China's rise has been a dominant topic in such Global writings including Global History and Global Political Economy.

Years ago, I purchased two Chinese versions of *Considérations sur les causes de la grandeur des Romains et de leur décadencee* (in English, *Reflections/Considerations on the Causes of the Grandeur and Decline of the Romans*) by French political philosopher, the author of The Spirit of Laws, Mon-

10 I reference two great books for better interpreting the Laozi or Lao Tze: David Hinton, *TAO TE CHING* (without the original Chinese version of the *Dao De Jing*), Berkeley: COUNTERPOINT, 2015. And, *THUS SPOKE LAOZI: A New Translation with Commentaries of DAODEJING* by Charles Q. Wu (吴千之), University of Hawaii (Honolulu) and Foreign Language Teaching and Research Press (Beijing), 2016. Dr Wu apparently was so inspired by German philosopher Friedrich Nietzsche's Thus Spoke Zarathustra (1885). It is Dr Wu's uniqueness in helping seizing the original meaning of the Laozi in Dao De Jing.

tesquieu(1689-1748). It is <罗马盛衰原因论>[11]. An old Chinese edition of the book before China's Cultural Revolution in 1962 and a new one in 2016 before the COVID-19, the both deeply reflect the Chinese long attachment to the greatest value of a most important European intellectual work. In other words, Chinese scholars well drew lessons of the rise and fall of a big power from the ancient Rome. At least since Montesquieu, the rise and fall of a big power or an empire or the rise and fall of many big powers (comparatively) has been an endurable paradigm not only for World Historians but for many strategists in various fields.

In the great era of China's opening up and reform (late 1970s to 2010s), China introduced many important literature of humanities and social sciences which dealt with the theme of the rise and fall of a big power or many big powers. Since the late 1990s, prominently, there have been several Chinese editions of Paul Kennedy's *The Rise and Fall of the Great Powers: Economic Change and Military Conflict* from 1500 to 2000, not surprisingly, almost all Chinese editions of the book have the same Chinese title: Da Guo Xing Shuai(大国兴衰)[12].

Since the early 1990s, while the world obsessed with the rise of China, few have also explored the fall of China[13]. Currently, the retreat of the rise of China structurally and partially is seen finally. At least China's economy faces huge challenges and even crises, the opponent narrative of the rise of China: the fall of China, is publicly opposed by both the official Chinese and the supporters of the rise of China narrative both homely and internationally.[14] Officially, China still publicizes its big or strategic judgement or theory that "the East (read: China) rises and continues to rise and the West (read: the United States) declines and continues to decline" (Dong Sheng Xi Jiang , 东升西降)[15]. The decline of China

11 See, (法) 孟德斯鸠: 《罗马盛衰原因论》(1951), 婉玲译(1962),许明龙译(2016), 商务印书馆.

12 I failed to check which one is the first Chinese edition of the Paul Kennedy's The Rise and Fall of the Great Powers: Economic Change and Military Conflict from 1500 to 2000, but I believe it was translated based on its first original English in 1990s.

13 For example, in 2001, Gordon G. Chang, *The Coming Collapse of China*, New York, Random House, 2001, 344 pp.

14 Interestingly, Singapore's Lee Kuan Yew School of Public Policy organized an event entitled "The China Shock at the End of History". Singapore's former ambassador to the the US Chan Heng Chee (former head of Political Science Department at NUS) argued the world including Singapore is now facing the 3rd "China Shock" - the falling of China's economy after the 1st and 2nd ones - totally the Rise of China at least economically since the early 1990s.

15 China's new theory of the "Dong Sheng Xi Jiang" is part of the Xi Jinping Thought. Amid the COVID-19 crisis, China's theorists in social sciences endorsed the Xi Jinping Thought by

as a narrative is clearly and even absolutely politically not correct in China. Still, no Montesquieu-like or Paul Kennedy-like frameworks to consider not only the rise but also fall of China.

In fact, take China as an example, its rise has been so complex in the past 4 decades since the late 1970s. It is a truly non-linear evolution. The fall too. The rise of China from 1979 to 1989 was stopped in 1989-1991. In 1992, just after the collapse of Soviet Union, China (read: Deng Xiaoping) relaunched its economic opening up and reform (**改革和开放**, Gai Ge and Kai Fang) towards a "Socialist Market Economy" (**社会主义市场经济**). The rise re-emerged in the whole 1990s amid globalization's popularization driven by the US-led neoliberalist economic and foreign policy. A fundamentally improved relationship between Beijing and Washington, which lasted in nearly 2 decades, Clinton-Jiang and Bush-Hu years (1992-2011), made China the "2nd largest economy"in a changed world economy: Global Economy. China was internationally accepted to be a full member of the post-Cold War international institution World Trade Organization (WTO) in 2001. In Hu Jintao years (2001-2011), China declared it has reached to a (world) great power (Da Guo, **大国)** status[16]. China joined the others in the form of the Group of the Twenty (G20) to help deal with the "Global Financial Crisis" triggered by the Wall Street during the Bush administration in 2008. The COVID-19 was tragic but unfortunately served as a turning point of history. Actually, since the end of the pandemic in late 2022, China has to deal with the big economic trouble.

IV. DOES THE AI IN CHINA VERIFY THE THEORY "DONG SHENG XI JIANG"?

The "Dong Sheng Xi Jiang" marks a China's "paradigm shift" – one power's rise and fall to the rise of a power (read: China) and fall of a power (read: the US). Simultaneously, China finishes its ideological shift from the Euro-centric or West-centric to the Sino-centric: China only sees the rise of itself and the fall of the other (read: the US).

But, the AI in China, seemingly, does not strongly verify the "Dong Sheng Xi Jiang".

arguing that China as the today's representation of the East continues to rise and the US as the today's representation of the West continues to decline.

16 Zheng Bijian, "China's "Peaceful Rise" to Great-Power Status", New York and Washington: *Foreign Affairs*, 2005. 84(5):18. in addition, Bijian, Zheng. *China's Peaceful Rise: Speeches of Zheng Bijian 1997-2005*. Brookings Institution Press, 2005.

The issue of AI, perhaps provides a unique window opportunity to examine the rise of China. Has China truly risen in the field of AI? Why has the AI sector in China developed? How about private AI investment in China? If the AI sector in China is not private economy (Min Ying Jing Ji, 民营经济) driven, how about the prospects of the AI in China?

Ninghui Sun, from the Institute of Computing Technology, Chinese Academy of Sciences (CAS) and an academician at the Chinese Academy of Engineering (CAE), is regarded as a pivotal figure in AI development in China. At the Standing Committee of the National People's Congress (China's Parliament), Sun gave a lecture/hearing entitled "The Development of Artificial Intelligence and Intelligent Computing" on April 26, 2024. In the lecture, Sun's conclusion is that no doubt the rise of China's AI continues but China faces tremendous challenges and uncertainties in the field of the AI.

"It was a strategic manifesto outlining China's AI ambitions and the associated anxieties. He emphasized AI's dual-edged nature, noting its potential for groundbreaking advancements while also highlighting security risks, particularly concerning politically sensitive information.The tech rivalry with the United States was another focal point in Sun's lecture. With superior talent, foundational algorithms, and computational power, the U.S. holds a commanding lead in AI. China's current role as a fast follower is a status quo that Beijing is desperate to disrupt. U.S.-imposed restrictions on high-performance computing products and advanced semiconductor technologies are significant obstacles, argued Sun. The ban on advanced chips like the A100 and H100 stymies China's AI progress, highlighting the strategic chokehold the United States maintains over China's tech aspirations. This impact is particularly harsh for companies like Huawei and SMIC. Sun proposed that China innovate indigenously and diversify supply chains to reduce dependency on U.S. technology, echoing Beijing's call to safeguard its high-tech future.He outlined three strategic pathways for China's AI development. The first involves aligning with U.S.-led systems, a practical but restrictive path given current geopolitical tensions"[17].

17 A text of Sun's lecture is published by China's National People's Congress as 十四届全国人大常委会专题讲座第十讲讲稿《 人工智能与智能计算的发展》,April 30, 2024, accessed at http://www.npc.gov.cn/c2/c30834/202404/t20240430_436915.html. A report on Sun's lecture, see, Felicity Hurley, "China's AI Ambitions: Prof. Sun Ninghui Highlights Strategic Blueprint", June 15, 2024, accessed at https://thefinancialanalyst.net/2024/06/15/chinas-ai-ambitions-prof-sun-ninghui-highlights-strategic-blueprint/

I guess Sun and his team in China's AI science and technology, is not necessarily for doing the "Dong Sheng Xi Jiang" and intentionally frame the rise of AI in China according to the new paradigm. His lecture may be interpreted as several meanings (Ji Ge Yi Si, 几个意思): Sun and his team still think China's AI or AI in China is rising or developing but in terms of the very old slogan since 1950s "Catching Up" (Zhui Gan, 追赶)[18] to the levels of AI innovation in the US, firstly; Secondly, he fully acknowledged China's AI development benefited greatly from the originality of AI in the US; Thirdly, he indirectly predicted a scenario of AI in China including the AI competition with the US is not optimistic. These meanings are not congruent with the "Dong Sheng Xi Jiang".

V. CHINA'S RESPONSE TO THE INDUSTRIAL REVOLUTION

While, in the US, since the dawn of the 21st century, China has been defined as a leading strategic competitor and rival in many aspects.

At a White House Roundtable on U.S. Leadership in AI Infrastructure on September 12, 2024, Jensen Huang (黃仁勳), an Chinese-American from Taiwan, the founder and CEO, Nvidia[19], re-stressed his firm argument and prediction that AI is the driving force of a new industrial revolution across the globe[20]. Huang has repeatedly addressed the "AI is the driving force of a new industrial revolution" in the globe[21].

China's last dynasty Qing (1644-1911) in the late 19th century much belatedly had to meet the challenges from the (1st) industrial revolution, which happened in Britain in the 18th century. Qing's industrialization - Westernization Movement (洋务运动, 1860s-1900s) really made something but generally failed: the movement did not narrow "the great divergence" between then China and Europe/West[22].

18 Sun as a generation of 1960s, still remembers a Mao Zedong slogan "Catching Up the United States" (赶上美国，Gan Shang Mei Guo) during the *Great Leap Forward*, one of the political movements before the Cultural Revolution in 1950s.

19 https://www.nvidia.com/en-us/about-nvidia/board-of-directors/jensen-huang/

20 https://www.whitehouse.gov/briefing-room/statements-releases/2024/09/12/readout-of-white-house-roundtable-on-u-s-leadership-in-ai-infrastructure/

21 Jensen Huang, A Keynote Speech at the COMPUTEX 2024, on June 2 in Taipei, Taiwan, see, https://www.nvidia.com/en-us/events/computex/

22 A global consequence of the Britain based Industrial Revolution in 18 century is the emergence of the so-called "The Great Divergence", which is suggested by the California

After 1949, China conducted a Soviet Union model of industrialization. The model largely did not work with huge human and natural resources costs in 1960s and 1970s. by "reform and opening up" towards a "socialist market economy with Chinese characteristics", China seized the opportunity of the coming of the information age in 1980s and 1990s. By joining the World Trade Organization (WTO) in 2001, China was seemingly successfully towards a market forces driven economy. China becomes a major "emerging economy" in the global economy.

Especially, exaggeratedly by ignoring existing big problems in China's research and development, China is widely regarded as a scientific superpower[23].

Is China taking advantages of the AI opportunity to succeed in the 21st Century Industrial Revolution?

With a sense of history, I have no immediate and irresponsible answer, but I have to leave this question open here: may China's record of responses and results by markets, societies and states to the previous industrial revolutions, both the failure and success - repeat in the future?

China is seen by some influential media in the US as a rising power of AI. New York based The Time100 names "Leaders", "Innovators", "thinkers", "Shapers" of AI annually. This following are from China[24]:

- Zhuang Rongwen, Director, Cyberspace Administration of China;
- Liang Rubo, CEO and co-founder, ByteDance;
- Wang Xiaochuan, Founder, Baichuan;
- Zhang Linghan, Professor at the Institute of Data Rule of Law, China University of Political Science and Law;
- Andrew Chi-Chih Yao, Tsinghua University.

Particularly, Andrew Chi-Chih Yao, 1946-, a recipient of The ACM A. M. Turing Award in 2000. Yao used the minimax theorem to prove what is now known as

School.Kenneth Pomeranz's The Great Divergence: China, Europe, and the Making of the Modern World Economy (2000, PUP) led to a global (including Chinese scholars) debate in the past 2 decades.

23 Many comments in the West think so. For example, **https://www.economist.com/science-and-technology/2024/06/12/china-has-become-a-scientific-superpower.** Such comments worry China will overtake the US in the future, see, Jeff Tollefson & Richard Van Noorden, "The US is the world's science superpower–but for how long?", **https://www.nature.com/articles/d41586-024-03403-4**

24 https://time.com/collection/time100-ai-2024/

Yao's Principle[25]. Yao was a native Chinese born in Shanghai. He returned to China in 2004, currently a professor and the dean of Institute for Interdisciplinary Information Sciences (IIIS) at one of China's top 2 universities Tsinghua in Beijing. Yao was a naturalized U.S. citizen and worked for many years in the US but renounced his U.S. citizenship in 2015 and became an academician of the Chinese Academy of Sciences[26]. Yao helped create China's first school of AI at Tsinghua University in April 2024 and was appointed as its founding dean[27].

Yao argues China's AI is already globally competitive, almost 50% talents of the global AI are trained in or from China, the country is a largest AI talents exporter in the world[28].

Yao's strategic importance has been realized by China's political leadership. Reportedly, Yao wrote a letter to report his works to Xi Jinping. The top leader replied Yao to congratulate his 20 anniversary returnee. Xi not only praised Yao but hoped Yao continues to help lift China's home-grown and autonomy-oriented innovation of AI science and technology[29].

This story between a laureate of the A.M Turing Award and China's top leader may imply many things strategically:

Some globally leading scholars in science and technology in China like Yao are well treated as China's "Strategic Scientists" (**战略科学家**). China develops the AI as a most strategic project like the country pursued its nuclear weapons in the late 1950s to 1960s, when two Communist states China and Soviet Union were on the brink of nuclear conflict. Beijing critically faced Soviet Union's use of nuclear weapons in the late 1960s[30]. Soviet Union obtained the nuclear weapons in 1950s after the United States in 1940s[31]. China's nuclear program got key assistance from Moscow.

---

25 "In recognition of his fundamental contributions to the theory of computation, including the complexity-based theory of pseudorandom number generation, cryptography, and communication complexity". See, https://amturing.acm.org/award_winners/yao_1611524.cfm

26 See, https://en.wikipedia.org/wiki/Andrew_Yao

27 After Tsinghua, many state-owned leading Chinese universities are establishing their schools of AI.

28 See, China's populist newspaper *Global Times under the People's Daily*, Beijing, June 29 2024, https://world.huanqiu.com/article/4IOhiaWznCs.

29 http://www.moe.gov.cn/jyb_xwfb/s6052/moe_838/202406/t20240612_1135187.html

30 "The Sino-Soviet Border War: Why the USSR nearly nuked China", **https://www.historyanswers.co.uk/history-of-war/the-sino-soviet-border-war-why-the-ussr-nearly-nuked-china/**

31 See, https://www.britannica.com/technology/nuclear-weapon/The-Soviet-Union

In economic development, China has claimed its "system advantages" (Ti Zhi You Shi, 体制优势). Angela Zhang argues China's authoritarian governance model can not only support the entire AI development but enable AI companies to train their models on vast amounts of personal data. In this sense, China obtains some "AI advantage" (Ren Gong Zhi Neng You Shi, 人工智能优势)[32].

Talents and innovation capitals are the key of keys in well developing AI in China. China faces challenges from the AI investors or entrepreneurs'"Great Escape"[33]. Reportedly, there is a wave of Chinese companies to the United States, particularly California's Silicon Valley to re-develop their enterprises, although they have to face the restrictions by the US government. Such Chinese have lost their confidence and opportunities in China. They believe they can take advantage of many opportunities artificial intelligence presents in the United States[34].

VI. THE "GREAT DIVERGENCE" RATHER THAN THE "GREAT CONVERGENCE" IN AI BETWEEN THE US AND CHINA[35]

Look at A Glossary of AI terms developed by Washington DC based think tank The Brookings Institution and its research partner at Tsinghua University in Beijing. The glossary shows some not small difference between the two AI[36].

Look at a newly release book *Nexus: A Brief History of Information Networks from the Stone Age to AI* by Yuval Noah Harari, who depicts a most dangerous future of China-US compared with yesterday's Soviet-US:

In CHAPTER 11 The Silicon Curtain: Global Empire or Global Split?, Harari's a key word is 'Silicon Curtain'. What does the "Silicon Curtain" mean? The UK-based Guardian comments that the geopolitics between China and the US is:

32 https://www.project-syndicate.org/commentary/how-china-uae-became-artificial-intelligence-powerhouses-by-angela-huyue-zhang-2024-09

33 The Great Escape, American war film, released in 1963, that was loosely based on the true story of an ambitious escape by Allied prisoners of war during World War II. See, https://www.britannica.com/topic/The-Great-Escape

34 Li **Yuan**, "Can China Tech Find a Home in Silicon Valley?",

35 The Great Divergence means the socioeconomic shift in which the Western world overcame pre-modern growth constraints and emerged during the 19th century as the most powerful and wealthy world civilizations, eclipsing previously dominant or comparable civilizations such as 19th century China. See, https://en.wikipedia.org/wiki/Great_Divergence

36 https://ciss.tsinghua.edu.cn/upload_files/atta/1725035303351_4A.pdf

extrapolating from the Great Firewall of China, which prevents most Chinese citizens from accessing sites such as Google and Wikipedia, Harari supposes that, in time, Chinese and US computer systems might be completely prevented from interoperating or even communicating with one another, so "ending the idea of a single shared human reality"[37].

The term "Silicon Curtain" is not creatively coined by Harari. New generation in the 21st century may totally forget the 'Iron Curtain' (**铁幕**, Tie Mu)[38]. Lindsay Gorman wrote in 2019 when the 30th anniversary of 1989, "A Silicon Curtain is descending: Technological Perils of the next 30 years"[39]. I believe most of Chinese don't know the term although they have known the "new Cold War" (Xin Leng Zhan , **新冷战**). However, Silicon Curtain is already interpreted into Chinese as **硅幕**, or Gui Mu).

Can China and the US avoid dropping a 'Silicon Curtain' to pursue their technological goals without abandoning open markets and compromising global innovation?[40]

VII. CHINA'S TRYING TO GOVERN THE AI

China initiates so-called Global AI Governance (Quan Qiu Ren Gong Zhi Neng Zhi Li, **全球化人工智能治理**). What does China's Global AI Governance mean and imply?

Put Governing AI rather than developing AI first. China has no Global AI Development Initiative but issued a Global AI Governance Initiative (GAIGI) in October 2023. To China, it seems, governing AI at multi-levels, including the global levels (for example, at the United Nations), is more pressing than developing AI.

---

37 See, "Nexus by Yuval Noah Harari review – end of days?", September 11, 2024, https://www.theguardian.com/books/article/2024/sep/11/nexus-by-yuval-noah-harari-review-the-ai-apocalypse

38 The Soviet not only refers to the Soviet Union. It represented the System of the "International Communist Movement" controlled by Moscow. People's Republic of China was part of the system from its founding 10 years to the early 1960s. As Beijing-Moscow split lasted from late 1960s to 1980s, it seemed China was no longer part of it. But, the PRC, in its nature, after the end of the Soviet Union, is still a full Soviet and even Stalinist state.

39 https://www.gmfus.org/find-experts/lindsay-gorman

40 Marina Yue Zhan (the author of Demystifying China's Innovation Machine: Chaotic Order), "The US and China must avoid dropping a 'Silicon Curtain'", Hong Kong: **South China Morning Post**, September 10, 2024.

It is a new China's new initiative after its *Big Three Initiatives*. The Big Three Initiatives have already been regarded as most important parts of *Xi Jinping Thought of Foreign Policy/Diplomacy* (习近平外交思想): Global Development Initiative (GDI), Global Security Initiative (GSI) and Global Civilization Initiative (GCI)[41]. Currently, China still not revised the Big Three Initiatives as the Bit Four Initiatives to include its GAIGI. But, GAIGI potentially becomes the fourth big initiative after the big Three. The GAIGI defines "governance" as "security" together with "development" of AI[42] and calls for stakeholders at the international levels to "develop AI governance frameworks, norms and standards based on broad consensus, so as to make AI technologies more secure, reliable, controllable, and equitable"[43].

Before the issue of the GAIGI, China began to organize the World AI Conference (WAIC) in 2018[44]. The Conference is now a government's *channel to* practice its Global AI Governance Initiative but serves as an AI learning platform for many Chinese elites. The 2019 WAIC was so impressed globally because China's Jack Ma and Elon Musk conducted a maybe historic MUSK-Ma Dialogue[45]. In the intense interaction, "Elon Musk and Jack Ma disagree about AI's threat"[46]. Jack Ma ignored the importance of AI and misinterpreted the AI. China's leading AI scholars and investors (most of them so young) largely praised Musk, China's official news agency Xinhua praised Ma very much[47].

The WAIC was paused during the COVID, but the WAIC 2024 in Shanghai continued to explore Global AI Governance.

41 "A Global Community of Shared Future: China's Proposals and Actions" by The State Council Information Office of the People's Republic of China, September 2023, see, https://www.gov.cn/zhengce/202309/content_6906335.htm

42 A fundamental "principle" of Xi Jinping Thought on Foreign Policy/Diplomacy is "equal emphasis on development and security" (发展和安全并重).

43 https://www.fmprc.gov.cn/web/ziliao_674904/1179_674909/202310/t20231020_11164831.shtml

44 https://www.worldaic.com.cn/wangjie?year=2019

45 https://www.youtube.com/watch?v=f3lUEnMaiAU

46 https://www.bbc.co.uk/news/technology-49508091

47 马斯克狠狠地给马云上了一课, https://user.guancha.cn/main/content?id=435375

THE SHANGHAI DECLARATION ON GLOBAL AI GOVERNANCE[48]:

"With the rapid development of AI technologies, we are also facing unprecedented challenges, especially in terms of security and ethics. We underline the need to promote the development and application of AI technologies while ensuring security, reliability, controllability and fairness in the process, and encourage leveraging AI technologies to empower the development of human society. We believe that only through global cooperation and a collective effort can we realize the full potential of AI for the greater well-being of humanity".

China promotes its *Global AI Governance Initiative* at the United Nations and other inter-governmental organizations. At the General Debate of the 79th Session of the United Nations General Assembly 2024, Foreign Minister Wang Yi, who is a Member of the Political Bureau of the Chinese Communist Party Central Committee, repeated China's existing doctrine and practice of AI Governance as below:

"China is committed to taking a people-centered approach, developing AI for good, and putting equal emphasis on development and security. We are working to explore and establish widely-recognized international rules and standards. China supports the U.N.'s role as the main channel for global AI governance and is committed to strengthening international cooperation on AI capacity-building. China has put forth the *AI Capacity-Building Action Plan for Good and for All*, and is ready to make more contributions to the sound, orderly, fair and inclusive development of AI"[49].

VIII. IN THE PARALLEL WORLDS, THE DIFFERENT AI WORLDS?

In 1950s, under the worst conditionality of the Cold War between the Soviet Union and the United States, the newly created People's Republic of China in 1949 was absolutely part of the International Communist Movement world dominated by Moscow. But,the death of Joseph Stalin in 1953, the PRC-Soviet Union tensions in the late 1950s and 1960s, and the USSR-USA (and China-US) "detente" in the late 1960s and 1970s, made Mao Zedong finalize his theory of "three worlds" (三个世界) before his death in 1976. Since then China began to identify itself as "a Third World country" or "a developing country" to pursue

48 https://english.news.cn/20240704/b1ca8d45f8f54ba9b954fcd969a2304f/c.html
49 https://webtv.un.org/zh/asset/k1v/k1vx3nr7l1

its more "independent foreign policy" within and beyond the International Communist Movement system.

After the Cold War, China decided to re-reform and opening up to be a member of globalization driven world/international order. In 2008, in the name of One World, China hosted the One Olympics in Beijing. Currently, China's "one foot" is still on the One World. There was a Chinese delegation of the 2024 Olympic Games, Paris, France. But, China's "one foot" is already on the another world[50]. China has been a pillar of the BRICS since 2009, which has been seeking an alternative world order.

Recently, an American scholar in International Relations and World Affairs consider "a dividing world": "Three Worlds: the West, East and South"[51], which is so different from his previous only one liberal international or world order.

More than 10 years ago, G. John Ikenberry asserted that"the rise of China will be integrated into the existing international order". He even deeply believed China is part of the existing One World[52]. Then a Shanghai-based newspaper used an very optimistic title to cover Ikenberry's lecture trip to Beijing: "Professor John Ikenberry: China is not another world"[53].

IX. WHICH WORLD TODAY'S AND FUTURE'S CHINA BELONGS TO?

The death of Soviet Union does not mean the death of the Soviet System. During the rapid rise of China years from 1990s to 2010s, China assured and reassured

50 There is a Winston Churchill **statue** on the **grounds** of the British **Embassy** in Washington, **DC, the US. "**Sculpted by William McVey, the statue has one foot on British soil, the other on American, symbolizing Churchill's honorary United States citizenship, and the ties between the two nations". See, https://winstonchurchill.org/resources/in-the-media/churchill-in-the-news/historian-richard-norton-smith-on-the-churchill-statue-at-the-british-embassy-in-washington/. Here, I borrow the "one foot on a land" and "the other on a land" to describe China's current international standing.

51 G John Ikenberry, Three Worlds: the West, East and South and the competition to shape global order, *International Affairs*, Volume 100, Issue 1, January 2024, Pages 121–138.

52 G. John Ikenberry, *Liberal Leviathan: The Origins, Crisis, and Transformation of the American World Order* (Princeton: Princeton University Press, 2011, which is translated into Chinese, entitled 自由主义利维坦: 美利坚世界秩序的起源、危机和转型. On January 2014, Professor Ikenberry visited Beijing and strongly confirmed his conclusion on future China as part of the liberal order, see, **https://en.iiss.pku.edu.cn/fj/PDF/ciss_en/upload/docs/2016-09-06/doc_9281473125195.pdf.**

53 See, https://www.thepaper.cn/newsDetail_forward_1247078

the others in the international system, particularly the US, "today's China is not yesterday Soviet Union"; China is impossible to become a Soviet state. But, finally, China is emerging as a truly heir of the Soviet Union in the world.

The author of The dean of Shandong: Confessions of a minor bureaucrat at a Chinese university, "a westerner's defence of China's political system"[54], Daniel Bell,thinks "The Communist Comeback"[55]. Bell's point of view is perhaps not accurate, but he observed a resurrection of yesterday's Soviet Union in today's China.In higher education of the ruling ideologies, China has run at least 1440 schools of Marxism until 2021 with more than 125 thousands staff professors[56]. Economically, the entire Chinese economy is no longer "market economy" but rapidly back to the centrally planed or command economy[57], although recently, at the 3rd the Third Plenary Session of the 20th Central Committee of the CCP, the market elements with "socialist" conditionality once again was mentioned[58].

AI, if developed over-rapidly, may accelerates the division of the dividing world. António Guterres, Secretary-General of the United Nations repeated his worry about AI at the general debate of the 79th Session of the General Assembly of the UN in New York on September 24, 2024:

"Artificial intelligence could lead to artificial divisions across the board –a Great Fracture with two internets, two markets, two economies– with every country forced to pick a side, and enormous consequences for all"[59].

António Guterres's words are not necessarily consistent with his deeds. The UNSG declined to attend the "(first global) Summit on Peace in Ukraine in Switzerland on June 15 2024 to enhance a high-level dialogue on pathways towards a comprehensive, just and lasting peace for Ukraine"[60]. But, he in the capacity of the UNSG spoke in in Kazan Russia on October 24, 2024 at the 16th BRICS Summit. His "We need peace in Ukraine. A just peace in line with the UN

54 https://www.ft.com/content/a2a8c734-49b4-4f23-b6bd-6b7f73fae250

55 Bell, Daniel. (2023) 2023. The Dean of Shandong. Princeton University Press.

56 The figure comes from China's Ministry of Education. see, http://www.moe.gov.cn/fbh/live/2022/54301/mtbd/202203/t20220317_608432.html

57 On the term "Command economy", see, https://www.britannica.com/money/command-economy

58 https://english.news.cn/20240718/e74d931886e64878ae6e5419d19a64da/c.html

59 **António Guterres, Secretary-General of the United Nations, Remarks at the Opening of the General Debate of the Seventy-ninth Session of the General Assembly, September 24, 2024.**

60 https://www.eda.admin.ch/eda/en/fdfa/fdfa/aktuell/dossiers/konferenz-zum-frieden-ukraine/Summit-on-Peace-in-ukraine-joint-communique-on-a-peace-framework.html

Charter, international law and General Assembly resolutions"[61] was responded differently in the world. Ukraine angrily criticized the UNSG's meeting with Russian President Vladimir Putin[62], who is a defendant accused of committing war crimes by the International Criminal Court (ICC)[63].

The case of China in development and governance of AI may strengthen the UN SG's concern and anxiety over the AI as a threat to human ontology in the exiting international order. In facts and trends, China has transformed itself as a representative of a new world: a politically unreformed Soviet-nature system with Chinese characteristics, a foreign policy to seek a multi-polar world order together with today's Russia, an intranet hard to connect with the global-nets, a dominant state-owned economy, a controlled society with the brainwashing as education for the next generations.

What are historic and strategic implications of China's new direction towards a new world for China's AI science and technology? Look at the future, do we witness the continuity of China's rise or do we see the end of the rise of China in the field of AI?

X. CONCLUDING REMARKS

By examining some neglected but important events, cases, figures and stories, this chapter looks at partially the progresses, problems and prospects of the AI in China.

I introduce one of China's current official theory entitled "the East rises and the West declines" to examine the AI in China, which has seemingly not demonstrated the theory.

Paul Samuelson (1919-2009), the author of a classic textbook Economics: an Introductory Analysis in 1948, an American economist who was awarded the Nobel Prize in Economic Science in 1970 for his fundamental contributions to nearly all branches of economic theory[64], who predicted the Soviet Union economy would overtake the US in 1980s[65]. But, his prediction did not happen.

61 https://www.un.org/sg/en/content/sg/statement/2024-10-24/secretary-generals-remarks-the-16th-brics-summit-delivered

62 https://news.sky.com/story/antonio-guterres-why-is-the-un-secretary-general-in-russia-incurring-ukraines-wrath-13240299

63 https://www.icc-cpi.int/defendant/vladimir-vladimirovich-putin

64 See, https://www.britannica.com/money/Paul-Samuelson

65 From The Vintage News, https://www.thevintagenews.com/2018/08/23/soviet-gnp/

The "Dong Sheng Xi Jiang" is not a Samuelson-like scientific research conclusion but a political ideology. The rise of the AI in China has been due to the rise of the AI in the US. This is an international result of mutually promoted interdependence. The primacy of the AI in the US is largely acknowledged by China's leading AI scientists, who have had connections with their American colleagues. The rise of AI in China not means the the decline of the AI in the US. AI development in China faces its sustainability challenge as China deviates truly market economy and re-embraces the Soviet Union centrally planed or command economy.

From China-US relations perspective, AI is so globally geostrategic. Currently, in the field of AI, the great divergence rather than the great convergence between China and the US is happening. In the "Age of AI", the globe is not an integrated and united world. In the future, the globe includes at least two worlds: China and the US are in different worlds.

THE IMPACT OF AI ON CHINA'S ECONOMIC DEVELOPMENT. ADVANCES AND CHALLENGES TO BE ADRESSED

SUMMARY

I. Introduction **II.** The Current State of China's Economy **III.** Main Advantages of AI Application in China **IV.** Challenges of AI Application in China **V.** Policy Recommendations **VI.** Bibliography

> JIN SONG
Director of the Development Research Center of the Institute of World Economics and Politics (IWEP) and Researcher at the Chinese Academy of Social Sciences (CASS)

I. INTRODUCTION

The unprecedented advancements in science and technology result profound transformations in the global political and economic landscape. A new wave of technological revolution is sweeping the world as evidence that the emerging fields such as artificial intelligence (AI), blockchain, and biotechnology are becoming pivotal engines of economic growth. AI, a branch of computer science, aims to simulate and extend human intelligence through technologies such as machine learning, natural language processing, and computer vision, and stands out as a transformative force with the potential to redefine industries and reshape economies. In recent years, significant breakthroughs in deep learning, natural language generation models (e.g., GPT), and autonomous systems have led the widespread application of AI in various fields such as healthcare, finance, energy, transportation, etc., which proves that AI is increasingly integrated into economic and societal operations and offes innovative solutions and driving efficiency.

In no exception, China's economic development in the global transformations is also at a historic crossroads. Having achieved decades of high-speed growth, China is now transitioning to high-quality development, which emphasizes innovation, sustainability, and inclusivity as the pillars of the progress. The Chinese government's focus on fostering "new quality productivity" aligns closely with the imperatives of the transition. It highlights the integration of advanced technologies like AI into economic systems to accelerate industrial upgrading, optimize supply chains, and build a modern, innovation-driven industrial framework. Besides, demographic changes in China further underscore the urgency of the transition. As the population growth slows and the labor force ages, the traditional model of growth driven by abundant low-cost labor faces

diminishing returns. It necessitates a shift toward productivity-led growth fueled by technological innovation. AI, as a key enabler of such productivity, could plays a leading role in addressing the challenges of an aging population, optimizing resource allocation, and enhancing economic competitiveness. Despite the fact that China's innovation-driven development strategy has already positioned the country as one of the global leaders in artificial intelligence research and application, the rapid integration of AI into economic systems also brings critical challenges such as ethical and regulatory concerns, and the impartiality to AI benefits. Responding effectively to these challenges will be core issues to China's high-quality development in the AI era.

This paper analyzes the primary advantages of artificial intelligence in China's economic transformation, lists the core challenges that need to be addressed and attempt to provide a systematic perspective on how AI can contribute to the high-quality development of China's economy.

II. THE CURRENT STATE OF CHINA'S ECONOMY

China's economy has entered a new phase of high-quality development after decades of rapid growth, and the economic operation now presents a series of new characteristics and challenges.

FIRST, ECONOMIC GROWTH RATE HAS DECLINED

The pace of China's economic growth has gradually slowed down from double-digit rates to a moderate-high range. Since the 2020s, the economic growth rate has noticeably decelerated with the gradual disappearance of the demographic dividend, weakening external demand, and the deepening of domestic economic structural adjustments. Though a slowdown is a normal phenomenon as an economy matures, the decelerating trend reveals deeper contradictions and issues such as insufficient domestic demand, slowing productivity growth, and uneven technological innovation capabilities. Additionally, uncertainties and complexities in the international macroeconomic environment, including geopolitical conflicts and global supply chain realignments, add further pressure on the growth rate. It necessitates that China should focus more on the development quality rather than the speed of economic growth to ensure a long-term sustainable development through optimizing resource allocation, strengthening innovation-driven strategies, and improving total factor productivity.

SECOND, TRADITIONAL INDUSTRIES NEED TRANSFORMATION AND UPGRADING

China's traditional industrial system, centered around manufacturing, has been driven by low-cost labor, resource inputs, and scale expansion. But with the adjustment of the international division of labor and changes in the stage of domestic economic development, the limitations of this model have become increasingly apparent. First, traditional manufacturing faces overcapacity. In certain industries, low-end and homogenized competition lead to resource wastage and declining profit margins. Second, the rise of emerging economies in mid-to-low-end manufacturing has intensified international competitive pressures. Meanwhile, the global economy is accelerating towards digitalization and greener practices, forcing traditional industries to leap towards higher-end, smarter, and greener operations through technological upgrades and structural adjustments. This not only involves nurturing high-tech industries represented by artificial intelligence, big data, and cloud computing, but also includes promoting the deep integration of manufacturing and services, establishing a modern industrial system powered by technological innovation.

THIRD, CHINA'S POPULATION AGING ISSUE HAS INTENSIFIED, AND THE PROPORTION OF THE WORKING-AGE POPULATION CONTINUES TO DECLINE

Statistics show that since 2012, the total number of China's working-age population has been decreasing, and it is projected that by 2035, the elderly will account for over 30% of the total population. Population aging impacts the economy in multiple ways. The decrease in the labor force directly weakens the potential momentum for economic growth, the pressure of pension payments and increased medical expenses further exacerbate the public fiscal burden, and aging significantly influences consumption patterns, with rising demands for healthcare and eldercare services, while growth in traditional consumer sectors slows. To address the challenges of aging, it is essential to increase labor productivity, to delay the retirement age, to expand the supply of skilled technical personnel, and to optimize the social security system to mitigate its impact on the economy.

FOURTH, THE LABOR SKILL STRUCTURE HAS BEEN SIGNIFICANTLY TRANSFORMED

In the past few years, the overall university enrollment rate in China has shown an upward trend, gradually stabilizing at a high level. Since the implementation of the higher education expansion policy in 1999, the gross enrollment ratio (GER) in tertiary education has experienced a significant increase, rising from 30% in 2012 to 59.6% in 2022. The underlying drivers behind the soaring rate in higher education enrollment include robust support from educational policies and a strong inclination of households to invest in education. China has consistently promoted the development of education, striving to enhance the accessibility of education and narrow the gaps between urban and rural areas, as well as among different regions. Simultaneously, with the ongoing economic transformation and technological advancements, there has been an increasing demand for highly skilled labor in society. Higher education is widely recognized as a crucial pathway to enhancing individual competitiveness. Given the high value placed on education by Chinese families, an increasing number of households are willing to provide support for their children to pursue higher education. All these have led to a notable shift in the skill structure of the labor supply.

FIFTH, CHINA FACES SIGNIFICANT PRESSURE FOR CARBON REDUCTION

Being one of the main carbon emitters in the world, China confronts significant pressures to reduce its carbon footprint. To achieve the goals of carbon peak and carbon neutrality, it is required to advance the quality of its economic development. Nonetheless, the endeavor to reduce carbon emissions is beset by numerous challenges. Firstly, the predominance of fossil fuels in China's energy mix, coupled with the relatively low share of clean energy, presents a formidable barrier. Secondly, the transition of traditional, high-energy-consuming, and high-emission industries is impeded by both technological hurdles and employment. The fact that these sectors often serve as major employers, where rapid reform could precipitate substantial job displacement. Furthermore, disparities in economic development and resource endowments across regions complicate the uniform implementation of carbon reduction objectives. To foster a green development trajectory, making the best use of technology progress is imperative for achieving a balance between economic growth and environmental sustainability.

III. MAIN ADVANTAGES OF AI APPLICATION IN CHINA

AI could play a significant role in China's economic development. Since the global economy has been characterized by sluggish performance, many countries urgently explore new sources or drives of growth. China has introduced the concept of "new quality productive forces", aiming to leverage the latest technological achievements, including AI, to promote economic development. Owing to its powerful capabilities in cross-domain data processing and analysis, AI technology can break down traditional barriers between fields, promote interdisciplinary integration, and spur innovative activities. It facilitates the research and development of new products, the improvement of modern services, and the innovation of business models, and gives rise to new economic opportunities. These innovative outcomes by AI not only significantly impact existing industries through creative destruction, transforming and elevating them, but also continuously stimulate economic expansion, fostering a new leap in social productivity and the overall high-quality development of the economy.

1. CONTRIBUTION TO INDUSTRIAL UPGRADING AND THE CONSTRUCTION OF A MODERN INDUSTRIAL SYSTEM

A modern industrial system requires the coordinated development of multiple sectors, including manufacturing and services, to achieve diverse and sustainable drivers of economic growth. In the past several years, Chinese traditional manufacturing and service industries have faced some major constraints due to a combination of internal inefficiencies and external pressures. The advent of AI has change the circumstances and it played a pivotal role in integrating and innovating these industries, driving the optimization and upgrading of industrial structures. AI has strengthened both internal advancements within industries and synergistic interactions across sectors, thereby enhancing overall economic efficiency, resilience, and competitiveness.

In the manufacturing sector, AI has significantly advanced automation and intelligence, fostering innovation, improving product quality, and enhancing delivery efficiency. These developments have enabled manufacturers to meet customized demands better from consumers, thereby increasing competitiveness in their products. AI empowers technological breakthroughs and provides positive influences, from production to resource allocation, in almost every

stage of the manufacturing process. Current research indicates that up to 45% paid work can now be automated through existing technologies, which facilitate real-time monitoring and adjustment of production lines. This leads to substantial improvements in product and service quality while reducing costs (Chui et al. 2016; World Economic Forum 2023). Moreover, AI enhances product delivery systems by analyzing large datasets to streamline distribution processes, significantly lowering operational costs, accelerating capital accumulation, and directing investments toward innovative areas. The shift encourages higher investment in research and development, thereby lays the foundation for sustained economic growth and expansion (Arntz et al. 2016; Brynjolfsson and McAfee 2014).

In the service sector, the integration of digitalization and intelligent technologies has enabled more efficient, accessible, and sophisticated service with improved quality and efficiency. For example, the labor substitution rate in the retail has increased from 53% to 60%, from 43% to 66% in finance and insurance services. Similarly, production and service models have shifted in healthcare and education. Across different industries, the service sector is providing more diversified support in high-quality to manufacturing and bringing new opportunities for Chinese manufacture. Correspondingly, manufacturing with great advancements offers new pathways for technological breakthroughs in services, creating a feedback loop that drives further mutual growth. This bidirectional synergy between manufacturing and services optimizes the structure of the labor market, generates higher-quality employment opportunities, and strongly supports sustainable economic development.

2. THE ENHANCEMENT OF LABOR PRODUCTIVITY

AI is reshaping the dynamics of labor productivity by complementing human efforts in the workplace. The synergy between AI and human labor has the potential to significantly boost efficiency, to enhance service quality, and to lower operational costs. In the context of China's economic transformation, where labor-intensive industries are giving way to knowledge- and technology-intensive sectors, AI represents a critical tool for achieving sustainable productivity gains.

First, AI helps to enhance efficiency through data processing and analysis.

AI improves operational efficiency across industries because it is able to process vast amounts of data and deliver real-time analytical insights significantly.

For example, in healthcare sector, AI helps in diagnostic processes and drug development. The AI-powered medical imaging systems have halved the time doctors spend on image analysis, reducing misdiagnosis rates and increasing the overall efficiency of healthcare delivery. In the financial sector, domestic banks have achieved up to 90% automation in transactions through AI simply because it allows financial professionals to focus only on complex investment decisions while minimizing operational costs. These advancements illustrate how does AI support routine tasks and how does it optimize high-value decision-making processes.

Second, AI leads to cost reduction through process automation.

AI excels in automating repetitive and time-consuming tasks, significantly lowering labor costs in various sectors. For instance, in customer service, AI-powered chatbots handle a vast number of routine inquiries, providing round-the-clock support while reserving human intervention for more complex issues. Major Chinese e-commerce platforms have adopted AI chatbots, cutting customer service costs by 30% while simultaneously improving customer satisfaction rates. This combination of cost savings and enhanced service quality highlights the transformative potential of AI in service-oriented industries.

Third, AI mitigates labor risks and increasing operational resilience.

The integration of AI and robotics further reduces labor risks and enhances productivity in high-risk and high-intensity environments. AI-powered robots are increasingly deployed in industries characterized by hazardous conditions, such as mining and chemical production, as well as in sectors requiring precision and endurance. These robots can operate continuously without fatigue, reducing workers' exposure to dangerous tasks and significantly lowering operational costs. Furthermore, the adoption of robotics minimizes labor disputes and reduces expenditures on welfare benefits and workplace safety, making it an attractive option for industries under cost and regulatory pressures.

Fourth, AI addresses China's demographic challenges.

AI's potential to replace and augment labor becomes increasingly valuable. Automation of routine and repetitive tasks helps mitigate the economic challenges posed by a shrinking labor pool. AI also complements human labor by enabling older workers to remain productive through assistive technologies, thus partially offsetting the demographic pressures on the labor market (Acemoglu and Restrepo 2018).

3. CATALYST FOR INNOVATION

AI stands out for its ability to simplify complex tasks and process vast datasets, a capability that distinguishes it from traditional approaches to innovation. Unlike human cognition, AI can search and analyze large-scale datasets within a remarkably short timeframe, leveraging its unparalleled computational power and expansive knowledge base. Such way of working enables AI to detect complex patterns through advanced techniques such as deep learning and predictive analytics, facilitating accurate forecasting, real-time decision-making, and an overall acceleration of the innovation process. Additionally, AI's robust cross-domain data processing capabilities break down the traditional disciplinary boundaries and foster interdisciplinary integration. The integration drives the development of new products, enhances modern services, and supports innovative business models, generating significant economic opportunities (Brynjolfsson and McAfee 2014). These outcomes from AI disrupt existing industries through mechanisms of creative destruction and stimulate continuous economic expansion. For instance, productivity gains driven by AI often lower the costs of goods and services, increasing real purchasing power and reshaping consumer demand elasticities, thereby contributing to economic growth (Bessen 2018).

AI significantly improves the model and efficiency of education and skills development of the talents, a cornerstone of innovation. Particularly in China, where rapid technological advancements are reshaping industries, traditional education systems often struggle with rigid curricula and inefficiencies, which is comparatively not conductive to labors' skills in rapidly evolving industrial demands. AI has significantly improved education and skill development through adaptive learning technologies. By personalizing learning content and methods, AI addresses individual learning needs, enhances skill acquisition and facilitates dynamic adaptation to new technological requirements. For example, AI-driven learning platforms in China are helping workers develop critical skills aligned with the demands of an innovation-driven economy. This capability elevates labor productivity and creates a robust talent pipeline to support sustained innovative activity (OECD 2019).

4. STIMULATING CONSUMPTION AND ENHANCES CONSUMER WELFARE

AI largely improves production efficiency and resource allocation and leads to enhanced product and service quality and lower prices, which are critical in

stimulating domestic demand. China's recent declining consumption results from a combination of structural and short-term factors. From the income perspective, slower growth in household incomes and widening income disparities have weakened consumption capacity and marginal propensity to consume. From the demographic perspective, accelerating aging and declining birth rates have altered consumption structures: a rising proportion of low-consumption elderly groups dampening overall consumption. Additionally, uneven urban-rural and regional development has constrained the release of potential consumption demand, while insufficient quality and alignment in product and service supply have limited the optimization and upgrading of consumption structures.

In general, AI reduces operating costs for businesses, drives down prices for products and services under competitive market conditions. It also optimizes supply chains and delivery experiences, increases consumer surplus and enhances overall welfare. The dynamic mechanism may even stimulate higher levels of consumption (Bessen 2018). The healthcare sector serves as a typical example. AI-driven technological advancements are profoundly transforming diagnostic and treatment methods, improving patient outcomes while effectively lowering medical costs and easing financial burdens, thereby enhancing welfare (Topol 2019). AI-enabled predictive analytics helps hospitals optimize resource allocation, reduce hospital stays, and improve patient welfare through multiple channels. In addressing the challenges posed by declining consumption demand in China's aging society, AI-powered health and eldercare services can meet the high-quality service needs of older populations, expand market participation among marginal consumers, and drive overall consumption growth.

Moreover, AI leverages sophisticated algorithms to analyze and interpret massive amounts of user data and then offer precise insights into consumer preferences and behaviors. This capability enables businesses to provide highly personalized products and services, significantly enhancing customer satisfaction and further improving consumer welfare (Davenport 2018; Deng et al. 2016). Technological innovation driven by AI presents opportunities for enhancing local product design and service quality, revitalizing domestic manufacturing and service industries, and partially substituting imported products and services, thus fostering domestic economic development. Through intelligent logistics, e-commerce platforms, and optimized allocation of public service resources, AI wonderfully boosts economic vitality and consumption capacity in underdeveloped regions..

5. ENHANCING RESOURCE EFFICIENCY, FACILITATING EMISSIONS REDUCTION, AND SUPPORTING ENVIRONMENTAL SUSTAINABILITY

AI also serves in a critical position in improving resource efficiency, reducing emissions, and promoting sustainable development. Its capabilities in predictive analytics, real-time monitoring, and resource optimization significantly enhance efficiency in key sectors such as transportation, manufacturing, construction, energy, and agriculture. In transportation, intelligent systems optimize traffic flow, and reduce fuel consumption and greenhouse gas emissions. In manufacturing, AI improves resource utilization by optimizing supply chains and production processes. Smart grids and AI-driven building energy management systems reduce energy waste. In agriculture, AI's precision minimizes water and fertilizer usage, and promote the sustainable practices in agricultural production. (World Economic Forum 2018; Chui et al. 2016).

For China, AI is particularly significant in advancing sustainable development. The country has committed to achieving carbon neutrality by 2060 and addressing environmental challenges, with a phased goal of peaking carbon emissions by 2030. These national objectives align closely with AI's potential to support the green transformation of traditional industries. AI-driven innovations, such as smart grids and green technologies like electric vehicles, contribute to reducing resource consumption, lowering pollution, and improving energy efficiency. These advancements support China's transition to a high-quality growth model that balances economic development with ecological conservation.

AI enables China to harmonize economic growth with environmental protection and avails the country advantageously in the global competition for sustainable development by promoting green technologies and circular economy practices. In the meantime, these efforts enhance societal well-being, demonstrating the dual benefits of AI-driven innovation in fostering economic prosperity and ecological sustainability.

IV. CHALLENGES OF AI APPLICATION IN CHINA

The widespread application of AI deeply reshapes the labor market in downsides, such as employment inequality, skill mismatches, income disparities, and governance issues. These challenges are adventurous for businesses, individuals, and policymakers. Unlike influences from any previous technological revolutions, the impacts of AI are faster, broader, and more complex. China without exception

also faces competitive risks stemming from global power rivalries though it leverages AI-driven advancements to foster economic development.

1. PROFOUND TRANSFORMATIONS IN THE LABOR MARKET

The adoption of AI is redefining skill requirements. In accordance with the latest report by the World Economic Forum, from 2023 to 2027, globally approximately 23% of jobs would be affected by technological disruption, involving 69 million new jobs and 83 million eliminated positions, which will result the profound changes in occupational structures across industries (World Economic Forum 2023).

AI-related positions, such as machine learning experts, robotics engineers, and business intelligence analysts, are in high demand that advanced skills like data analysis, machine learning, and algorithm development are urgently needed. Conversely, the existing clerical roles, such as bank tellers, postal workers, and ticket agents, are rapidly disappearing. Workers in traditional roles often face skill difficulties in meeting new demands, and the consequential mismatches between skill supply and demand appear,, which further require structural adjustments driven by automation and AI.

The paradox is that businesses cannot recruit their expected highly skilled talents, which consequently hampers the pace of AI deployment and innovation, and a large number of workers, who are not equipped with necessary skills, are forced out of the labor market and suffering from distress to reemployment. The mismatch is exacerbated by the slow adaptation of education and training systems, intensifies the labor market imbalances and leaves many workers in impotence in the evolving job demands.

AI's impact concentrates in technology-intensive industries and geographically works in relevant regions though its effects are broadly felt across sectors and locations due to its algorithm –and software-based nature. The productivity gains from AI also vary widely among firms, and the imbalance production returns exacerbate the inequalities of employment distribution. Capital –and technology-intensive companies achieve remarkable productivity improvements through AI-driven supply chain optimization and process innovation, but labor-intensive and low-tech firms often lack the resources and capabilities to integrate AI effectively, so they sometimes got stranded to stagnant or declining productivity and reduced competitiveness.

The productivity gap, often leading to widespread closures, enables high-tech firms to capture larger market shares, while low-tech firms face survival

challenges. High-tech enterprises experience surging demand for skilled talent, whereas traditional industries reliant on low-skilled labor face severe unemployment issues. Such polarization highlights the uneven distribution of opportunities and risks within the labor market.

Additionally, the distribution of displaced and created jobs is uneven across labor groups, potentially leading to significant structural unemployment. AI primarily targets cognitive tasks, including both routine tasks, such as customer support and data entry, and non-routine tasks, such as financial forecasting, medical diagnosis, and legal consulting (Webb 2020). By altering task structures, AI substitutes some human tasks while creating new tasks requiring different skills (Trajtenberg 2018).

Following employment disruptions, many workers are likely to enter informal employment. Some may face temporary unemployment and require retraining to secure new jobs, while others with weaker adaptability may experience long-term unemployment or even exit the labor market altogether. These outcomes depend on workers' attributes and the degree of technological disruption within their industries (Chui et al. 2016).

2. POTENTIAL RISKS OF INCOME INEQUALITY

Even though AI and intelligent robotics significantly enhance labor productivity, their impact on labor income growth is more complex. Unlike traditional automation, which predominantly affects low-income, low-skill workers, AI's influence on income is multifaceted. It disrupts jobs ranging from simple, routine tasks like data entry (typically middle-income white-collar roles) to high-income professions such as doctors, designers, and financial analysts (Trajtenberg 2018). This suggests that AI may disproportionately impact middle- and high-income workers.

Studies indicate that wage inequality may decrease during this process (Webb 2020). However, as some middle- and high-income workers shift to lower-skill jobs during reemployment, the labor oversupply also depresses low-skill workers' wages, suppressing overall income growth. The final distribution of labor income depends on the reemployment outcomes of displaced workers. It is a key policy challenge to ensure the workers share equally in the dividends of AI technology to eventually achieve equitable income distribution.

Regarding specific job-level income changes, AI has created numerous new technology-related roles and could create new high-income groups, such as AI specialists, who may temporarily lead to wage growth in some sectors. Current

high-income groups, such as doctors and financial analysts, must re-adapt their job roles or employment models in response to AI advancements to avoid temporary income declines. Displaced workers unfortunately may face direct income losses.

During the period of significant employment transition, retained workers may find themselves in disadvantageous position during labor-capital negotiations and have to take the downward wage pressures. In the long run, labor supply will gradually adjust in line with the new skills and job demands, and would form a convergence in wage advantages for newly created roles.

Overall, the introduction of conventional robotics shifts income from labor to capital, while AI and intelligent robotics may enable wealth accumulation for those with access to capital, data, innovative algorithms, and specialized expertise. Nevertheless, automation could lower the costs of goods and services, potentially enhance the effective purchasing power of existing incomes if cost reductions are passed on to consumers effectively.

3. COMPETITIVE RISKS ARISING FROM GREAT POWER RIVALRIES

Over the past several years, technological competition among major powers was fierce and insensitive and becomes more intricate and strategic. Higher levels of innovation and technological sophistication are critical for industrial advancement and enhance national competitiveness despite increasingly insurmountable technological barriers in developing countries. To secure a competitive edge in global industries, developed nations such as the United States, European Union members, and Japan have implemented a range of technology and industrial development strategies. These strategies position technological innovation and digitalization as the driving forces for future industrial development, and emphasizes the cultivation and attraction of highly skilled and qualified talent and fostering advancements in AI, the Internet of Things, and cloud computing. Developed nations leverage their mature innovation ecosystems and higher education institutions to strategically enhance the cultivation and attraction of top talent, particularly in science, technology, engineering, and mathematics (STEM). In 2020, over half of the PhD students in the US STEM programs were international students, with 73% of them staying in the United States after graduation. High retention rates among students who stay in the US are from China and India. For example, nearly 90% of Indian students earning STEM doctorates in 2020 stayed in the US. Following the 2018 US-China trade

frictions, the United States further intensified economic decoupling measures, including restricting technology exchanges, safeguarding critical technologies and sensitive data, and enhancing technology controls through patents, intellectual property protections, and market entry barriers. Developed nations also impose stringent market access standards and certification requirements to limit developing countries' products and services in their markets. Export controls and restrictions on technology transfer further inhibit the flow of critical technologies to developing countries. These measures exacerbate global technological inequalities, hinder industrial upgrading in developing nations, and reduce their participation in global value chains.

AI advancements compound these challenges, significantly shrinking the labor cost advantages of developing countries. AI, in conjunction with automation, reduces the demand for human labor, lowers production costs, and shifts capital investments toward innovation and efficiency improvements in manufacturing and services. These dynamics diminish developed nations' reliance on low-skilled labor from developing countries, and result in reduced investment and slower industrial upgrading in those nations. Furthermore, developed countries' monopolistic control over high-tech industries limits the ability of developing countries to transition to higher-value-added industries, and stalls employment shifts toward higher-skilled roles. Without domestic innovation capabilities, developing countries may over-rely on imported AI technologies, which will cost additional expenses and increasing production vulnerabilities, potentially lead themselves to technology-related debts. As the primary contributors to global population growth, developing countries face unique challenges under an international landscape marked by development asymmetries and monopolized opportunities. Population surpluses may fail to contribute to economic potential. Instead, it may cause widespread unemployment, deprivation of economic and educational opportunities, overconsumption of resources, and regional instability. Domestically, high unemployment, social security pressures, and governance challenges escalate fiscal burdens And internationally the diminished economic and technological leverage of developing countries further erodes their voice in global affairs (Manyika and Sneader 2018; Reijnders and Vries 2018).

4. SOCIAL AND ETHICAL CHALLENGES

AI holds immense potential to drive economic and social development, yet several challenges must be addressed before its positive impacts can be fully

realized. From a technological perspective, AI systems still require significant refinement. Errors in analysis results during reinforcement and deep learning processes hinder reliability and stability, which are critical for industrial applications. The discrepancy slows the pace of AI adoption relative to technological advancements. A key controversy surrounding deep learning lies in the opacity of its outcomes. Application scenarios such as healthcare demand transparency in the mechanisms underlying AI-generated outputs. The lack of clarity imposes a constraint on the pace of AI application as it is crucial to ensure its development does not outstrip human understanding.

AI also raises significant social and ethical concerns, which can limit its adoption. Beyond structural impacts on employment, AI's discriminatory effects may arise from its operational mechanisms. AI systems heavily rely on available data for reinforcement learning, and biases in such data can amplify algorithmic prejudice, exceeding that of human biases. This distorts production, service delivery, and resource allocation. Unchecked AI applications, such as generating fake content, pose threats to economic and social stability. Ethical principles guiding AI operations, as well as its interaction with humans, remain contentious. On a technical level, safeguarding data privacy is a major challenge, with data breaches posing risks to public safety and providing opportunities for criminal activities. Establishing robust regulatory frameworks to manage AI development and application is essential to address these issues (Brynjolfsson and McAfee 2014; Acemoglu and Restrepo 2019; Topol 2019).

V. POLICY RECOMMENDATIONS

AI serves as a critical driver for China's economic advancement, reshaping the modern industrial ecosystem. To maximize AI's positive impacts while mitigating its challenges, China must establish a comprehensive system that accelerates industrial upgrading and ensures efficient talent cultivation and allocation.

1. ESTABLISHING AN AI-COMPATIBLE TALENT STRUCTURE

Firstly, cultivating diverse talent profiles is important. On one hand, China needs leading talents with innovative thinking, entrepreneurial spirit, and global vision, as well as interdisciplinary talent capable of integrating advanced technologies and achieving "new combinations" of production factors for profit growth. Such individuals play pivotal roles in adapting to dynamic markets, driving industry innovation, and expanding global market presence. On the other hand, it is

essential to have ordinary workers with solid technical skills and digital literacy form the foundation for AI integration. Workers must maintain competitiveness, master their roles while developing skills in AI and data analysis to enhance efficiency, adapt to technological changes, and transition to tech-intensive jobs.

Secondly, education system reforms are needed. Traditional education in China appears some defects in flexibility and efficiency in responding to rapidly evolving labor market needs. There are two directions that the education system could be improved. One is innovative talent development: Shifting educational objectives from knowledge transmission to fostering creativity, critical thinking, and problem-solving skills; creating an open, interactive education ecosystem emphasizing interdisciplinary learning to break traditional academic silos; Incorporating cutting-edge technology trends and societal needs into curricula, and using AI tools for personalized learning experiences and progress monitoring. The other is collaborative education mechanisms: promoting dynamic collaboration among government, enterprises, schools, training institutions, scientific representatives, and families; aligning education with future industry trends and emerging occupations, and encouraging lifelong learning through policy incentives like subsidies to ensure th workforce adaptability to evolving industry demands.

2. ESTABLISHING A MARKET ENVIRONMENT FACILITATIVE OF AI DEVELOPMENT AND APPLICATION

It is important to cultivate a competitive market environment too. A competitive market environment for AI includes four components.

Firstly, establishing a healthy market ecosystem that can benefit countries to adapt to the AI era.

China boasts a large and mature market for the development and application of AI technologies, supported by a robust digital infrastructure – a unique advantage not available in many other countries. Efforts should be made to further cultivate this market towards maturity and healthiness. It involves opening up and jointly building an environment conducive to international cooperation in AI technology, which includes platforms that promote mutual benefits and win-win outcomes among nations. It can create a relaxed international exchange environment by facilitating international exchanges of experiences, technologies, and talents through diversified channels such as talent exchange programs, cross-border research platforms, remote work systems, and international academic conferences. Regular interaction between domestic and international talents,

which is thought to be contributive to a prosper global innovation ecosystem, will form a global network of talent exchange characterized by mutual benefit.

Secondly, ensuring orderly market competition during the process of transition.

Enterprises mastering AI and robotics technologies gain first-mover advantages, which can reshape the competitive landscape within industries. Given the high entry barriers associated with these technologies, small and medium-sized enterprises (SMEs) may find it challenging to enter the market, potentially leading to increased industry concentration. To maintain a competitive market structure, it is imperative to strengthen antitrust regulation, particularly focusing on enterprises with core AI and robotics technologies, especially leading firms and major internet companies, to prevent abuse of dominant market positions. Support for SMEs' development is also crucial. Providing policy support to assist SMEs in acquiring funds, technologies, and talents can be adopted to enhance their competitiveness. Promoting open sharing of technologies can encourage technological innovation and application. Nevertheless, the government should ensure smooth mechanisms for enterprise exit, avoiding excessive local government intervention or protection that leads to resource misallocation, particularly unnecessary subsidies for inefficient manufacturing sectors, to prevent hindrance to industrial upgrading and labor reallocation.

Thirdly, enhancing labor market flexibility and resilience and mitigating the challenges brought by AI to employment.

Addressing changes in employment structures brought about by new forms of productivity requires enhancing the flexibility and resilience of the labor market. Technological innovations and industrial upgrades will likely increase volatility in the job market and raise the proportion of informal employment and posing short-term or long-term occupational transition challenges for some workers. Employment management strategies need to balance maintaining employment stability within ensuring healthy market development and avoiding the accumulation of ineffective employment. Concurrently, efforts should be made to build proactive employment support and social security systems. Measures such as optimizing the market environment by accelerating the construction of a unified, open, and orderly national market, reducing administrative barriers, lowering enterprise operating costs, strengthening property rights and intellectual property protection, and enhancing the attractiveness of the business environment can indirectly promote the increase and improvement of employment opportunities. And means of improving the functionality of employment information service platforms by broadening channels for employment information release

and offering customized re-employment training services can accelerate skill upgrades and job matching for workers, mitigating adverse effects of frictional unemployment.

Fourthly, the refinement of a dynamic and adaptive social security system is essential.

China has already established a comprehensive and well-organized social protection network. However, in the era of rapid advancements in AI, it is imperative to bolster the system's dynamism and precision to effectively support diverse population segments. It helps individuals navigate through the challenges posed by AI, ensures that the system remains responsive to the evolving socio-economic landscape, and thereby support the healthy development of new forms of productivity. The system should smooth transition mechanisms and facilitate the seamless transition of affected workers into new roles or industries. The process includes providing personalized career counseling, retraining programs, and vocational education opportunities tailored to individual needs, ensuring that displaced workers can quickly adapt to emerging job markets. Proactive measures should be taken to reduce the impact on vulnerable groups, and then to prevent social unrest. Also, by protecting workers' livelihoods during transitions, it encourages greater willingness to embrace new technologies and participate in lifelong learning, which is critical for sustaining long-term economic growth.

VI. BIBLIOGRAPHY

Acemoglu, Daron, and Pascual Restrepo. 2019. "Automation and New Tasks: How Technology Displaces and Reinstates Labor". *Journal of Economic Perspectives* 33(2): 3-30.

Acemoglu, Daron, and Pascual Restrepo. 2018. "Artificial Intelligence, Automation, and Work". In *The Economics of Artificial Intelligence: An Agenda*, edited by Ajay Agrawal, Joshua Gans, and Avi Goldfarb. University of Chicago Press.

Arntz, M., Terry Gregory, and Ulrich Zierahn. 2016. "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis". OECD Social, Employment and Migration Working Papers, No. 189. Paris: OECD Publishing.

Bessen, James. 2018. "Artificial Intelligence and Jobs: The Role of Demand". In *The Economics of Artificial Intelligence: An Agenda*, edited by Ajay Agrawal, Joshua Gans, and Avi Goldfarb. University of Chicago Press.

Brynjolfsson, Erik, and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company.

Chui, Michael, James Manyika, and Mehdi Miremadi. 2016. "Where Machines Could Replace Humans–and Where They Can't (Yet)". *McKinsey Quarterly*.

Davenport, Thomas H. 2018. *The AI Advantage: How to Put the Artificial Intelligence Revolution to Work*. The MIT Press.

Deng, Yue, Feng Bao, Youyong Kong, Zhiquan Ren, and Qionghai Dai. 2016. "Deep Direct Reinforcement Learning for Financial Signal Representation and Trading". *IEEE transactions on neural networks and learning systems 28(3): 653-64*.

Manyika, James, and Kevin Sneader. 2018. "AI, Automation, and the Future of Work: Ten Things to Solve For". McKinsey Global Institute. https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for.

OECD, 2019. *Education at a Glance 2019: OECD Indicators*. OECD Publishing.

Reijnders, Laurens S., and Gaaitzen J. de Vries. 2018. "Technology, Offshoring and the Rise of Non-Routine Jobs". *Journal of Development Economics* 135: 412–432.

Topol, Eric J. 2019. "High-Performance Medicine: The Convergence of Human and Artificial Intelligence". *Nature Medicine* 25 (1): 44–56.

Trajtenberg, Manuel. 2018. "AI as the Next GPT: A Political-Economy Perspective". National Bureau of Economic Research Working Paper No. 24245. National Bureau of Economic Research. https://doi.org/10.3386/w24245.

Webb, S., 2020, "The Impact of Artificial Intelligence on the Labor Market". https://web.stanford.edu/~mww/webb jmp.pdf.

World Economic Forum, 2018. *Future of Jobs Report 2018*. World Economic Forum. https://www.weforum.org/publications/the-future-of-jobs-report-2018/.

World Economic Forum, 2023. *Future of Jobs Report 2023*. World Economic Forum. https://www3.weforum.org/docs/WEF_Future_of_Jobs_2023.pdf

ADVANCED GENERATIVE ARTIFICIAL INTELLIGENCE AND AI GOVERNANCE IN JAPAN

SUMMARY

I. Introduction **II.** The Rise of LLM Generation AI and Multimodal AI **III.** Hiroshima AI Process **IV.** Biden Harris Presidential Decree **V.** AI User Guidelines by Japan Government **VI.** Consideration of more advanced AI governance and AI legislation **VII.** Toward a New Vision of Social Development **VIII.** Bibliography

> OSAMU SUDOH
Ph.D. and Professor, Chuo University

I. INTRODUCTION

This paper reviews the technological trends in generative and multimodal AI, which are bringing about rapid changes in social systems and will be a paradigm shift that will mark an epoch in human history. Based on this, the significance of the "Hiroshima AI Process" proposed and approved at the G7 Summit 2023, which Japan chaired, as a trend that attempts to govern cutting-edge AI, and the Biden-Harris Presidential Decree (October 30, 2023) of the United States that supports this process will be discussed. The significance and challenges of the "AI User Guidelines", which were adopted as policy in Japan to implement these guidelines in a more concrete manner, and the basic policy of the AI legislation currently under consideration will be described. In doing so, we will describe the progress of Japan's international contributions to advanced generative and multimodal AI.

II. THE RISE OF LLM GENERATION AI AND MULTIMODAL AI

GPT3.5, a large-scale model for natural language processing generated by Open AI, was released in November 2022, and GPT4 was incorporated into Microsoft's search engine Bing this year; Bing is now able to write natural language prompts for Dall-e generated images.

Meanwhile, Google released Bard, a generative natural language large-scale data model, and PaLM2 was released on May 10, 2023.

Since then, the trends surrounding generative AI have had a tremendous impact on people to this day. AI as a superior tool (optimization tool) that optimizes for a given objective using a given set of means has changed dramatically. Major changes have been prepared since around 2018 at leading American companies and research and educational institutions such as Google, OpenAI,

Microsoft, Meta, IBM, Stanford University, and MIT (Massachusetts Institute of Technology) (1).

One of these was the Generative Pre-Transformer (GPT), the core of which was the development of the Transformer (2).

Transformer is a five-stage natural language processing mechanism that processes natural language, understands context, and generates words through large data sets and high-speed computer processing. The first stage consists of "tokenization of words", the second stage "embedding by vector transformation", the third stage "positional encoding", the fourth stage "transformer blocking", and the fifth stage "softmaxing using probability. These processes understand the context and generate the words that should be displayed next. In addition, a convolutional neural network developed by Google is used for the generative image processing (see Fig. 1).

FIG.1 Five elements of a transformer

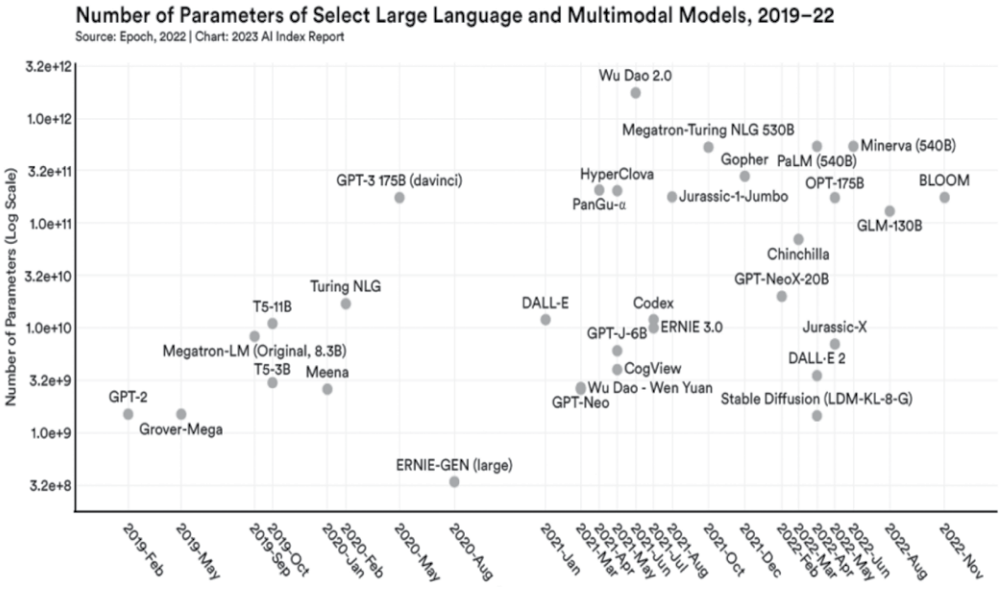
- Tokenization
- Embedding
- Positional encoding
- Transformer block
- Softmax

The text is generated through the above five steps of operation.
Mathematical processing with vector databases, context dependence and probability distributions are the main features.

This process requires a large database and fast computing resources. GPT-3, released by Open AI in the spring of 2020, has 175 billion parameters. The improved GPT-3.5, which appeared in November 2022, is also estimated to have 175 billion parameters. Even more astonishing is Google's transformer PaLM, which has 5400 parameters, and Minerva, which performs mathematical inference, which has 5400 parameters (see Fig. 2). Incidentally, Google unveiled its fine-tuned PalM-2 at Google I/O 2023 in May 2023. Gemini1.0, an early model of integrated multimodal AI R&D for mathematical reasoning and analysis of chemical formulas, DNA sequences, etc., was also released in 2023 (3).

In Japan, the Ministry of Economy, Trade and Industry has a well-known supercomputer called ABCI, which is currently being augmented. Noteworthy trends in the U.S. are those in the enhancement of mathematical and logical reasoning capabilities. For example, Minerva, a generative AI for research that was at a high level of capability as of spring 2023, is an AI that specializes in computational logical reasoning in mathematics, physics, pharmacology, and chemistry, but this was an artificial intelligence capable of running 540 billion parameters.

FIG.2 Parameters of LLM Generative Models



[SOURCE] Stanford University Human Centered AI Institute ed. Artificial Intelligence Index Report 2023, p.60

The parameters of GPT-4, which have not been made public, are said to be in excess of 1 trillion according to American researchers, and some estimate that GPT-5, which is said to be in the study phase by 2024, will be in excess of 4 trillion. The computing power of Japan's supercomputers is currently unsurpassed.

We believe that the construction and strategic use of Retrieval Augmented Generation (RAG) will become increasingly important in the future, The technology extracts information from a variety of knowledge sources that are not publicly available on the Internet and generates text from them. In the next step, the "Generator Model" uses this information to generate answers to questions and text.

After 2022, LLM-generated AI using huge supercomputers is on the rise, but recently Retrieval Augmented Generation (RAG) has been attracting a lot of attention. Instead of collecting a large amount of data from all over the world to make judgments or write sentences, for example, a RAG can process a university's confidential information, such as management, training and education, know-how, and personal information of each faculty member, using only that information. With such a volume of information, there is no need to use a high-performance supercomputer like GPT4, and RAG will be able to produce

documents with high efficiency and quality. It will also be able to evaluate reports. If this happens, we can have artificial intelligence do all the evaluation of reports on behalf of the teacher. In Japan, not only IT companies, but also many manufacturing companies, financial institutions, and local governments are focusing on RAG construction and operation (4).

With RAG, Japanese companies do not need a supercomputer and can operate with a computer of a certain level of capability, and can process important information without exposing it to the outside world. Cloud computing allows for thorough management of personal and confidential information, processing can be done in a closed environment, and copyright management is easy.

Microsoft, AWS, Oracle, Mistral, and others are also focusing on RAG. In Japan, NTT Data, Hitachi, Fujitsu, NEC, and many other IT companies are focusing on RAG development and operation. However, technically, it is not an easy task, as it requires the competence to create advanced vector databases. The way to create vector database attributes and other information is important (see Fig. 4).

FIG.3 What is Retrieval Augmented Generation (RAG)?

RAG is a widely used and important technology in natural language processing (NLP). This technology extracts relevant information from large knowledge sources and generates text from it.

The RAG consists of two components: the <retrieval model> and the <generator model>.

>The "Retrieval Model" searches for relevant information from a large amount of vector data for a given query (an imperative in DB). The Generator Model then uses this information to generate an answer or sentence to the question.

✓ GPT4 services using closed organizational data are expanding: MS, Google, AWS, Oracle, Meta, Anthropic, Mistral, NTT Data, Hitachi, Fujitsu, NEC, etc.

With the utilization of this RAG, advanced AI services can be provided for the strategic use of generative AI inside local governments and for strategic management inside corporate organizations. In this sense, RAGs are likely to steadily have a significant impact on governments, municipalities, and other administrative agencies, financial institutions, and other corporate management LLM generative AI and multimodal AI, especially RAGs, will have a significant impact on human creative and strategic actions.

Then there is the RAG system, which uses the RAG database in a cozy way and can run in conjunction with a huge LLM-generated AI system. However, if you do not manage copyrights well, you will face copyright issues when information that is supposed to be in a closed system goes out to an outside LLM, so you need to be very careful.

Furthermore, NTT and Sakana AI are working together to create a new AI. It is called MoE (Mixture of Experts). Each company has training data in its area of expertise, and the MoE, a coalition of SLM-generated AI, is an attempt to build and operate a highly capable generative AI through the cooperation of companies and research institutions that do not have the financial and human resources to gather such resources.

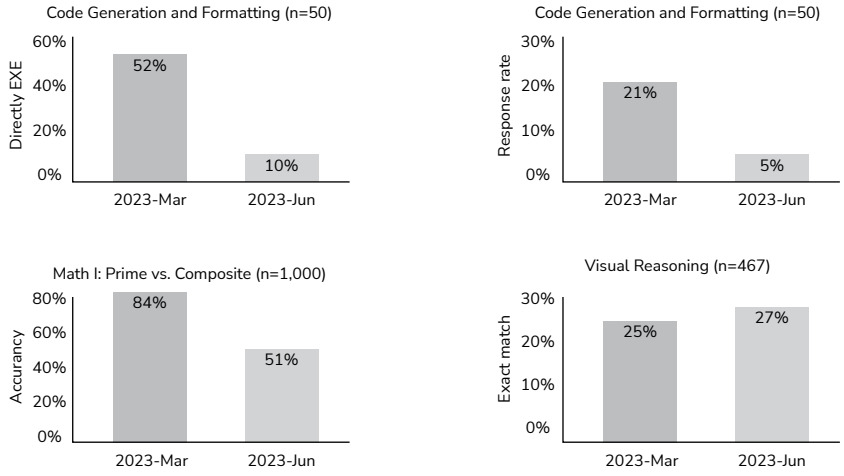
A. GENERATION AI CAPABILITIES

Wall Street Journal editors and Wall Street Journal columnists in the U.S. have compared the document production capabilities of high-profile generative AI (5).

For health information, ChatGPT is superior, followed by Google's Gemini, then Perplexity. For financial information, Gemini is the best, followed by Google's Gemini, then Perplexity. For food menus and cooking, ChatGPT is the best, and Claude, an LLM-generating AI released by Anthropic, is the best at writing work-related documents. Microsoft's Copilot is the best at creative writing, Perplexity is the best at summarizing, and Perplexity is the best at current events. In terms of coding, Perplexity was rated the highest, and ChatGPT was rated the fastest in terms of response time. Finally, from the user's perspective, the overall evaluation was Perplexity, ChatGPT, Gemini, Claude, and Copilot, in that order.

The next evaluation indicator is an interesting one (see Fig. 4).

FIG.4 Output change of GPT4.
Changes in LLM Performance Over Time



SOURCE: Lingjiao Chen, Matei Zaharia, James Zou, How is ChatGPT's Behavior Changing Over Time?

The research presented here is the work of researchers at Stanford University and UC Berkeley. It has been shown that the performance of certain commonly available LLMs can vary significantly over time.

For example, the June 2023 version of the GPT-4 was 42 percentage points lower than the March 2023 version in code generation, from 52% to 10% (n=50); in answering subtle questions, from 21% to 5%, 16 percentage points lower (n=100); in mathematical questions, from 84% to 51%, 33 points worse (n=1000). In visual reasoning, they improved by 2 points, from 25% to 27% (n=467).

The reason for this change is not clear, but may be a complex combination of various factors, including prompts, vector databases, feature settings, and token processing (6).

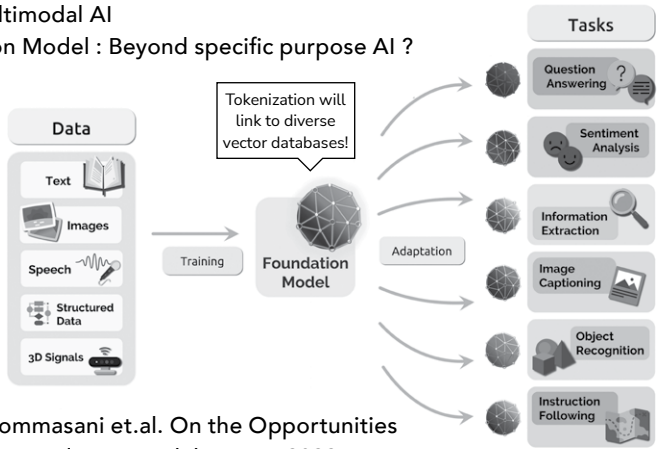
B. EVOLUTION OF GENERATION CAUSING MAJOR SOCIAL CHANGE

What has happened so far is called the democratization of AI or the popularization of AI. Until now, if you wanted to use AI, you had to learn advanced skills, the very thing you needed to be an expert in. However, with the advent of generative AI, anyone can write prompts in ordinary everyday language and use artificial intelligence to extract advanced knowledge. Hence, it is called the democratization of AI.

In the future, advanced generative AI is likely to bring enormous changes to society. And in the near future AGI (general-purpose AI) could be born.

There is a study that is highly significant in the perspective of AGI. Researchers at Stanford University in the U.S. have written a paper that describes multimodal generative AI as a foundation model. This AI is an artificial intelligence that can handle various forms of data, including language data (semi-structured data), voice data, images, videos, and structured data (numeric data) (see Fig. 5) (7).

FIG.5 Multimodal AI
Foundation Model : Beyond specific purpose AI ?



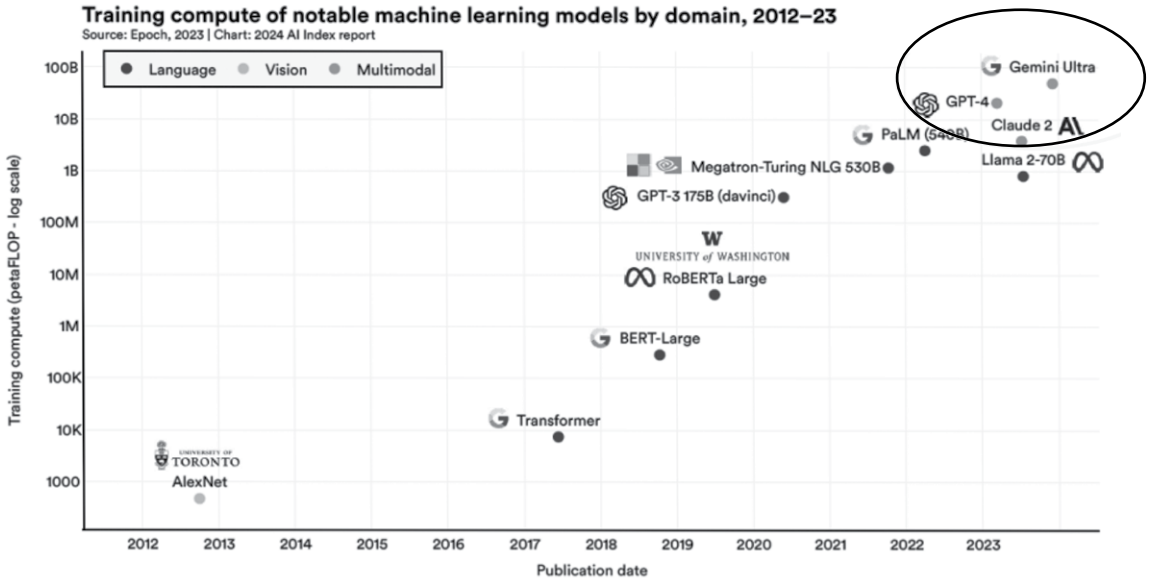
A foundation model can centralize the information from all the data from various modalities. downstream tasks.

SOURCE: Rishi Bommasani et.al. On the Opportunities and Risks of Foundation Models, arXiv, 2022

In October 2023, then US President Biden emphasized FM, or foundation model, as the most important AI.

Running a multimodal AI requires a high-performance supercomputer; see Fig. 6. If you look at this chart, the horizontal is a time series and the vertical is a scale of computation speed per second, with the highest value being petaflops x 100 billion, and the graph and scale are in logarithmic notation. The light blue color is the computational power required for multimodal AI. Compared to this, the power of ordinary LLM-generating AI, which is a hot topic in Japan, is at a much lower level.

FIG.6 Required Computational Power for Multimodal AI.

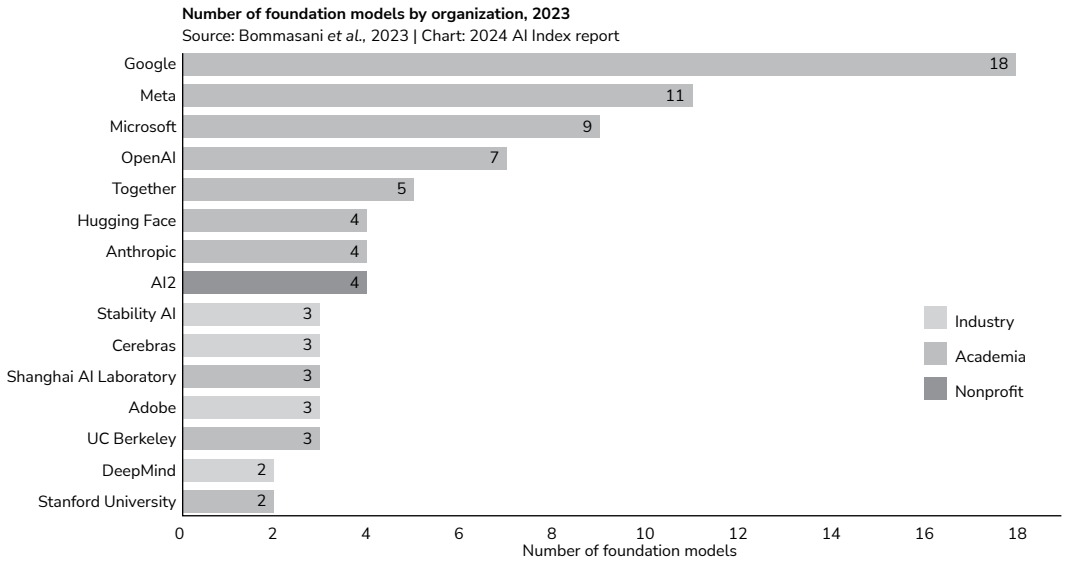


SOURCE: Stanford University Human Centered AI Institute ed. Artificial Intelligence Index Report 2024, p.25

The computational power required for multimodal AI, on which Google, Open AI Inc. and Anthropic are focusing their efforts, is in the order of awesome muscle; as of February 2025, Japanese companies and research institutions will be at a level that cannot be followed.

See Exhibit 7a. Institutions developing multimodal AI include Google, Meta, Microsoft, OpenAI, Together, Hugging Face, Anthropic, Shanghai AI Laboratory, UC Berkeley, and Stanford University.

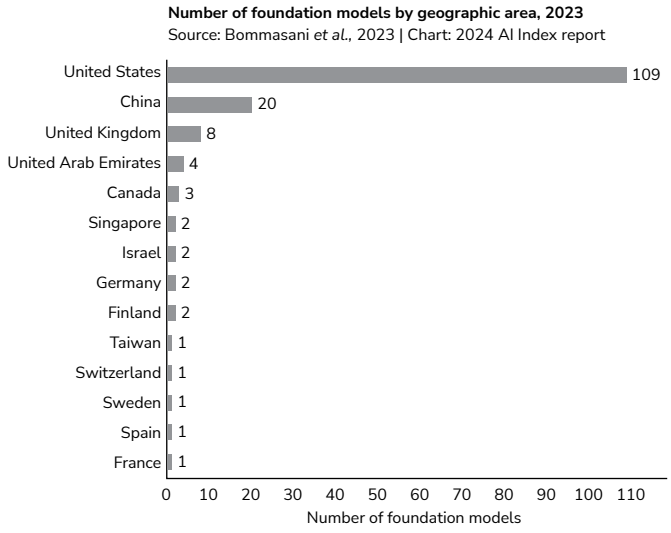
FIG.7A Development status of multimodal AI.



SOURCE: Stanford University Human Centered AI Institute ed. Artificial Intelligence Index Report 2024, p.33

And by region, where are the strongest; see Fig. 7b. The United States is by far the strongest. China is next, followed by the United Kingdom and the United Arab Emirates. Canada, Singapore, Israel, Germany, Finland, Taiwan, Switzerland, Sweden, Spain, and France follow, but Japan's name does not appear. In other words, we must say that is quite a backward country in multimodal AI R&D (8).

FIG.7B Comparison of Multimodal AI by Region.



SOURCE: Stanford University Human Centered AI Institute ed. Artificial Intelligence Index Report 2024, P35

C. POTENTIAL OF MULTIMODAL AI

Google's demonstration at GoogleI/O2024 featured a radical change in the field of education using Gemini 2.0, a multimodal AI system that is still in the prototype stage, but seeks to flexibly teach the "universal law of gravity" according to the academic level of students and pupils. Students and pupils of various academic levels can be accommodated, with high level classes for high level students, and college level studies for high school and junior high school students. For students who are not good at studying, they can include a game element and use games to increase concentration and interest in teaching. This would make them more responsive than human teachers, so it is easy to imagine that quite drastic changes could occur in the field of education. There is a growing possibility that a paradigm shift could occur not only in education, but also in healthcare, business organizations, government agencies, and the social system as a whole.

In the United States, various benchmarks for measuring the logical reasoning ability of advanced generative AI have been developed and are being used in practice. MMMU is presented here as an example (see Fig. 8).

FIG. 8 AI evaluation benchmark MMMU
Measuring the General Reasoning Ability of Multimodal AI MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Sample MMMU questions
Source: Yue et al., 2023

Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly? **Options:** (A) Major third <*image 1*> (B) Diminished fifth <*image 2*> **(C) Minor seventh <*image 3*>** (D) Diminished sixth <*image 4*>	**Question:** ... The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82_ **Options:** (A) 0 (B) 0.2142 **(C) 0.3571** (D) 0.5	**Question:** <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. **Options:** **(A)** $\int_{0}^{1.5} [f(x) - g(x)]dx$ (B) $\int_{1.5}^{0} [g(x) - f(x)]dx$ (C) $\int_{0}^{2} [f(x) - g(x)]dx$ (D) $\int_{0}^{2} [g(x) - x(x)]dx$
Subject: Music; **Subfield:** Music; **Image Type:** Sheet Music; **Difficulty:** Medium	**Subject:** Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts; **Difficulty:** Medium	**Subject:** Math; **Subfield:** Calculus; **Image Type:** Mathematical Notations; **Difficulty:** Easy
Health & Medicine	**Humanities & Social Science**	**Tech & Engineering**
Question: Youe are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. Whats is the etiology of the finding in the left breast? **Options:** (A) Susceptibility artifact (B) Hematoma **(C) Fact necrosis** (D) Silicone granuloma	**Question:** In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image 1*> **Options:** (A) Opressor (B) Imperialist **(C) Savior** (D) Isolationist	**Question:** Find the VCE for the circuit shown in <*image 1*>, Neglect VBE **Answer: 3.75** **Explanation:** ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10V - 6.25 V = 3.75 V
Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.; **Difficulty: Hard**	**Subject:** History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons; **Difficulty:** Easy	**Subject:** Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams; **Difficulty: Hard**

SOURCE: Stanford University Human Centered AI Institute ed. Artificial Intelligence Index Report 2024, p.113

MMMU is an acronym for A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI MMMU is a measure of reasoning ability at the college level of acquisition, It is designed to measure overall competence in Art & Design, Business, Science, Health & Medicine, Humanity & Social Science, and Tech & Engineering. Various other benchmarks have been developed, but so far each has its own strengths and weaknesses and is still in the trial-and-error stage(9).

Entering the year 2024, the multimodal AI Gemini 2.0 was reportedly compared to human logical reasoning ability. According to the report, Gemini achieved a score of 90.0%, outperforming human experts in 30 of the 32 academic benchmarks used in LLM R&D (note that Open AI's O2 also reportedly surpassed human capabilities in the winter of 2024).

Remarkably, Alpha Fold3, a multimodal AI advancement by Google's DeepMind published in the scientific journal Nature in May 2024, had a huge impact on the state of medicine. It is now able to simulate protein 3D structures in a short period of time and predict hierarchical interactions and binding between DNA, mRNA, and ligands. It is highly anticipated that the detailed elucidation of the mechanisms of protein interactions and binding will revolutionize the research and development of drugs for the treatment of cancer and other diseases.

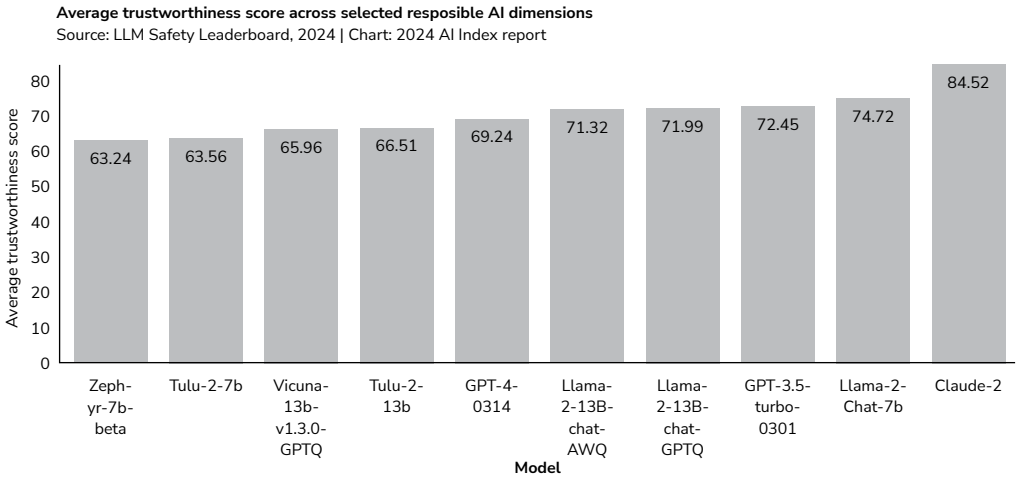
It is now clear that advanced generative AI is surpassing human logical reasoning ability. This is what Artificial General Intelligence (AGI) is all about. Masayoshi Son, founder of Softbank, has repeatedly emphasized that Japan must promote the development and social application of AGI.

D. IMPORTANT PERSPECTIVES IN ADVANCED AI GOVERNANCE

When AI talks about Vance, one of the most important concepts is Responsible AI (AI). This concept is at the core of AI governance. Its components are privacy and data governance, transparency and accountability (in short, as much as possible, transparency and accountability must be constantly pursued, since black boxes and incomprehensible things are processed and output in artificial intelligence), and security and safety, The key is security and safety, as well as fairness. Therefore, studies that attempt to measure these elements are well researched internationally. However, while institutional studies related to AI are often discussed in Japanese academic societies and government conferences, studies on meaningful measurement methods and benchmarks have unfortunately not been conducted much. In this sense, research and studies on AI governance are not at the international level.

So, let me mention a few examples of AI-in-vance quantitative studies: there is a benchmark called Decoding Trust, which is a quantitative check on how reliable a generative AI system is (see Fig. 10). According to it, GPT4 has a clear tendency to generate biased output and leak personal information from the training data set. It is also apparent that GPT4 is more susceptible to misleading prompts due to its improved data fidelity (10).

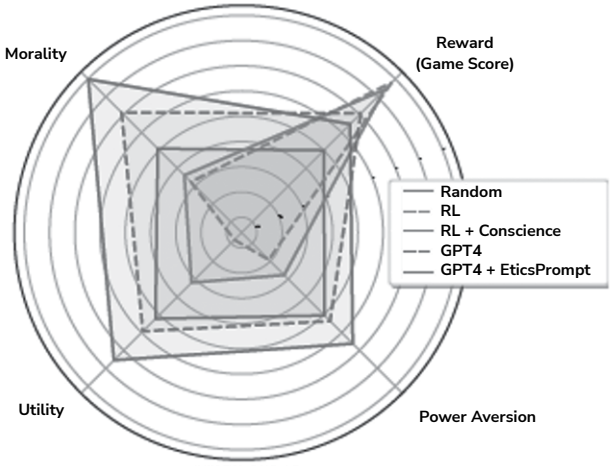
FIG.10 Overall Reliability of AI: Benchmark named Decoding Trust



SOURCE: LLM Safety Leaderboard

We should also mention a simulation study that is important for AI governance. It is a game-theoretic approach to measuring fairness called the "Machiavelli Benchmark" (see Fig. 11).

FIG.11 Machiavelli Benchmark: Fairness Measurement



SOURCE: Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside Jonathan Ng, Hanlin Zhang, Scott Emmons, Dan Hendrycks, Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark, arXiv, 16 July 2023

The model is a behavioral assessment benchmark for AI agents, creating a collection of 134 choice adventure games and building over 500,000 diverse social decision-making scenarios.

The goal is to measure and evaluate the extent to which AI agents pursue power, engage in deception, induce adversity, and commit ethical violations. Throughout the study, the authors found that the model faces a tradeoff between gain maximization and ethical decisions. However, they also reveal that there are decision-making patterns that balance the pursuit of gain and adherence to ethics in order to achieve objectives. The Machiavelli Benchmark may be evaluated as a highly useful study that can measure the fairness of multimodal AI (11).

By the way, as you know, Niccolò Machiavelli (May 3, 1469 - June 21, 1527) was an Italian Renaissance diplomat. Machiavellism, the concept named after him, is the idea that any means or unethical behavior should be socially acceptable if it results in the advancement of the interests of the state or organization. Although there is often a trade-off between fairness and the promotion of economic interests, depending on the options and the way resources are invested, they may form a socially rational choice. It is a measurement benchmark for AIs that make such flexible rational choices. Simulation studies on fairness of AI that effectively use such an agent model are also useful in considering how AI should be used in society, but unfortunately, to my knowledge, I have not seen such studies in Japanese academic societies or government review meetings.

III. HIROSHIMA AI PROCESS

In Japan, research and development of generative AI is not advanced by any means, but it is already widely used by many companies and universities. However, there are many things that need to be kept in mind now and in the future. Leakage of confidential information, inappropriate use of personal information, protection of intellectual property rights, and criminal sophistication must be addressed. Artificial intelligence has developed considerable wisdom, so we are likely to see more sophisticated crimes for which evidence is not easy to obtain. Disinformation may wreak havoc on society. In particular, we need to think carefully about the negative impact on elections, which are the institutional foundation of democracy. There will be a phase in which cyber attacks will become more sophisticated and the battle will be artificial intelligence vs.

Therefore, I would like to introduce the Hiroshima AI Process, which will define Japan's future AI-related efforts.

In 2023, the G7, chaired by the Japanese government, announced the "Hiroshima AI Process" 12 items for generating AI governance (see Fig. 12). In parallel, international organizations and major countries are developing specific initiatives to develop this framework in the OECD, the EU, and the UN.

FIG.11 G7 AI Hiroshima Process 12 items

Take appropriate measures throughout the development of advanced AI systems, including prior to and throughout their deployment and placement on the market, to identify, evaluate, and mitigate risks across the AI lifecycle.

Identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment including placement on the market. Identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment including placement on the market.

Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring 3. Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increase accountability.

Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia.

Develop, implement and disclose AI governance and risk management policies, grounded in a risk-based approach - including privacy 5. Develop, implement and disclose AI governance and risk management policies, grounded in a risk-based approach - including privacy policies, and mitigation measures, in particular for organizations developing advanced AI systems.

Invest in and implement robust security controls, including physical security, cybersecurity and insider threat safeguards across the AI lifecycle.

Develop and deploy reliable content authentication and provenance mechanisms, where technically feasible, such as watermarking or other techniques to enable users to identify AI-generated content.

Prioritize research to mitigate societal, safety and security risks and prioritize investment in effective mitigation measures.

Prioritize the development of advanced AI systems to address the world's greatest challenges, notably but not limited to the climate crisis, global health and education. Prioritize the development of advanced AI systems to address the world's greatest challenges, notably but not limited to the climate crisis, global health and education.

Advance the development of and, where appropriate, adoption of international technical standards.

Implement appropriate data input measures and protections for personal data and intellectual property.

Encourage and contribute to safe and responsible use of AI.

SOURCE: Japan Government, The Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems, Dec. 2023

At the 2023 G7, the Japanese government, as chair, issued the Hiroshima Process International Guidelines for All AI Stakeholders. We emphasize the responsibility of all AI stakeholders in promoting safe, secure, and reliable AI in an appropriate and relevant manner. We recognize that stakeholders across the

lifecycle have different responsibilities and different needs regarding the safety, security, and reliability of AI. And all 12 principles of the Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems "require that all AI stakeholders be given the opportunity to participate in the design, development, implementation, provision, and use of advanced AI systems in a timely manner and to the appropriate extent, recognizing that some elements may only be applicable to organizations developing advanced AI systems.. should be applied in an appropriate and proportionate manner and to an appropriate extent" (12).

In particular, mention should be made of items that will become important in the future. Note the seventh item in Figure 11. How do we manage copyrights? For images, watermarking is already available; Microsoft's Copilot, GPT-4, and Gemini all support it. However, as of spring 2024, reliable content authentication and illumination of text was not possible. However, research is making progress on the use of SynthID in the fall of 2024 (13).

Look at item 9: The G7 expects AI to contribute to three events: contribution to climate change issues, medicine and health care, and education.

In Item 10, technical standardization will be necessary to ensure international compatibility in the future, and active efforts will be required by internationally trusted organizations with high evaluation and auditing capabilities, such as ISO, NIST, IEEE (Institute of Electrical Engineers of America), and OECD. These organizations are collectively referred to as SDOs (Standards Development Organizations), but it will also be important to establish a relationship between SDOs and the organizations that determine technical standards in each country.

In Item 11, we must envision how to create an international technical audit mechanism, and a national technical audit mechanism linked to it. Generating AI will have a great influence in the future, and technical and activity audits will be very important.

Based on the 11 items of the "Hiroshima Process International Guideline for Organizations Developing Advanced AI Systems" published in October 2023, the final version of the guideline in December 2023 will include the following items to be included in the final version, which are based on the concept that each party involved in the entire AI life cycle has different responsibilities to realize safe, secure, and reliable AI. The 12th item was added to the final version of the Hiroshima Process International Guidelines for Organizations Developing Advanced AI Systems in December, based on the 11 items of the Hiroshima Process International Guidelines for Organizations Developing Advanced AI Systems.

In the process of this study, President Biden of the United States of America issued a Presidential Decree on AI in October 2023 that will have a significant impact on the world. Therefore, it is important to mention the AI Presidential Decree by President Biden at.

IV. BIDEN HARRIS PRESIDENTIAL DECREE

On October 30, 2023, the United States issued a presidential decree as an AI policy to complement the Hiroshima AI process in Japan. At the time, this was an important AI policy with global influence, and although it is long, it is quoted here (14).

"Today, President Biden is issuing a landmark Executive Order to ensure that America leads the way in seizing the promise and managing the The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil liberties. The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.

As part of the Biden-Harris Administration's comprehensive strategy for responsible innovation, the Executive Order builds on previous As part the Biden-Harris Administration's comprehensive strategy for responsible innovation, the Executive Order builds on previous actions the President has taken, including work that led to voluntary commitments from 15 leading companies to drive safe, secure, and trustworthy The Executive Order builds on previous actions the President has taken, including work that led to voluntary commitments from 15 leading companies to drive safe, secure, and trustworthy development of AI.

THE EXECUTIVE ORDER DIRECTS THE FOLLOWING ACTIONS: 1.

New Standards for AI Safety and Security

With this Executive Order, the President directs the most sweeping actions ever taken to protect Americans from the potential risks of AI systems: AI's capabilities grow, so do its implications for Americans' safety and security. With this Executive Order, the President directs the most sweeping actions ever taken to protect Americans from the potential risks of AI systems: the Require

that developers of the most powerful AI systems share their safety test results and other critical information with the U.S. government. In accordance with the Defense Production Act, the Order will require that companies developing any foundation model that poses a serious risk to national security, national economic security, or national public health and safety must notify the federal government when training the model. In accordance with the Defense Production Act, the Order will require that companies developing any foundation model that poses a serious risk to national security, national economic security, or national public health and safety must notify the federal government when training the model, and must share the results of all red-team members. These measures will ensure AI systems are safe, secure, and trustworthy before companies make them public.

Develop standards, tools, and tests to help ensure that AI systems are safe, secure, and trustworthy. The National Institute of Standards and Technology will set the rigorous standards for extensive red-team testing to ensure safety before public release. The Department of Homeland Security will apply those standards to critical infrastructure sectors and establish the AI Safety and Security Board. The Departments of Energy and Homeland Security will also address AI systems' threats to critical infrastructure, as well as chemical, biological, radiological, nuclear, and cybersecurity risks. Together, these are the most significant actions ever taken by any government to advance the field of AI safety.

Protect against the risks of using AI to engineer dangerous biological materials by developing strong new standards for biological synthesis screening Agencies that fund life-science projects will establish these standards as a condition of federal funding, creating powerful incentives to ensure Agencies that fund life-science projects will establish these standards as a condition of federal funding, creating powerful incentives to ensure appropriate screening and manage risks potentially made worse by AI.

Protect Americans from AI-enabled fraud and deception by establishing standards and best practices for detecting AI-generated content and The Department of Commerce will develop guidance for content authentication and watermarking to clearly label AI- generated content. Federal agencies will use these tools to make it easy for Americans to know that the communications they receive from their government are authentic-and set an example for the AI- generated content. Federal agencies will use these tools to make it easy for Americans to know that the communications they receive from their government are authentic-and set an example for the private sector and governments around the world.

Establish an advanced cybersecurity program to develop AI tools to find and fix vulnerabilities in critical software, building on the Biden-Harris Together,

these efforts will harness AI's potentially game-changing cyber capabilities to make software and networks more secure.

Order the development of a National Security Memorandum that directs further actions on AI and security, to be developed by the National Security Council and White House Chief of Staff. This document will ensure that the United States military and intelligence community use AI safely, ethically, and This document will ensure that the United States military and intelligence community use AI safely, ethically, and effectively in their missions, and will direct actions to counter adversaries' military use of AI.

Protecting Americans' Privacy

Without safeguards, AI can put Americans' privacy further at risk. AI not only makes it easier to extract, identify, and exploit personal data, but it also heightens incentives to do so because companies use data to train AI systems. AI not only makes it easier to extract, identify, and exploit personal data, but it also heightens incentives to do so because companies use data to train AI systems. To better protect Americans' privacy, including from the risks posed by AI, the President calls on Congress to pass bipartisan data privacy legislation to protect all Americans, especially To better protect Americans' privacy including from the risks posed by AI, the President calls on Congress to pass bipartisan data privacy legislation to protect all Americans, especially kids, and directs the following actions: 1.

Protect Americans' privacy by prioritizing federal support for accelerating the development and use of privacy-preserving techniques -including ones that use cutting-edge AI and that let AI systems be trained while preserving the privacy of the training data.

Strengthen privacy-preserving research and technologies, such as cryptographic tools that preserve individuals' privacy, by funding a The National Science Foundation will also work with this network to promote the adoption of leading-edge privacy-preserving technologies by federal agencies. The National Science Foundation will also work with this network to promote the adoption of leading-edge privacy-preserving technologies by federal agencies.

Evaluate how agencies collect and use commercially available information-including information they procure from data brokers -This work will focus in particular on commercially This work will focus in particular on commercially available information containing personally identifiable data.

Develop guidelines for federal agencies to evaluate the effectiveness of privacy-preserving techniques, including those used in AI systems. guidelines will advance agency efforts to protect Americans' data.

Advancing Equity and Civil Rights

Irresponsible uses of AI can lead to and deepen discrimination, bias, and other abuses in justice, healthcare, and housing. Administration has already taken action by publishing the Blueprint for an AI Bill of Rights and issuing an Executive Order directing agencies to combat To ensure that AI advances equity and civil rights, the President directing agencies to combat AI advances and civil rights, the President directing agencies to combat AI advances and civil rights, the President directing agencies to combat AI advances and civil rights. To ensure that AI advances equity and civil rights, the President directs the following additional actions: To ensure that AI advances equity and civil rights, the President directs the following additional actions

Provide clear guidance to landlords, Federal benefits programs, and federal contractors to keep AI algorithms from being used to exacerbate discrimination.

Address algorithmic discrimination through training, technical assistance, and coordination between the Department of Justice and Federal civil rights offices on best practices for investigating and prosecuting civil rights violations related to AI.

Ensure fairness throughout the criminal justice system by developing best practices on the use of AI in sentencing, parole and probation, pretrial release and detention, risk assessments, surveillance, crime forecasting and predictive policing, and forensic analysis.

Standing Up for Consumers, Patients, and Students

AI can bring real benefits to consumers-for example, by making products better, cheaper, and more widely available. To protect consumers while ensuring that AI can make Americans better off, the President directs the following actions: 1. To protect consumers while ensuring that AI can make Americans better off, the President directs the following actions: 1.

Advance the responsible use of AI in healthcare and the development of affordable and life-saving drugs. The Department of Health and Human Services will also establish a safety program to receive reports of-and act to remedy - harms or unsafe healthcare practices involving AI.

Shape AI's potential to transform education by creating resources to support educators deploying AI-enabled educational tools, such as Shape AI's potential to transform education by creating resources to support educators deploying AI-enabled educational tools, such as

Supporting Workers

AI is changing America's jobs and workplaces, offering both the promise of improved productivity but also the dangers of increased To mitigate these risks, support workers' ability to bargain collectively, and To mitigate these risks, support workers' ability to bargain collectively, and invest in workforce training and development that is accessible to all, the President directs the following actions: To reduce the number of workers in the labor market, support workers' ability to bargain collectively, and

Develop principles and best practices to mitigate the harms and maximize the benefits of AI for workers by addressing job displacement; labor standards; labor standards; and data collection. These principles and best practices will benefit workers by providing guidance to prevent These principles and best practices will benefit workers by providing guidance to prevent employers from undercompensating workers, evaluating job applications unfairly, or impinging on workers' ability to organize.

Produce a report on AI's potential labor-market impacts, and study and identify options for strengthening federal support for workers Produce a report on AI's potential labor-market impacts, and study and identify options for strengthening federal support for workers facing labor disruptions, including from AI.

Promoting Innovation and Competition

America already leads in AI innovation-more AI startups raised first-time capital in the United States last year than in the next seven The Executive Order ensures that we continue to lead the way in innovation and competition through the following actions: a. Increase the number of AI startups in the United States

Catalyze AI research across the United States through a pilot of the National AI Research Resource-a tool that will provide AI researchers and students access to key AI resources and data-and expanded grants for AI research in vital areas like healthcare and climate change.

Promote a fair, open, and competitive AI ecosystem by providing small developers and entrepreneurs access to technical assistance and resources, helping small businesses commercialize AI breakthroughs, and encouraging the Federal Trade Commission to exercise its authorities. Promote a fair, open, and competitive AI ecosystem by providing small developers and entrepreneurs access to technical assistance and resources, helping small businesses commercialize AI breakthroughs, and encouraging the Federal Trade Commission to exercise its authorities.

Use existing authorities to expand the ability of highly skilled immigrants and nonimmigrants with expertise in critical areas to study, stay, and work Use existing authorities to expand the ability of highly skilled immigrants and nonimmigrants with expertise in critical areas to study, stay, and work.

Advancing American Leadership Abroad

The Biden-Harris Administration will continue working with other nations to support safe, secure, and trustworthy deployment and use of AI worldwide. To that end, the President directs the following actions: To support safe, secure, and trustworthy deployment and use of AI worldwide.

The State Department, in collaboration, with the Commerce Department will lead an effort to establish robust international frameworks for harnessing AI's benefits and managing its risks and ensuring safety. In addition, this week, Vice President Harris will speak at the UK Summit on AI Safety, hosted by Prime Minister Rishi Sunak.

Accelerate development and implementation of vital AI standards with international partners and in standards organizations, ensuring that the Accelerate development and implementation of vital AI standards with international partners and in standards organizations, ensuring that the technology is safe, secure, trustworthy, and interoperable.

Promote the safe, responsible, and rights-affirming development and deployment of AI abroad to solve global challenges, such as advancing sustainable development and mitigating dangers to critical infrastructure.

Ensuring Responsible and Effective Government Use of AI

It can expand agencies' capacity to regulate, govern, and disburse benefits, and it can cut costs and enhance the security of government systems. However, use of AI can pose risks, such as discrimination and unsafe decisions. To ensure the responsible government deployment of AI and modernize federal AI infrastructure, the President directs the following actions The President directs the following actions: The President directs the following actions

Issue guidance for agencies' use of AI, including clear standards to protect rights and safety, improve AI procurement, and strengthen AI deployment.

Help agencies acquire specified AI products and services faster, more cheaply, and more effectively through more rapid and efficient contracting.

Accelerate the rapid hiring of AI professionals as part of a government-wide AI talent surge led by the Office of Personnel Management, U.S. Digital Service, U.S. Digital Corps, and Presidential Innovation Fellowship. Agencies will provide AI training for employees at all levels in relevant fields.

As we advance this agenda at home, the Administration will work with allies and partners abroad on a strong international framework to govern the The Administration has already consulted widely on AI governance frameworks over the past several months-engaging with Australia, Brazil, Canada, Chile, the European Union, France, Germany, India, Israel, Italy, Italy, Chile, India, Israel, and the United States. The Administration has already consulted widely on AI governance frameworks over the past several months- engaging with Australia, Brazil, Canada, Chile, the European Union, France, Germany, India, Israel, Italy, Japan, Kenya, Mexico, the Netherlands, New The actions taken today support and complement Japan's The actions taken today support and complement Japan's leadership of the G-7 Hiroshima Process, the UK Summit on AI Safety, India's leadership as Chair of the Global Partnership on AI, and ongoing discussions at the United Nations".

The actions that President Biden directed today are vital steps forward in the U.S.'s approach on safe, secure, and trustworthy AI. More action will be required, and the Administration will continue to work with Congress to pursue bipartisan legislation to help America lead the way in responsible innovation. more action will be required, and the Administration will continue to work with Congress to pursue bipartisan legislation to help America lead the way in responsible innovation.

Biden= The significance of the Harris Executive Order is that it sought to advance this AI agenda within and beyond the United States and to create a strong international framework to manage the development and use of AI. He sought to collaborate with Australia, Brazil, Canada, Chile, the European Union, France, Germany, India, Israel, Italy, Japan, Kenya, Mexico, the Netherlands, New Zealand, Nigeria, the Philippines, Singapore, South Korea, the United Arab Emirates, the United Kingdom, and others to create a G7 Hiroshima Japan's leadership in the process, the UK Summit on AI Safety, India's leadership as Chair of the Global Partnership on AI (GPAI Summit 2023), and complemented discussions on AI at the UN.

In particular, it should be noted that it was influential not only domestically, but also for the Japanese government and industry.

First, it was important to present new standards for AI security. It requires developers of the most powerful AI systems to share safety test results and other critical information with the U.S. government, and in accordance with the Defense Production Act, companies developing foundation models that pose a significant risk to national security, national economic security, or the health and safety of the public (Foundation Model), and that companies developing such

models must notify the federal government when training them and must share the results of all Red Team safety testing.

As part of this effort, it will protect Americans from AI-based fraud and deception by establishing standards and best practices for detecting AI-generated content and authenticating public content. Directs the Department of Commerce to develop guidance on content authentication and watermarking to clearly label AI-generated content.

The second point is the clarification of policy on AI and privacy. In short, it makes clear that it will protect Americans' privacy by prioritizing federal support for the use of cutting-edge AI and accelerating the development and use of privacy-preserving technologies.

It will enhance privacy-preserving research and technology, including cryptographic tools that protect personal privacy. The National Science Foundation will also work with the network to promote the adoption of cutting-edge privacy-preserving technologies by federal agencies.

Third, it issued a strong directive on Advancing Equity and Civil Rights to federal agencies.

Irresponsible use of AI can lead to and deepen discrimination, bias, and other abuses in justice, health care, and housing. The Biden-Harris Administration has already taken action by releasing the Blueprint for an AI Bill of Rights and issuing a Presidential Executive Order directing agencies to combat discrimination by a algorithm while enforcing existing authorities to protect people's rights and safety, but AI can also be used to improve equity and civil rights, the President has directed additional, more g-aggressive actions to federal agencies to ensure that AI improves fairness and civil rights.

Fourth, it seeks to further advanced AI innovation and international competition. The U.S. already leads the world significantly in AI innovation, but this Executive Order seeks to keep the U.S. leading in innovation and competition by taking the following actions

It seeks to promote AI research more strongly through piloting the National AI Research Resource, a tool that provides AI researchers and students with access to key AI resources and data, and by expanding funding for AI research in critical areas such as healthcare and climate change. The National AI Research Resource is a tool that provides access to key AI resources and data for students and researchers.

It seeks to provide technical assistance and access to resources to small developers and entrepreneurs, help small companies commercialize AI breakthroughs, and promote a fair, open, and competitive AI ecosystem.

Although these directives are not comprehensive AI legislation like the EU, nor are they soft-law thinking with an emphasis on guidelines like the traditional Japanese approach, they can be evaluated as an attempt to secure the Hiroshima AI process by making competition and cooperation with the G7, EU, UK, India, Japan, and others more concrete. However, when President Donald Trump took office on January 20, 2025, he immediately scrapped Biden-Harris' Executive Order on Advanced Generation AI. It is not clear at this point (February 2025) what kind of advanced AI strategy the U.S. will implement in the future, but it is easy to imagine that the U.S. will considerably ease regulations to further promote innovation and strengthen the international competitiveness of U.S. AI companies.

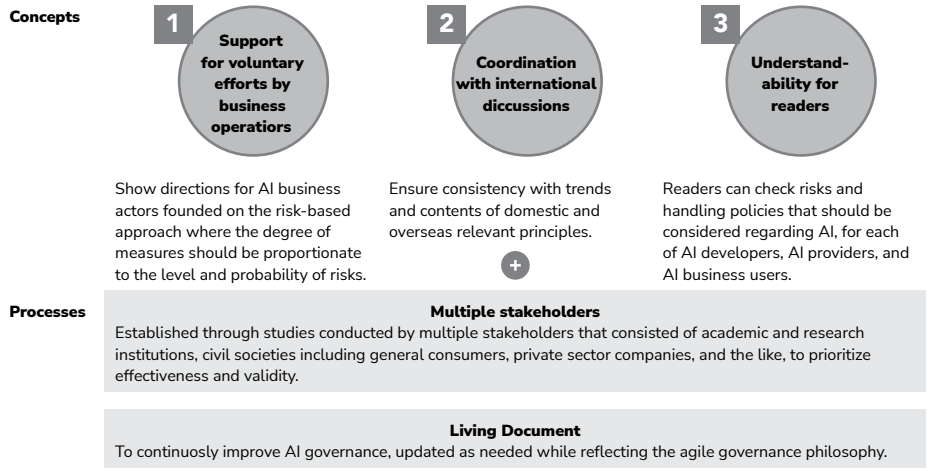
At any rate, 2023 was a year of further international cooperation on the governance of generative AI, with Japan's efforts as chair of the G7.

Next, it will be important for Japan to find a way to get Japanese society to understand and comply with the 12 items of the "Hiroshima AI Process. Let us look at the efforts to achieve this goal (15).

V. AI USER GUIDELINES BY JAPAN GOVERNMENT

In Japan, it was decided to create and publish soft-law "Guidelines for AI Operators" that abolish the rigidity of rules and respect the autonomy of operators as much as possible. The basic framework is described below (see Fig. 12)(16).

FIG. 12 Basic Concepts of the Guidelines



SOURCE: AI Guidelines for Business Ver. 1.01, December 25, 2024. Ministry of Internal Affairs and Communications Ministry of Economy, Trade and Industry

These AI Business Operator Guidelines, in cooperation with the MIC Guidelines Study Council and the METI Guidelines Study Council, provide unified guidelines for AI governance in Japan to address risks associated with the expanded use of generative AI technologies and to promote safe and secure use of generative AI. In light of international trends and stakeholder concerns, the objective is to promote the safe and secure use of AI by business operators so that they can correctly identify AI risks and voluntarily implement necessary countermeasures throughout the entire AI lifecycle.

AI technology is developing day by day, contributing to the creation of innovation in industry and the solution of social issues. In particular, the rise of generative AI has promoted the "democratization of AI" and made it easier for many people to use AI. At the same time, however, new social risks have arisen, such as infringement of intellectual property rights and generation and dissemination of false or misinformation. It is against this background that the Hiroshima AI Process is being promoted.

These Guidelines provide uniform guidelines for AI governance in Japan in order to promote the safe and secure use of AI. This will help those who utilize AI to correctly recognize AI risks in light of international trends and stakeholder concerns, and to voluntarily implement necessary countermeasures throughout the AI lifecycle.

The use of AI can create significant risks to society, depending on the field and form of use. However, excessive countermeasures may hinder AI utilization itself and its benefits. Therefore, these guidelines adopt a "risk-based approach", in which the degree of countermeasures is adjusted according to the magnitude of the risk.

In addition, amidst the rapidly changing trends surrounding AI, these guidelines are planned to be updated as a Living Document from time to time with multi-stakeholder involvement, with reference to the concept of agile governance, in order to continuously improve AI governance.

The Guidelines first define three broad categories of entities responsible for AI business activities: "AI developers", "AI providers", and "AI users", and then define each of these categories.

1. AI developers: Businesses that develop AI systems (including those that research and develop AI)
2. AI provider: An entity that provides AI systems to AI users as services embedded in applications, products, existing systems, business processes, etc.
3. AI Users: Businesses that use AI systems or AI services in their business activities.

This categorization helps to clarify the role and guidelines to keep in mind during the generative AI lifecycle.

Structure of the Guidelines
These guidelines consist of the following
Part 1: What is AI (Definition of Terms)
Part 2: The society that should be aimed at by AI, and the issues to be tackled by each entity
A. Basic Philosophy
B. Principles
C. Common Guidelines
D. Common guidelines for businesses involved in advanced AI systems
E. Building AI Governance
Part 3: AI Developer Matters
Part 4: AI Provider Matters
Part 5: Matters Related to AI Users

Part 2 in particular has some important statements, so I will give a few details.
A. Basic Philosophy: 1.
Based on the Principles for a Human-Centered AI Society proposed by the Japanese government in March 2019, we aim to create a society that respects the following three values

A society where human dignity is respected
A society where people of diverse backgrounds can pursue diverse happiness.
Sustainable Society

B. Principles
In order to realize the basic principles, the items to be addressed by each entity and the items that are expected to be addressed in cooperation with society are summarized. It is important for each entity to promote the development, provision, and use of AI systems and services, and to realize the objectives of AI, such as creating value in business and solving social issues, while protecting human dignity.
C. Common Guidelines
When each entity develops, provides, and uses AI systems and services,

We believe that people-centeredness, safety, fairness, privacy protection, security assurance, transparency, and accountability should be respected.

D. Common guidelines for businesses involved in advanced AI systems
It states that entities involved in advanced AI systems should comply with the international guidelines developed through the Hiroshima AI Process, in addition to the common guidelines, as well as the additional items for each entity indicated in these guidelines.

E. AI Governance
It believes that the early establishment of AI governance is important to manage AI-related risks at a level acceptable to stakeholders while maximizing the benefits that come from it. Agile governance practices are particularly important.

Let us now consider the significance of this guideline and its remaining significance. The Guidelines are considered to be of great significance in the following respects

First, it can be highly evaluated as a comprehensive AI governance guideline. Until now, the Ministry of Internal Affairs and Communications (MIC), the Ministry of Economy, Trade and Industry (METI), and other ministries have developed individual guidelines for AI, but these guidelines integrate and review these guidelines and present them as a comprehensive AI governance guideline.

Next, the adoption of a risk-based approach similar to that of the EU is highly commendable from the viewpoint of effectiveness: the risks of AI vary greatly depending on the field and form of its use. The Guidelines adopt a "risk-based approach" that adjusts the degree of countermeasures according to the magnitude of risk, which is expected to lead to highly effective countermeasures. And while it encourages the adoption of agile governance, AI technology is evolving on a daily basis and cannot be handled with fixed rules. By introducing the concept of agile governance and encouraging continuous improvement, these guidelines aim to build governance that can flexibly respond to AI risks.

Third, the Guidelines advocate multi-stakeholder formulation, which is emphasized by the OECD and the EU. The Guidelines are highly effective and legitimate because they were formulated with the participation of a diverse range of stakeholders, including not only governments, but also educational and research institutions, civil society including general consumers, and private companies.

However, we must also point out the remaining challenges.

While these guidelines are an important step toward the safe and secure use of AI, the following issues remain to be addressed.

First, since these guidelines are non-legally binding soft law and are left to the voluntary efforts of operators, the issue is how to ensure compliance with the guidelines.

Second, agile governance is a concept to promote continuous improvement, but how exactly to put it into practice is left to the ingenuity of each business. Therefore, specific methodologies and tools will have to be provided to support the practice of agile governance.

Third, while these guidelines refer to advanced generative AI systems, they do not go far enough to address their risks and countermeasures. More detailed analysis of the benefits and risks of advanced generative AI systems, such as multimodal AI and Graph-RAG, and specific countermeasures would be needed.

Fourth, specifics on support measures for SMEs are needed. Although these guidelines are intended for all businesses, especially for SMEs, which make up the majority of companies in Japan, they will often not be able to devote sufficient resources to building AI governance. Support measures need to be enhanced to help SMEs establish AI governance.

Fifth, it must be said that technical explanations are lacking. The Guidelines focus mainly on organizational aspects and principles and lack technical details. For example, it must be said that descriptions of privacy protection techniques (differential privacy, secret calculations, etc.) and specific algorithms to ensure fairness are limited.

And as a sixth, the mechanism of auditing and evaluation must be mentioned. The mechanism for how compliance with the guidelines will be audited and evaluated is not clear. Transparency and credibility could be enhanced by considering the introduction of a third-party auditing and certification system, but this is an important issue for the future, and we believe that careful consideration is needed.

Although we have made issues over six points, we expect that overcoming the issues through version upgrades and continuous improvement of this guideline in the future will promote the safe and secure use of AI and contribute to the realization of a sustainable society. In this sense, we are trying to equip the guideline with the ability to respond to rapid changes in technology, and this is highly commendable.

In addition, disinformation and intellectual property rights may attract very important issues. In this regard, a basic law that can serve as the basis for an industry law and AI guidelines is considered necessary, and in 2024, the AI Strategy Council and AI System Study Group organized by the Cabinet Office began to study AI legislation. An interim report was released on February 4, 2025. Based on this report, the legislation is scheduled to be discussed at the ordinary session of the Diet in 2025. The following is an overview of the AI legislation in Japan.

VI. CONSIDERATION OF MORE ADVANCED AI GOVERNANCE AND AI LEGISLATION

The "Interim Summary" of the AI Strategy Council and AI System Study Group was prepared through discussions and public comments at the AI System Study Group, which has met a total of seven times since July 2024.While AI has the potential to make a significant contribution to Japan's development, the background of the various risks that have emerged underscores the need to ensure transparency and other aspects of AI The need to ensure appropriateness and to promote the development and utilization of AI is emphasized (17).

First, it has been shown that the performance of generative AI will improve dramatically after the fall of 2022, and that it has the potential to greatly improve efficiency and convenience in various areas of industry and people's lives. On the other hand, however, various risks are emerging, such as the increasing sophistication of crimes in which fake AI websites and synthetic voices are used for fraud, etc., and information manipulation through the use and spread of AI to create false and misleading information, etc. In light of trends in AI regulations overseas, including the EU AI Act and trends in federal and state governments in the United States The report suggests the need for institutional design of AI in Japan, taking into account the trends of AI regulations overseas, such as the EU AI Act and the federal and state governments in the U.S.

First, the following four points are fundamental to the system.

The importance of balancing the promotion of innovation and risk response. The report states that it is necessary to promote innovation through R&D support, human resource development, and the development of data and computational resources, as well as to address risks through an appropriate combination of laws, regulations, and guidelines. In addition, he stressed the importance of ensuring international consistency and interoperability in the formation of AI governance. The report also states that the system should be considered from the standpoints of three entities: AI developers, AI providers, and AI users, and that foreign operators should also be subject to the system.

A major feature of this interim report is the compatibility between the application of laws and regulations and the use of soft law. Further systemic consideration should be given to risks, based on the assumption that existing laws and regulations address certain risks to a certain extent. The report emphasizes the appropriate combination of laws and regulations and soft law such as guidelines.

Next, the following three points are listed as the direction of specific systems and measures.

First, the report emphasizes the strengthening of the government's command post function and the formulation of strategies. It states that the government's command post function, which overlooks the entire process, will be strengthened, and a strategy (basic plan) for the safe and secure research, development, and utilization of AI will be formulated.

Secondly, it describes the improvement of safety, etc. Development of guidelines by the government (in compliance with the Hiroshima AI Process), co-operation by operators, surveys and information gathering by the government, guidance and advice to operators and the public, provision of information, etc.

Third, the government will formulate guidelines to address AI-specific risks in government procurement and promote the use of AI to improve the quality and efficiency of government services and operations.

Response to AI related to securityThe ministries and agencies will respond to AI in areas that have a significant impact on people's lives and social activities, such as medical devices, self-driving cars, and infrastructure services, based on existing business laws, and will determine the need for additional responses.

Particularly important points of this "interim report" include: the Ministry will support the safe use of AI by promoting information sharing among AI developers, providers, and users; the introduction of a certification system for safety evaluation of AI systems and governance of organizations using AI; and the government will take the lead in investigating the cause of a major incident, i.e., a major incident caused by the use of AI, and formulating measures to prevent its recurrence. occurrence, i.e., in the event of a serious incident resulting from the use of AI, the government will take the lead in investigating the cause and formulating measures to prevent recurrence.

The "Interim Report" concludes that, in order to balance the promotion of innovation and response to risks, the government should develop guidelines for widely used AI, etc., and encourage business operators to take the initiative in ensuring transparency and appropriateness, while at the same time ensuring the safety and security of life and limb as a matter of course, At the same time, the report concludes that the government should investigate and understand the status of responses to various risks, including the actual development and use of AI, and if a serious problem arises or is likely to arise, the government should take actions based on existing laws and regulations and provide necessary support.

The significance of this interim report can be summarized in the following three points.

First, it provides a common ground for a comprehensive discussion of the ethical, legal, and social issues of AI for a comprehensive discussion of AI and suggests directions for future discussions.

Second, it aims to balance risk and innovation, providing a framework for effectively addressing potential risks without impeding the advancement of AI technology, and outlining a path for safe AI development and use.

Third, it emphasizes the importance of international collaboration. It indicates Japan's contribution to the formation of global AI governance and describes how it has established a platform to promote international collaboration.

We must also state a few requests for this "interim report".

There is always the possibility that excessive regulation will impede the progress of AI technology. We urge a strong sense of flexibility to promote innovation. Furthermore, the birth of the Trump administration is creating significant turbulence in international bridge relationships. With the different trends in AI regulations in different countries, we would like to see an environment that helps to ensure international consistency. While more detailed analysis and studies with specific examples are needed in the future, we believe that this interim report is certainly an important step forward in determining the direction of future AI policy.

VII. TOWARD A NEW VISION OF SOCIAL DEVELOPMENT

I believe that Japan needs to make more forward-looking efforts, both in the executive branch, industry, research institutions, and educational institutions. (18) OECD expert meetings have discussed Risks and Benefits, and AI Policy with regard to AI, including AGI. Effective governance of AI requires a complex but closely interrelated examination of data management and protection, consumer protection, digital platforms, the global environment, digital finance, medicine and healthcare, education and human resource development, finance, and many other factors. And independent third-party audits would require a detailed study of what authority they should have, what role they should play, and how they should be exercised.

It is fair to say that AI to date has been primarily optimized for specific purposes and has not been a threat to human creative or strategic acts. However, generative and multimodal AI is becoming more involved in creative and strategic actions.

Thus, the social deployment of generative and multimodal AI has the potential to bring about major changes in human conduct. It also has the potential to bring about major upheavals in existing social systems. For this reason, international conferences and international organizations such as the G7, UN, UNESCO, and OECD have been conducting fundamental studies on advanced AI. Although Japan's R&D achievements in advanced generative AI are not outstanding, it has a world-class track record in human-centered principles and innovation policies related to advanced generative AI, and it will have to build up and demonstrate its strength in R&D and human resource development in the future.

VIII. BIBLIOGRAPHY

Pan, Alexander, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, Dan Hendrycks, [2023]. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark, arXiv, 16 July 2023

AI Strategy Council and AI System Study Group ed.[2025] Interim Summary, Cabinet Office in Japan Goverment,4 Feb. 2025 (in Japanese).

Anzai, Yuichiro [2021] Learning and Interaction: From Cognitive Theories to Epistemology, Keio University Press

Bommasani, Rishi et.al. [2022] On the Opportunities and Risks of Foundation Models, arXiv:2108.07258v3

Heaven, Will Douglas [2023] ChatGPT is everywhere. Here's where it came from, MIT Technology Review, Feb. 8th 2023

Hinton, G. E. [2014] Where do features come from?, Cognitive Science, Vol. 38(6).

Japan Government [2023] The Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems, Dec. 2023

Japan AI Safety Institute (J-AISI) ed.[2025] National Status Report on AI Safety in Japan 2024, 5 Feb. 2025

Kai-Fu, Lee [2018] AI Superpowers: China, Silicon Valley, and the New World Order, Harper Business

LeCun, Y., Bengio, Y. and Hinton, G. E. [2015] Deep Learning, Nature, Vol. 521

Lingjiao Chen, Matei Zaharia, James Zou [2023] "How is ChatGPT's Behavior Changing Over Time?", arXive, 30 Oct. 2023

Ministry of Internal Affairs and Communications Ministry of Economy, [2025] Trade and Industry ed., AI Guidelines for Business Ver1.01, 25 December 2024

OECD ed. [2023] Advancing Accountability in AI : Governing and Managing Risks Throughout The Lifecycle For Trustworthy AI, OECD Publishing

OECD ed.[2024] Assessing Potential Future Artificial Intelligence Risks, Benefits and Policy Imperatives 2024, OECD Publishing

Pan Pan, Alexander, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, Dan Hendrycks, [2023]. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark, arX-iv, 16 July 2023

Radford, Alec et.al. [2021] Learning Transferable Visual Models From Natural Language Supervision, arXiv: 2103.00020 (cs).

…Simon, H. A. [1st edition 1947] Administrative Behavior, (4th ed., The Free Press,1997)

Schwartz, Reva et.al. [2022] Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, NIST

Stanford University Human-Centered Institute for Artificial Intelligence [2023] Artificial Intelligence Index Report 2023, Stanford University

Stanford University Human Centered AI Institute ed. [2024] Artificial Intelligence Index Report 2024

NOTES:

(1) See Bommasani, Rishi et.al. [2022] and Stanford University Human-Centered Institute for Artificial Intelligence [2023].

(2) See Heaven, Will Douglas [2023].

(3) View Google I/O 2023 (https://io.google/2023/pro-gram/396cd2d5-9fe1-4725-a3dcc01bb2e2f38a/intl) on May 11, 2023.

(4) Even in RAG, there is a risk of halcination. GraphRAG is attracting attention as a deterrent to halcination occurrence.See Darren Edge et.al, From Local to Global: A Graph RAG Approach to Query-Focused Summarization, ArXiv, 24 Apr. 2024

(5) I have seen https://www.wsj.com/tech/personal-tech/ai-chat-bots-chatgpt-gemini-copilot-perplexity-claude-f9e40d26 (28 May 2024)

(6) See Rishi Bommasani et.al. On the Opportunities and Risks of Foundation Models, arXiv, 2022

(7) See Lingjiao Chen, Matei Zaharia, James Zou, "How is ChatGPT's Behavior Changing Over Time?", arXive, 30 Oct. 2023

(8) Stanford University Human Centered AI Institute ed. Artificial Intelligence Index Report 2024, p.33 and p.35
(9) Stanford University Human Centered AI Institute ed. Artificial Intelligence Index Report 2024, p.113
(10) See LLM Safety Leaderboard
(11) See Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, Dan Hendrycks Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark, arXiv, 16 July 2023
(12) See Japan Government, The Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems, Dec. 2023
(13) See Sumanth Dathathri, Pushmeet Kohli et.al, Scalable watermarking for identifying large language model outputs (in) Nature 634, 818 -823 (2024)
(14) See FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, URL:http//:AI.gov, USA White House House, 30 Oct, 2023
(15) AISI (AI Safety Institute), a government organization overseeing AI safety and security, was established in February 2024. This is a very important move in terms of the global development platform of advanced generation AI, ensuring safety, and international compatibility of systems. National Status Report on AI Safety in Japan 2024,5 Feb. 2025
(16) See Ministry of Internal Affairs and Communications Ministry of Economy, Trade and Industry ed., AI Guidelines for Business Ver1.01, 25 December 2024
(17) See AI Strategy Council and AI System Study Group ed., Interim Summary, Cabinet Office in Japan Goverment,4 Feb. 2025 (in Japanese)
(18) We have talked on Risks and Benefits of AI Futures at OECD Expert Meeting on AI Futures. See OECD ed., Assessing Potential Future Artificial Intelligence Risks, Benefits and Policy Imperatives 2024, OECD Publisher, Nov. 2024

ARTIFICIAL INTELLIGENCE IN INDIA: ADVANCES AND PENDING CHALLENGES

SUMMARY

I. Introduction **II.** Literature review **III.** Historical context of AI in India **IV.** Advancements in AI **V.** Challenges facing AI in India **VI.** Case studies **VII.** Future prospects and recommendations **VIII.** Conclusion **IX.** Bibliography

> BADAR ALAM IQBAL
Research Fellow, Department of Economics Stellenbosch University South Africa

> MOHAMAD ZREIK
Postdoctoral Fellow, School of International Studies, Sun Yat-sen University.

> ARTI YADAV
Assistant Professor, Department of Commerce, Ramanujan College, University of Delhi, India

I. INTRODUCTION

Artificial Intelligence (AI) is revolutionizing industries and societies around the globe, emerging as a critical component of the Fourth Industrial Revolution (Park, 2018). In recent years, India has recognized the transformative potential of AI and has embarked on an ambitious journey to integrate AI into various sectors of its economy (Chakraborty, 2020; Davenport & Mittal, 2023). The importance of AI in India's technological landscape cannot be overstated, as it promises to drive innovation, improve efficiency, and address complex challenges in healthcare, agriculture, education, and beyond. By leveraging AI, India aims to achieve significant socio-economic benefits and establish itself as a global leader in AI research and application.

This chapter aims to provide a comprehensive overview of the advances and pending challenges in AI within the Indian context. It begins by exploring the historical context of AI development in India, highlighting early milestones, government initiatives, and the contributions of key organizations and research institutions. The chapter then delves into the remarkable advancements made in AI across various sectors, showcasing the impact of AI-driven solutions in healthcare, agriculture, education, and other industries. Despite these advancements, several challenges hinder the full potential of AI in India. These include data privacy and security concerns, infrastructural and technological limitations, skill gaps, and regulatory hurdles.

The central research questions guiding this chapter are: What are the key advancements in AI across various sectors in India? What are the primary challenges that hinder the full potential of AI in India? How can these challenges be addressed to maximize the benefits of AI? By investigating these questions, this chapter seeks to offer a nuanced understanding of the current state

of AI in India and provide strategic recommendations for overcoming existing obstacles.

The methodology adopted for this chapter is descriptive and analytical, focusing on a thorough examination of existing literature, government reports, and case studies. Descriptive analysis helps in providing a detailed account of the advancements and challenges, while analytical methods are employed to interpret data, identify patterns, and draw meaningful conclusions. Theoretical frameworks related to technology adoption, data privacy, and regulatory policies will be utilized to analyze the findings and propose solutions. By combining descriptive and analytical approaches, this chapter aims to offer a comprehensive and insightful examination of AI in India.

By examining both the progress and the obstacles, this chapter aims to provide a balanced perspective on the current state of AI in India. It will also include case studies of successful AI projects to offer practical insights and lessons learned. Finally, the chapter will present future prospects and strategic recommendations for overcoming the existing challenges, emphasizing the role of the government, private sector, and academia in fostering the growth of AI. Through this comprehensive analysis, the chapter seeks to contribute to the ongoing discourse on AI development in India and highlight pathways for leveraging AI to achieve sustainable and inclusive growth.

II. LITERATURE REVIEW

The field of Artificial Intelligence has seen substantial growth and development globally, with significant contributions from various countries, including India. The literature on AI in India highlights a diverse range of advancements, applications, and challenges. This review synthesizes key findings from recent studies and reports to provide a comprehensive understanding of the current state of AI in India.

ADVANCEMENTS IN AI IN INDIA

India's AI landscape has been significantly shaped by government initiatives, research institutions, and private sector investments. According to Chatterjee (2020), the Indian government has launched several initiatives to promote AI, including the establishment of the National AI Strategy and the creation of AI research institutions. These efforts aim to position India as a global leader in AI and leverage AI for socio-economic development.

In the healthcare sector, AI has been employed to improve diagnostic accuracy, enhance patient care, and streamline administrative processes. Kumar *et al*. (2023) highlight the use of AI in diagnostic imaging, where machine learning algorithms assist in detecting diseases such as cancer and tuberculosis with high precision. Moreover, AI-driven tools are being used to manage patient records and predict disease outbreaks, contributing to better healthcare delivery and public health management.

Agriculture is another sector where AI has made significant inroads. Hassan, Malhotra, & Firdaus (2022) discuss how AI-powered technologies, such as precision farming and predictive analytics, are being used to optimize crop yields, manage pests, and conserve resources. These technologies provide farmers with real-time data on soil health, weather conditions, and crop growth, enabling more informed decision-making and sustainable agricultural practices.

Education in India has also benefited from AI. According to Jaiswal & Arun (2021), AI-driven educational platforms offer personalized learning experiences, adaptive assessments, and intelligent tutoring systems. These tools cater to the diverse needs of students, enhance engagement, and improve learning outcomes. Additionally, AI is being used to streamline administrative tasks in educational institutions, such as grading and student record management, allowing educators to focus more on teaching.

CHALLENGES FACING AI IN INDIA

Despite the notable advancements, several challenges impede the full potential of AI in India. One of the primary concerns is data privacy and security. As Sharma *et al*. (2022) note, the widespread adoption of AI involves the collection and processing of vast amounts of personal data, raising significant privacy concerns. There is a growing need for robust data protection regulations to safeguard individuals' privacy and ensure ethical AI deployment.

Infrastructural and technological limitations also pose significant challenges. According to Nedungadi *et al*. (2018), India's digital infrastructure is still developing, with varying levels of internet connectivity and access to advanced computational resources across different regions. This disparity hinders the uniform adoption of AI technologies and limits their impact, particularly in rural areas.

The skill gap in AI and related fields is another critical issue. Eubanks (2022) emphasize the need for a skilled workforce to develop, implement, and manage AI systems. While there are initiatives to enhance AI education and training, there remains a substantial gap between the demand for AI professionals and

the availability of qualified individuals. This skill gap restricts the growth of AI-driven industries and innovation.

Regulatory and policy challenges further complicate the AI landscape in India. Smuha (2021) discusses the lack of comprehensive AI-specific regulations and standards, which creates uncertainty and hampers the development of AI technologies. There is a need for clear and consistent policies that address ethical considerations, standardize AI practices, and promote innovation while ensuring public trust.

THEORETICAL PERSPECTIVES

The application of theoretical frameworks can provide a deeper understanding of the challenges and opportunities in AI. The Technology Acceptance Model (TAM), proposed by Davis (1989), is relevant in examining how users come to accept and use AI technologies. According to TAM, perceived usefulness and ease of use significantly influence users' acceptance of new technologies. In the Indian context, addressing the infrastructural and educational challenges can enhance the perceived ease of use and usefulness of AI, thereby promoting its wider adoption.

Another relevant framework is the Data Privacy Framework, which addresses the ethical and legal aspects of data collection, storage, and usage. According to Solove (2006), effective data privacy frameworks are essential for maintaining public trust and ensuring the ethical deployment of AI. In India, the implementation of robust data privacy laws and regulations can mitigate privacy concerns and foster a secure environment for AI innovation.

CASE STUDIES AND PRACTICAL INSIGHTS

Several case studies illustrate the successful application of AI in India and provide valuable lessons for future initiatives. For instance, the AI-powered healthcare project by Wadhwani AI in Maharashtra has significantly improved tuberculosis detection rates. According to the project report (Jain, Garg, & Khera, 2022), the use of AI algorithms in analyzing chest X-rays has enabled early diagnosis and treatment, reducing the disease's spread and improving patient outcomes.

In agriculture, the Precision Farming Project by Microsoft and ICRISAT has demonstrated the potential of AI in enhancing agricultural productivity. As reported by Kumar *et al.* (2020), the project uses AI models to analyze weather data, soil health, and crop conditions, providing farmers with actionable insights to optimize farming practices and improve yields.

FUTURE PROSPECTS AND RECOMMENDATIONS

The future of AI in India holds immense potential, provided the existing challenges are addressed effectively. A collaborative approach involving the government, private sector, and academia is essential for fostering AI growth. The government should continue to support AI research and development through funding and policy initiatives. Additionally, investing in digital infrastructure and AI education can bridge the skill gap and ensure equitable access to AI technologies across the country.

Private sector companies can play a crucial role by developing innovative AI solutions tailored to the Indian context and collaborating with academic institutions for research and training. Furthermore, establishing clear and comprehensive regulatory frameworks can provide a conducive environment for AI development while addressing ethical and privacy concerns.

III. HISTORICAL CONTEXT OF AI IN INDIA

The evolution of Artificial Intelligence in India is marked by several significant developments and milestones that have shaped its current landscape. The journey began in the late 20th century when Indian research institutions and universities started exploring the potential of AI. One of the earliest milestones was the establishment of the Centre for Artificial Intelligence and Robotics (CAIR) in 1986 under the Defence Research and Development Organisation (DRDO) (Mallick, 2018). CAIR focused on developing AI technologies for defense applications, setting the stage for AI research in the country (Basrur & Wu, 2023).

As AI gained global prominence, the Indian government recognized its strategic importance and began formulating policies to foster AI development. In 2018, the National Institution for Transforming India (NITI Aayog) launched the National Strategy for Artificial Intelligence, also known as #AIforAll (George, 2024). This strategy outlined India's vision to leverage AI for inclusive growth and economic development, focusing on key sectors such as healthcare, agriculture, education, smart cities, and smart mobility. The policy emphasized the need for creating a robust AI ecosystem, promoting research and innovation, and addressing ethical and privacy concerns associated with AI.

Government initiatives have played a pivotal role in promoting AI research and development. The Ministry of Electronics and Information Technology (MeitY) has been instrumental in supporting AI projects through various schemes and programs (Chatterjee, S., & NS, 2022). One such initiative is the establishment

of the National AI Portal, a one-stop digital platform for AI resources, news, and research (Chatterjee, 2020). Additionally, the government has launched the Atal Innovation Mission (AIM), which aims to create an environment of innovation and entrepreneurship, encouraging the development of AI-based solutions (Ellikkal & Rajamohan, 2023).

Several key organizations and research institutions have contributed significantly to AI advancements in India. The Indian Institutes of Technology (IITs) and Indian Institutes of Information Technology (IIITs) are at the forefront of AI research and education (George, 2024). IIT Bombay, IIT Delhi, and IIIT Hyderabad, in particular, have established dedicated AI research centers that focus on cutting-edge AI technologies and their applications (Sanyal & Chakrabarti, 2020). These institutions collaborate with industry partners and international organizations to drive AI innovation and address real-world challenges.

In addition to academic institutions, various government and private sector organizations have made substantial contributions to AI research and development. The Indian Space Research Organisation (ISRO) has integrated AI into its satellite and space exploration programs, enhancing data analysis and mission planning capabilities (Saxena, 2023). The National Association of Software and Service Companies (NASSCOM) has also played a crucial role by establishing the Centre of Excellence for Data Science and Artificial Intelligence (CoE-DSAI) (Panigrahi, 2024). This center aims to accelerate AI adoption in industry and promote collaboration between academia, industry, and government.

Furthermore, the private sector has been a key driver of AI innovation in India. Leading technology companies such as Tata Consultancy Services (TCS), Infosys, and Wipro have invested heavily in AI research and development (Chakrabarti & Sanyal, 2020). These companies collaborate with academic institutions and startups to develop AI solutions that cater to various industries, including healthcare, finance, and manufacturing.

IV. ADVANCEMENTS IN AI

Artificial Intelligence has witnessed remarkable advancements in India, transforming various sectors and driving innovation. This section explores the significant impact of AI in healthcare, agriculture, education, and other sectors, highlighting specific applications, case studies, and success stories.

AI IN HEALTHCARE

AI has revolutionized the healthcare sector in India, offering solutions that enhance diagnostic accuracy, streamline administrative processes, and improve patient outcomes. One notable application is in diagnostic imaging, where AI algorithms assist in analyzing medical images to detect diseases such as cancer, tuberculosis, and cardiovascular conditions. For example, Roy (2022) reported that AI-driven tools, like Qure.ai's qXR, have been instrumental in diagnosing tuberculosis by analyzing chest X-rays with high precision. This technology not only speeds up the diagnostic process but also ensures early detection, leading to timely treatment and better patient prognosis.

AI applications in healthcare also extend to predictive analytics and personalized medicine. Predictive analytics uses AI to forecast disease outbreaks and patient deterioration, enabling healthcare providers to take proactive measures. Personalized medicine, on the other hand, involves tailoring treatment plans based on individual patient data, such as genetic information and medical history. According to Sharma *et al*. (2022), companies like PharmEasy and 1mg are leveraging AI to offer personalized healthcare services, improving treatment efficacy and patient satisfaction.

Several case studies illustrate the impact of AI in healthcare. For instance, the AI-powered project by Wadhwani AI in Maharashtra significantly improved tuberculosis detection rates. The project used machine learning algorithms to analyze chest X-rays, enabling early diagnosis and reducing the spread of the disease. Another notable example is the use of AI by Apollo Hospitals to predict heart disease risk using patient data, which has led to better preventive care and reduced incidence of heart attacks (Agarwal *et al.,* 2022).

AI IN AGRICULTURE

The agricultural sector in India has also benefited significantly from AI, with smart farming techniques and predictive analytics playing a crucial role in enhancing productivity and sustainability. AI technologies provide farmers with real-time data on soil health, weather conditions, and crop growth, enabling them to make informed decisions and optimize resource use.

Smart farming techniques, such as precision agriculture, leverage AI to monitor and manage crops efficiently. Shaikh, Rasool, & Lone (2022) highlight the use of AI-powered drones and sensors to collect data on crop health, soil moisture, and pest infestations. This data is then analyzed to provide actionable

insights, helping farmers apply the right amount of water, fertilizers, and pesticides at the right time. As a result, farmers can increase crop yields, reduce costs, and minimize environmental impact.

Success stories in AI-driven agriculture include the Precision Farming Project by Microsoft and ICRISAT. This project uses AI models to analyze weather data, soil health, and crop conditions, providing farmers with actionable insights to optimize farming practices and improve yields (Kumar *et al.,* 2020). The project has demonstrated significant improvements in crop productivity and resource management, showcasing the potential of AI in transforming agriculture.

Another success story is the use of AI by startups like CropIn and Ninjacart. CropIn leverages AI to provide farmers with real-time data on crop health and market trends, helping them make informed decisions and maximize profits (Wolfert *et al.,* 2017). Ninjacart uses AI to streamline the supply chain, connecting farmers directly with retailers and reducing post-harvest losses (Suresh *et al.,* 2024). These initiatives have not only improved agricultural productivity but also enhanced the livelihoods of farmers.

AI IN EDUCATION

AI is transforming the education sector in India by offering personalized learning experiences, adaptive assessments, and intelligent tutoring systems. These AI-driven tools cater to the diverse needs of students, enhance engagement, and improve learning outcomes.

Personalized learning involves tailoring educational content and teaching methods to individual student needs and preferences. AI-powered platforms like BYJU's and Vedantu use machine learning algorithms to analyze student performance data and provide customized learning paths (Goswami & Sharma, 2024). These platforms offer interactive lessons, quizzes, and feedback, helping students learn at their own pace and improve their understanding of complex concepts.

AI-driven tools also support adaptive assessments, which adjust the difficulty of questions based on student performance. This approach ensures that assessments are neither too easy nor too difficult, providing an accurate measure of student knowledge and skills. According to Karan & Angadi (2023), AI-based assessment tools like Embibe are helping educators identify learning gaps and provide targeted interventions to address them.

Several examples of AI in Indian educational institutions illustrate the transformative potential of these technologies. For instance, IIT Bombay has developed an AI-powered chatbot, Asha, to assist students with academic queries

and provide personalized guidance (Shrivastava & Mahajan, 2016). The chatbot uses natural language processing (NLP) to understand student questions and deliver relevant answers, improving student support and engagement.

Another example is the use of AI by Amity University to enhance the learning experience. The university has integrated AI-powered tools into its curriculum, offering students access to intelligent tutoring systems, virtual labs, and personalized feedback. These tools have improved student performance and satisfaction, demonstrating the value of AI in education.

OTHER SECTORS

Beyond healthcare, agriculture, and education, AI has also made significant inroads in other sectors such as finance, transportation, and manufacturing.

In the finance sector, AI is being used to enhance fraud detection, automate customer service, and improve investment strategies. For example, Bajaj Finserv employs AI algorithms to detect fraudulent transactions and protect customer accounts (Patil & Kulkarni, 2019). AI-powered chatbots and virtual assistants are also being used by banks and financial institutions to provide 24/7 customer support and streamline operations.

The transportation sector has seen the adoption of AI in areas such as traffic management, autonomous vehicles, and logistics. According to Samim (2023), AI-driven traffic management systems are being implemented in cities like Bangalore and Delhi to reduce congestion and improve traffic flow. Autonomous vehicle initiatives by companies like Tata Motors and Mahindra & Mahindra are also exploring the potential of AI to enhance road safety and driving efficiency

In manufacturing, AI is being used to optimize production processes, improve quality control, and predict maintenance needs. AI-powered robots and automation systems are enhancing productivity and reducing costs in manufacturing plants. For instance, Tata Steel uses AI to monitor and analyze production data, identifying areas for improvement and ensuring high-quality output (Gajdzik & Wolniak, 2021).

V. CHALLENGES FACING AI IN INDIA

Despite the remarkable advancements in AI, several challenges impede its full potential in India. These challenges span across data privacy and security concerns, infrastructural and technological limitations, skill gaps, and regulatory

and policy hurdles. Addressing these challenges is crucial for the sustainable and ethical development of AI in the country.

DATA PRIVACY AND SECURITY CONCERNS

Data privacy and security are paramount concerns in the AI landscape, given the vast amounts of personal and sensitive data that AI systems process. The adoption of AI technologies often involves collecting and analyzing data from various sources, raising significant privacy issues. Dixon (2017) emphasize that without robust data protection measures, individuals' privacy can be compromised, leading to misuse of personal information and identity theft. Moreover, the increasing instances of data breaches highlight the urgent need for enhanced security protocols.

Legal and ethical implications of data usage in AI are profound. The lack of comprehensive data protection laws in India exacerbates these concerns. The Personal Data Protection Bill, introduced in 2019, aims to address some of these issues by establishing guidelines for data collection, storage, and processing (Hoofnagle, Van Der Sloot, & Borgesius, 2019). However, its implementation has faced delays, leaving gaps in the legal framework. Ethical considerations, such as bias in AI algorithms and the potential for discrimination, also require stringent oversight to ensure AI systems are fair and unbiased.

Public perception and trust issues further complicate the adoption of AI. According to a survey by Sharma *et al*. (2022), a significant portion of the Indian population is wary of AI due to concerns about data privacy and job displacement. Building public trust in AI technologies necessitates transparent practices, robust data protection measures, and clear communication about the benefits and risks of AI.

INFRASTRUCTURAL AND TECHNOLOGICAL LIMITATIONS

Infrastructural and technological limitations present significant barriers to the widespread adoption of AI in India. One major challenge is the access to high-quality data, which is essential for training effective AI models. Data in India is often fragmented, inconsistent, and of varying quality, making it difficult to derive actionable insights. Additionally, data silos within organizations hinder the sharing and integration of data across platforms.

The need for better computational resources is another critical issue. Advanced AI algorithms require substantial computing power, which is not

readily available in many parts of India. According to Lins *et al.* (2021), the lack of high-performance computing infrastructure limits the ability of researchers and businesses to develop and deploy sophisticated AI solutions. Investing in state-of-the-art data centers and cloud computing services is essential to overcome these limitations.

SKILL GAP AND WORKFORCE READINESS

The skill gap in AI and related fields is a significant hurdle in India's AI journey. While there is a growing demand for AI professionals, the supply of skilled individuals remains limited. Sharma *et al.* (2022) highlight that a considerable portion of the workforce lacks the necessary skills to develop, implement, and manage AI systems. This gap restricts innovation and slows down the adoption of AI technologies across industries.

Educational and training requirements need to be addressed to bridge this skill gap. Enhancing AI education in schools, colleges, and universities is crucial. Incorporating AI and machine learning courses into curricula, providing hands-on training, and encouraging research in AI are necessary steps. Online platforms offering AI courses and certifications, such as Coursera, edX, and Udacity, are also valuable resources for upskilling the workforce.

Several initiatives aim to bridge the skill gap in AI. The National AI Portal, launched by the Indian government, offers resources and training programs to build AI capabilities (Chatterjee, 2020). Additionally, the Atal Innovation Mission (AIM) and the NITI Aayog's AI for All initiative focus on promoting AI literacy and encouraging innovation. Collaborations between academia, industry, and government are essential to create a skilled workforce ready to harness the potential of AI.

REGULATORY AND POLICY CHALLENGES

Regulatory and policy challenges significantly impact the development and deployment of AI in India. The current regulatory framework is fragmented and lacks comprehensive guidelines specific to AI. Smuha (2021) notes that the absence of clear regulations creates uncertainty and hampers the growth of AI technologies. There is a need for policies that address ethical considerations, standardize AI practices, and promote innovation while ensuring public trust.

Current regulations, such as the Information Technology Act of 2000, provide a general framework for data protection but do not specifically address

the unique challenges posed by AI (Basu & Jones, 2005). The Personal Data Protection Bill, once implemented, is expected to fill some of these gaps by establishing stricter data protection norms. However, the dynamic nature of AI technologies necessitates continuous updates to the regulatory framework to keep pace with technological advancements.

Suggestions for policy improvements include developing a comprehensive AI strategy that outlines ethical guidelines, standards for data usage, and mechanisms for accountability. Establishing regulatory sandboxes can allow for the testing of AI applications in a controlled environment, facilitating innovation while ensuring compliance with ethical and legal standards. Additionally, fostering international collaborations can help India align its AI policies with global best practices and stay competitive in the global AI landscape.

ADDRESSING ETHICAL CONCERNS

Ethical concerns in AI revolve around issues such as bias, fairness, and accountability. AI algorithms, if not carefully designed, can perpetuate existing biases and lead to discriminatory outcomes. Ensuring fairness in AI systems requires rigorous testing and validation to identify and mitigate biases. Transparency in AI decision-making processes is also essential to build trust and accountability.

The ethical use of AI also involves considering the societal impact of AI technologies. The potential for job displacement due to automation is a significant concern. Policies must address the need for reskilling and upskilling workers to adapt to the changing job market. Additionally, ethical frameworks should guide the development and deployment of AI in sensitive areas such as healthcare and criminal justice to prevent harm and ensure equitable outcomes.

INFRASTRUCTURE DEVELOPMENT

Developing the necessary infrastructure to support AI is critical for its widespread adoption. This includes investing in high-speed internet connectivity, particularly in rural areas, to ensure that AI technologies can be accessed and utilized by all. Enhancing data storage and processing capabilities through the establishment of modern data centers is also essential.

Public-private partnerships can play a crucial role in infrastructure development. Collaborations between the government and private sector can accelerate the deployment of AI infrastructure and ensure that resources are utilized

effectively. For instance, initiatives like the Digital India program aim to improve digital infrastructure and promote digital literacy, creating a conducive environment for AI growth (Chatterjee, 2020).

PROMOTING AI RESEARCH AND INNOVATION

Promoting AI research and innovation is vital for maintaining India's competitive edge in the global AI landscape. Encouraging interdisciplinary research that combines AI with other fields, such as healthcare, agriculture, and environmental science, can lead to innovative solutions for complex problems. Providing funding and resources for AI research projects can foster a culture of innovation and drive technological advancements.

Creating innovation hubs and incubators can support startups and entrepreneurs working on AI technologies. These hubs can provide access to mentorship, funding, and networking opportunities, enabling the growth of a vibrant AI ecosystem. Additionally, fostering collaborations between academic institutions, research organizations, and industry can facilitate knowledge exchange and accelerate innovation.

VI. CASE STUDIES

AI has been applied successfully across various sectors in India, with several projects standing out for their innovative approaches and impactful results. Detailed analysis of these successful AI projects provides valuable insights and lessons that can guide future initiatives.

One notable AI project is the Wadhwani AI initiative focused on improving tuberculosis (TB) detection rates in Maharashtra. This project utilized machine learning algorithms to analyze chest X-rays, significantly enhancing the accuracy and speed of TB diagnosis. Before the implementation of AI, TB detection relied heavily on manual analysis, which was time-consuming and prone to errors (Jain, Garg, & Khera, 2022). The AI-driven approach not only accelerated the diagnostic process but also ensured early detection, which is crucial for effective treatment and control of the disease. According to the project report, the use of AI in TB detection led to a 20% increase in the identification of positive cases, highlighting the potential of AI to transform healthcare diagnostics.

Another successful AI application is seen in the agricultural sector through the Precision Farming Project by Microsoft and ICRISAT (International Crops Research Institute for the Semi-Arid Tropics). This project leverages AI models

to analyze weather patterns, soil health, and crop conditions, providing farmers with actionable insights to optimize their farming practices. For instance, AI-driven recommendations on irrigation schedules, fertilizer application, and pest management have helped farmers increase crop yields while reducing resource use. The project's impact is evident in the significant improvements in productivity and sustainability observed in participating farms. Kumar *et al*. (2020) reported that the project led to a 30% increase in crop yields and a 20% reduction in water usage, demonstrating the transformative potential of AI in agriculture.

In the education sector, AI-powered platforms like BYJU's have revolutionized personalized learning. BYJU's uses machine learning algorithms to assess individual student performance and provide customized learning paths. The platform offers interactive lessons, quizzes, and feedback tailored to each student's learning pace and style. This personalized approach has resulted in improved engagement and learning outcomes. According to a study by Goswami & Sharma (2024), students using BYJU's showed a 25% improvement in their academic performance compared to those using traditional learning methods. This case study underscores the effectiveness of AI in enhancing educational experiences and outcomes.

AI's application in finance is exemplified by HDFC Bank's AI-driven customer service initiative. The bank implemented an AI-powered chatbot named EVA (Electronic Virtual Assistant) to handle customer queries and provide instant support. EVA uses natural language processing (NLP) to understand and respond to customer questions, offering a seamless and efficient customer service experience. Since its launch, EVA has handled millions of customer interactions, reducing response times and improving customer satisfaction. Patil & Kulkarni (2019) reported a 50% reduction in call center workload and a 20% increase in customer satisfaction scores, highlighting the benefits of AI in customer service.

LESSONS LEARNED

Several lessons can be drawn from these successful AI projects in India. Firstly, collaboration between various stakeholders, including government bodies, private sector companies, and research institutions, is crucial for the successful implementation of AI projects. The involvement of multiple stakeholders ensures access to diverse expertise, resources, and perspectives, facilitating comprehensive and effective solutions.

Secondly, the importance of context-specific AI solutions cannot be overstated. AI technologies must be tailored to address the unique challenges and needs of different sectors and regions. For example, the Precision Farming Project's success is largely attributed to its focus on the specific agricultural conditions of semi-arid regions in India.

Another key lesson is the necessity of robust data infrastructure. High-quality, reliable data is the backbone of effective AI systems. Projects like the Wadhwani AI TB detection initiative relied on extensive datasets of chest X-rays to train their machine learning models, highlighting the importance of data accessibility and quality.

Additionally, continuous monitoring and evaluation are essential to measure the impact of AI projects and identify areas for improvement. Regular assessments help in fine-tuning AI models, ensuring their accuracy and relevance over time.

Finally, addressing ethical and privacy concerns is vital for building public trust and ensuring the responsible use of AI. Transparent practices, clear communication, and adherence to data protection regulations are necessary to mitigate public concerns and promote the ethical deployment of AI technologies.

VII. FUTURE PROSPECTS AND RECOMMENDATIONS

The future prospects of AI in India are promising, with the potential to revolutionize various sectors and drive significant socio-economic growth. As AI technologies continue to advance, they are expected to become more integrated into everyday applications, offering innovative solutions to complex problems. In healthcare, future developments may include more sophisticated AI-driven diagnostic tools and personalized treatment plans that leverage genetic data and patient history. In agriculture, AI could enable precision farming on a larger scale, optimizing resource use and boosting crop yields. In education, AI-driven personalized learning platforms may become even more adaptive, offering customized educational experiences that cater to individual student needs and learning styles.

To fully realize these potential developments, strategic recommendations must be implemented to overcome current challenges. Addressing data privacy and security concerns is paramount. The implementation of robust data protection laws, such as the Personal Data Protection Bill, should be prioritized to safeguard personal information and build public trust in AI technologies. Additionally, investing in high-quality data infrastructure and computational resources is

essential to support the development and deployment of advanced AI models. Enhancing internet connectivity, particularly in rural areas, will ensure that AI technologies are accessible to all, bridging the digital divide.

Another critical area is the skill gap in AI and related fields. Educational institutions should integrate AI and machine learning courses into their curricula and provide hands-on training opportunities. Online platforms offering AI courses and certifications can complement formal education and help upskill the workforce. Collaborative efforts between the government, private sector, and academia are crucial in creating a skilled workforce ready to harness the potential of AI.

The government plays a pivotal role in fostering AI growth through policy initiatives, funding for research and development, and the creation of a conducive regulatory environment. Policies that promote innovation while addressing ethical and legal considerations will be vital in guiding the responsible development of AI. The private sector, on the other hand, can drive AI adoption by investing in research, developing AI solutions tailored to the Indian context, and fostering partnerships with academic institutions and startups.

Academic institutions are instrumental in advancing AI research and training the next generation of AI professionals. By establishing AI research centers, promoting interdisciplinary research, and collaborating with industry partners, academia can contribute significantly to AI innovation. Furthermore, creating innovation hubs and incubators can support startups and entrepreneurs working on AI technologies, providing them with the resources and mentorship needed to succeed.

VIII. CONCLUSION

This chapter has explored the significant advancements and ongoing challenges in the field of Artificial Intelligence in India, highlighting the transformative potential of AI across various sectors such as healthcare, agriculture, education, and more. The historical context provided an understanding of the early developments and government initiatives that have shaped India's AI landscape, while detailed case studies demonstrated the practical applications and impacts of AI technologies.

Despite the remarkable progress, several challenges hinder the full potential of AI in India. Data privacy and security concerns, infrastructural and technological limitations, skill gaps, and regulatory and policy hurdles present significant barriers that must be addressed. The analysis underscored the importance of

robust data protection laws, investment in high-quality data infrastructure, and enhanced computational resources. Additionally, the need for a skilled workforce and comprehensive regulatory frameworks was emphasized as critical to fostering AI growth.

Future prospects for AI in India are promising, with the potential to revolutionize various sectors and drive socio-economic growth. Strategic recommendations for overcoming current challenges include prioritizing data privacy, investing in infrastructure, bridging the skill gap, and fostering a collaborative ecosystem involving the government, private sector, and academia. The role of the government in providing policy support, the private sector in driving adoption and innovation, and academia in advancing research and training is crucial in realizing the full potential of AI.

IX. BIBLIOGRAPHY

Agarwal, Y., Jain, M., Sinha, S., & Dhir, S. (2020). Delivering high-tech, AI-based health care at Apollo Hospitals. *Global Business and Organizational Excellence*, *39*(2), 20-30.

Basrur, R., & Wu, S. S. (2023). India's conventional strategy in a nuclear environment: a neglected link. *Defence Studies*, *23*(3), 457-476.

Basu, S., & Jones, R. (2005). Indian Information and Technology Act 2000: review of the regulatory powers under the Act. *International Review of Law, Computers & Technology*, *19*(2), 209-230.

Chakrabarti, R., & Sanyal, K. (2020). Towards a 'Responsible AI': Can India take the lead?. *South Asia Economic Journal*, *21*(1), 158-177.

Chakraborty, U. (2020). *Artificial Intelligence for All: Transforming Every Aspect of Our Life*. Bpb publications.

Chatterjee, S. (2020). AI strategy of India: policy framework, adoption challenges and actions for government. *Transforming Government: People, Process and Policy*, *14*(5), 757-775.

Chatterjee, S., & NS, S. (2022). Artificial intelligence and human rights: a comprehensive study from Indian legal and policy perspective. *International Journal of Law and Management*, *64*(1), 110-134.

Davenport, T. H., & Mittal, N. (2023). *All-in on AI: How smart companies win big with artificial intelligence*. Harvard Business Press.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). Technology acceptance model. *J Manag Sci*, *35*(8), 982-1003.

Dixon, P. (2017). A Failure to "Do No Harm"–India's Aadhaar biometric ID program and its inability to protect privacy in relation to measures in Europe and the US. *Health and technology*, *7*(4), 539-567.

Ellikkal, A., & Rajamohan, S. (2023). Unleashing the Entrepreneurial Spirit: An Analysis of Government Programs and their Impact on Entrepreneurship Education in India. *Journal of Management*, *10*(4), 28-37.

Eubanks, B. (2022). *Artificial intelligence for HR: Use AI to support and develop a successful workforce*. Kogan Page Publishers.

Gajdzik, B., & Wolniak, R. (2021). Transitioning of steel producers to the steelworks 4.0–Literature review with case studies. *Energies*, *14*(14), 4109.

George, A. S. (2024). India's Ascent as the Global Epicenter of Artificial Intelligence. *Partners Universal Innovative Research Publication*, *2*(1), 1-15.

Goswami, A., & Sharma, A. (2024). AI For Bridging Socio-Economic Inequities in Indian Education Space. *International Journal of Research and Scientific Innovation*, *11*(4), 890-935.

Hassan, M., Malhotra, K., & Firdaus, M. (2022). Application of artificial intelligence in IoT security for crop yield prediction. *ResearchBerg Review of Science and Technology*, *2*(1), 136-157.

Hoofnagle, C. J., Van Der Sloot, B., & Borgesius, F. Z. (2019). The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, *28*(1), 65-98.

Jain, R., Garg, N., & Khera, S. N. (2022). Adoption of AI-enabled tools in social development organizations in India: An extension of UTAUT model. *Frontiers in Psychology*, *13*, 893691.

Jaiswal, A., & Arun, C. J. (2021). Potential of Artificial Intelligence for transformation of the education system in India. *International Journal of Education and Development using Information and Communication Technology*, *17*(1), 142-158.

Karan, B., & Angadi, G. R. (2023). Artificial intelligence integration into school education: a review of indian and foreign perspectives. *Millennial Asia*, 09763996231158229.

Kumar, R., Yadav, S., Kumar, M., Kumar, J., & Kumar, M. (2020). Artificial intelligence: new technology to improve Indian agriculture. *International Journal of Chemical Studies*, *8*(2), 2999-3005.

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of ambient intelligence and humanized computing*, *14*(7), 8459-8486.

Lins, S., Pandl, K. D., Teigeler, H., Thiebes, S., Bayer, C., & Sunyaev, A. (2021). Artificial intelligence as a service: classification and research directions. *Business & Information Systems Engineering*, *63*, 441-456.

Mallick, P. K. (2018). Artificial Intelligence in Armed Forces: An Analysis. *CLAWS journal*, *11*(2), 63-79.

Nedungadi, P. P., Menon, R., Gutjahr, G., Erickson, L., & Raman, R. (2018). Towards an inclusive digital literacy framework for digital India. *Education+ Training*, *60*(6), 516-528.

Panigrahi, C. M. A. (2024). Impact of artificial intelligence on Indian economy. *Panigrahi A, Ahirrao SC, Patel A, Impact of artificial intelligence on Indian economy. J Manag Res Anal*, *11*(1), 33-40.

Park, S. C. (2018). The Fourth Industrial Revolution and implications for innovative cluster policies. *Ai & Society*, *33*, 433-445.

Patil, K., & Kulkarni, M. S. (2019). Artificial intelligence in financial services: Customer chatbot advisor adoption. *Int. J. Innov. Technol. Explor. Eng*, *9*(1), 4296-4303.

Roy, R. (2022). The COVID-19 opportunity for Qure. ai. *The International Journal of Entrepreneurship and Innovation*, *23*(2), 138-143.

Samim, A. (2023). Analysis of Artificial Intelligence in Traffic Congestion and Management System. *International Journal of Science and Research (IJSR)*, *12*(1), 679-684.

Sanyal, K., & Chakrabarti, R. (2020). *Artificial Intelligence and India*. Oxford University Press.

Saxena, A. (2023). India's Space Policy and Counter-Space Capabilities. *Strategic Analysis*, *47*(2), 146-158.

Shaikh, T. A., Rasool, T., & Lone, F. R. (2022). Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. *Computers and Electronics in Agriculture*, *198*, 107119.

Sharma, A., Sharma, R. K., Nanda, S., & Misra, A. (2022). Emerging real time analytics based health start-ups: opportunities during covid-19. *Journal of Medical Pharmaceutical and Allied Sciences*, *11*(4), 5017-5025.

Sharma, M., Luthra, S., Joshi, S., & Kumar, A. (2022). Implementing challenges of artificial intelligence: Evidence from public manufacturing sector of an emerging economy. *Government Information Quarterly*, *39*(4), 101624.

Shrivastava, R., & Mahajan, P. (2016). Artificial intelligence research in India: a scientometric analysis. *Science & Technology Libraries*, *35*(2), 136-151.

Smuha, N. A. (2021). From a 'race to AI'to a 'race to AI regulation': regulatory competition for artificial intelligence. *Law, Innovation and Technology*, *13*(1), 57-84.

Solove, D. J. (2005). A taxonomy of privacy. Univ. Pa. Law Rev. 154(3), 477–560

Suresh, D., Choudhury, A., Zhang, Y., Zhao, Z., & Shaw, R. (2024). The Role of Data-Driven Agritech Startups–The Case of India and Japan. *Sustainability*, *16*(11), 4504.

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big data in smart farming–a review. *Agricultural systems*, *153*, 69-80.

CHAPTER SIX

ARTIFICIAL INTELLIGENCE AND PUBLIC SECURITY

ARTIFICIAL INTELLIGENCE, DISINFORMATION AND PUBLIC SECURITY IN LATIN AMERICA

CONTENT:

I. Introduction **II.** Artificial intelligence in times of misinformation **III.** Artificial Intelligence and Public Safety in Latin America **IV.** Conclusion and findings

> FAUSTO CARBAJAL GLASS

Consultant in political risk and security. Internationalist from Universidad Iberoamericana and Master in War Studies from King's College London.

> AMAURY FIERRO GONZÁLEZ

Specialist in intelligence and geopolitics and instructor at Rabdan Academy.

I. INTRODUCTION

The emergence of artificial intelligence (AI) has profoundly transformed the way societies organise, communicate and interact. In Latin America and the Caribbean, this transformation presents both unique opportunities and challenges, particularly in two areas critical to democratic governance and social development: disinformation and public security. Although different in nature, these phenomena share a fundamental characteristic, namely their ability to erode public trust in institutions and alter the social fabric of communities.

The Latin American region faces a particular situation in which multiple crises converge: institutional, security, political legitimacy and social cohesion. In this context, AI emerges as a double-edged sword. On the one hand, it offers innovative solutions to combat deep-rooted problems. On the other hand, its inappropriate implementation can exacerbate existing inequalities and create new vulnerabilities for entire societies. This duality is clearly evident in the two thematic areas addressed in this chapter.

In the realm of disinformation, Latin America has become particularly fertile ground for the proliferation of fake news and manipulated content. As will be seen below, the massive penetration of social media, combined with low levels of digital literacy and highly polarised political contexts, has created ideal conditions for the spread of disinformation. AI, through technologies such as deepfakes and automated content generation, has amplified these challenges. However, it also provides sophisticated tools for detecting and combating disinformation, from automatic verification algorithms to dissemination pattern analysis systems. Thus, as the reader will best judge, this chapter will also focus on the friendly side of AI.

In terms of public safety, the region faces complex challenges, including high crime rates, transnational organised crime, and endemic homicidal violence in certain areas. AI offers promising solutions through predictive systems, intelligent video surveillance, and crime pattern analysis. However, the implementation of these technologies raises significant dilemmas around privacy, civil rights, and the ethical use of mass surveillance. The temptation of "technological solutionism" –the belief that technology alone can solve complex social problems– is particularly strong in this area, and the chapter will also revisit this concern.

The chapter therefore examines the intersection of AI with these two critical phenomena for Latin America, analysing both the promising innovations and the risks inherent in their implementation. The first section addresses how AI is transforming the landscape of disinformation, exploring both its role in creating misleading content and its potential to combat it. The second section analyses the application of AI in public security, examining cases of implementation in various countries in the region and assessing its impact on police effectiveness and citizens' rights.

The analysis is based on specific cases of implementation in various Latin American countries, highlighting how different socio-political contexts and regulatory frameworks lead to different outcomes. It also considers the ethical and social implications of these technologies, recognising that their implementation does not occur in a vacuum, but in societies marked by profound inequalities and structural challenges.

The ultimate goal is to provide a balanced understanding of how AI can contribute to addressing two of the most pressing challenges facing Latin America, without losing sight of the need for a comprehensive approach that considers the social, ethical, and political dimensions of its implementation. Only through this holistic understanding will it be possible to harness the transformative potential of AI while mitigating its risks and protecting the fundamental rights of citizens.

II. ARTIFICIAL INTELLIGENCE IN TIMES OF MISINFORMATION

The relationship between human beings and truth has been particularly complex throughout history. Perhaps even more so today with social media and artificial intelligence (AI). Not because we have necessarily achieved it before, but because it seems that we are moving away from it today. Our relationship with

truth today cannot be understood without considering the ambiguities of the VUCA world –more volatile, uncertain, complex and ambiguous. Everything can be or not be, depending on our beliefs, paradigms and, above all, emotions[1]. Moreover, in times of disinformation and fake news, it could be said that it is no longer about the search for truth itself, but rather the impulse to reaffirm our own cognitive biases[2]. And just a step away is relativism.

What is worrying about this way of relating to information –and to truth– is that it has direct and tangible implications for societies, particularly democratic ones. For example, the political use of artificial intelligence (AI), through the creation and dissemination of deepfakes and other biased content, makes it increasingly difficult for citizens to discern truth from falsehood, which inevitably weakens institutions and undermines the democratic process[3]. As Timothy Snyder pointed out: the problem begins when truth and hard data become irrelevant in a society that has ceased to be factual and has instead become an oracular society[4]. This transition from factual to oracular societies has laid the groundwork for disinformation –that is, the deliberate and systematic dissemination of lies[5]– to be present in major national and international events.

In fact, the World Economic Forum's (WEF) Global Risks Report 2024 identifies disinformation as one of the main risks facing the world precisely because of its severity in the short and long term[6]. The concern generated by disinformation

1 Para un estudio sobre el componente emocional de la posverdad en las redes sociales, ver: Durnová, Anna. 2019. "Unpacking emotional contexts of post-truth". Critical Policy Studies, 13 (4), 447-450.

2 De acuerdo con la literatura en la materia, la redes sociales exacerban primordialmente cinco sesgos cognitivos, a saber: rebaño, encuadre, exceso de confianza, confirmación y anclaje. Ver: French, Aaron M.; Storey, Veda C. & Wallace, Linda. 2023. "The impact of cognitive biases on the believability of fake news". European Journal of Information Systems, 1-22.

3 Labuz, Mateusz & Nehring, Christopher. 2024. "On the way to deep fake democracy? Deep fakes in election campaigns in 2023". European Political Science, DOI https://doi.org/10.1057/s41304-024-00482-9; Diakopoulos, Nicholas & Johnson, Deborah. 2020. "Anticipating and addressing the ethical implications of deepfakes in the context of elections". New Media & Society, 23 (7), 2072 – 2098. https://journals.sagepub.com/doi/abs/10.1177/1461444820925811.

4 Snyder, Timothy. 2017. On Tiranny: Twenty Lessons from the Twentieth Century. Nueva York: Crown Publishing.

5 Broda, Elena & Strömbäck, Jesper. 2024. "Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review". Annals of the International Communication Association, 48 (2), 139 – 166.

6 WEF. 2024. "Global Risks Report 2024". Disponible en: https://www.weforum.org/publications/global-risks-report-2024/, p. 8.

is justified by multiple cases in recent history. Whether during the most intense period of Covid-19, Russia's invasion of Ukraine, migratory flows to the United States, the conflict between the terrorist organisation Hamas and Israel, the strategic competition between the United States and China, or the growing political instability and social conflict in Latin America and the Caribbean.

It should be noted that, regardless of the case, disinformation follows a similar pattern: an atypical or crisis situation, in an environment characterised by fierce competition for power, in which multiple actors are interested in winning the hearts and minds of the population-audience. Disinformation has also made it increasingly difficult to learn about a specific situation and, more importantly, to anticipate its outcomes or implications[7].

In this regard, it should be noted that disinformation, amplified in the digital world through AI, is an essential part of what George Kennan called "organised political –warfare" in the mid-20th century[8]. More recently, the combination of disinformation and artificial intelligence has been used as part of a range of tactics in newer and more subtle expressions of conflict or intermediate states of peace and war, commonly referred to as hybrid wars or grey zone confrontations[9]. Broadly speaking, these new confrontations involve systematic efforts to undermine the stability, integrity and permanence of a state through information operations and fake news, coercive diplomacy, legal resources (lawfare), economic infiltration, political subversion, commercial blackmail, cyber attacks, as well as traditional military actions, both conventional and unconventional, including the mobilisation of violent non-state actors such as criminal, subversive, or terrorist groups[10].

7 Li, Cathy & Callegari, Agustina. 2024. "Stopping AI disinformation: Protecting truth in the digital world". World Economic Forum. Disponible en: https://www.weforum.org/agenda/2024/06/ai-combat-online-misinformation-disinformation/.

8 Kennan, George F. 1948. "The inauguration of organized political warfare". Wilson Center. Disponible en: https://digitalarchive.wilsoncenter.org/document/george-f-kennan-inauguration-organized-political-warfare.

9 Arreola García, Adolfo. 2024. "Inteligencia Artificial y Desinformación: Papel de los Conflictos del Siglo XXI". Centro de Estudios Estratégicos del Ejército de Perú, 3 (3), 87-113.

10 Ball, Joshua. 2022. "The changing face of conflict: What is hybrid warfare?". Global Security Review. Disponible en: https://globalsecurityreview.com/hybrid-and-non-linear-warfare-systematically-erases-the-divide-between-war-peace/

THE POTENTIAL OF AI TO COMBAT DISINFORMATION.

It is therefore of the utmost importance for democratic societies to review and rebuild their relationship with the truth, especially in the midst of global geopolitical transitions, global democratic recessions and increasingly polarised societies. In this regard, while it is true that technological advances, particularly AI, have amplified disinformation today, the potential of AI to combat it is highlighted below.

As noted, the proliferation of AI in the digital age has ushered in remarkable innovations and equally unique challenges, particularly in the realm of information integrity[11]. This is especially true when we consider that generative artificial intelligence (GAI) adds a new dimension to the problem of disinformation[12]. Even so, it is equally true that AI plays a major role in countering disinformation[13]. For example, advanced AI-powered machine learning systems can analyse patterns, language use and context to assist in content moderation, fact-checking and the detection of false information[14]. In this sense, AI algorithms make it possible to identify patterns and detect fake news, which helps with fact-checking more efficiently[15].

In this regard, one of the most tangible ways in which AI can combat misinformation is through its information analysis systems. According to a study by the Rady School of Management at the University of California San Diego, machine learning algorithms significantly outperform human judgement when it comes to detecting lies or fake news during high-risk strategic interactions[16].

11 Li & Callegari, 2024.

12 Endert, Julius. 2024. "Generative AI is the ultimate disinformation amplifier". Deutsche Welle. Disponible en: https://akademie.dw.com/en/generative-ai-is-the-ultimate-disinformation-amplifier/a-68593890

13 Moreno Espinosa, Pastora; Alsarayreh, Abdulsalam & Figuereo Benítez, Juan Carlos. 2024. "El big data y la inteligencia artificial como soluciones a la desinformación". Doxa Comunicación, 38.

14 Li & Callegari, 2024.

15 Soprano, Michael; Roitero, Kevin; La Barbera, David; Ceolin, David; Spina, Damiano; Demartini, Gianluca & Mizzaro, Stefano. 2024. "Cognitive Biases in Fact-Checking and Their Countermeasures: A Review". Information Processing and Managment, 61, 1-29.

16 Serra-García, Marta & Gneezy, Uri. 2023. "Improving human deception detection using algorithmic feedback". CESIFO Working Papers, 10518 (2023). Disponible en: https://www.cesifo.org/en/publications/2023/working-paper/improving-human-deception-detection-using-algorithmic-feedback

It should be noted that studies such as these open the door for AI to make significant contributions to combating misinformation, as machine learning could be used to reduce the amount of false content on major social networks, namely Facebook, YouTube, Instagram, and TikTok.

Additionally, although bots have generally become known for their ability to spread misinformation, these tools can also help to counteract disinformation. Specifically, "AI-powered bots play a key role in providing automatic responses to common queries that allow accurate information to be disseminated"[17]. In fact, some governments have already used these chatbots to combat misinformation during the COVID-19 pandemic. It should be noted that Latin American and Caribbean countries could well use this type of resource to mitigate the effects of misinformation in a society particularly affected by this phenomenon.

In this regard, the use of AI to combat misinformation in the region is essential to reverse the democratic erosion that several countries in the Western Hemisphere are experiencing. In this regard, Latin America and the Caribbean are a clear example of the scope and depth of misinformation. A high percentage of citizens in Latin American countries perceive that they are confronted with misleading or false news every day or almost every day. Brazil tops the list with 60%, followed by Ecuador with 58%. Additionally, citizens of Colombia, Chile, and Argentina also experience high levels of exposure to disinformation, with 53%, 52%, and 51%, respectively[18]. According to the same study, Mexico accumulated 43% of its citizens.

17 Alonso-González, Marián & Sánchez Gonzales, Hada M. 2024. "Inteligencia artificial en la verificación de la información política. Herramientas y tipología". Revista Más Poder Local, 56, p. 30.

18 ACTIVA & Worlwide Independent Network of Market Research. 2022. "Fake news: Desinformación en Chile y LatAm". Disponible en: https://chile.activasite.com/wp-content/uploads/2023/04/222287_Estudio-Desinformacio%CC%81n_0329_V1-2.pdf

PUBLIC-PRIVATE INITIATIVES

As noted, Latin America and the Caribbean is one of the regions where fake news and misinformation are most widely consumed[19]. This makes it imperative for the region to adopt concrete measures to combat this scourge which, as has been said, undermines the democratic governance of countries. In this regard, it is worth noting that some companies, such as Google, through its Google News Initiative, are taking steps to develop people's digital skills, especially with the support of the creation of the Latin American Network of Fact-Checking Trainers in 2022. In particular, this is a project whose objective is to provide training in fact-checking so that journalists and university professors in Argentina, Mexico, Colombia and Peru can act as multipliers[20].

On the other hand, various initiatives have focused specifically on addressing the problem of disinformation using artificial intelligence. A notable example is the Coalition for Content Provenance and Authenticity (C2PA), made up of leading technology companies such as Adobe, Arm, Intel, Microsoft, and TruePic. Its main objective is to combat the proliferation of misleading information online by developing technical standards to certify the origin, history, and provenance of multimedia content[21]. Another global initiative is the AI Governance Alliance, a flagship initiative of the World Economic Forum and part of the Centre for the Fourth Industrial Revolution, which brings together experts and organisations around the world to address the challenges of AI, including the generation of misleading or harmful content and the violation of intellectual property rights[22].

19 Duchiade, André. 2023. "Nuevo mapa destaca leyes contra la desinformación en América Latina y el Caribe". LatAm Journalism Review. Disponible en: https://latamjournalismreview.org/es/articles/nuevo-mapa-destaca-leyes-contra-la-desinformacion-en-america-latina-y-mas-alla/

20 Rosenzvit, Milena. 2022. "Lanzamos la Red latinoamericana de formadores en fact checking". Chequeado. Disponible en: https://chequeado.com/lanzamos-la-red-latinoamericana-de-formadores-en-fact-checking/

21 Para más informació sobre la iniciativa, ver: Coalition for Content Provenance and Authenticity. Disponible en: https://c2pa.org/

22 Li & Callegari, 2024.

DISINFORMATION AND POLITICS

The impact of disinformation on politics and institutions deserves special mention, as does how AI can contribute to reversing this problem. Moreover, these indicators reflect another problem associated with fake news and other expressions of disinformation: the disinterest and mistrust with which citizens relate to news reports in traditional media[23]. Of course, a poor flow of information will eventually lead to increased public mistrust of political institutions and electoral processes, essential components of a lasting democracy[24].

This is why AI helps detect and combat disinformation in politics. Advanced algorithms and machine learning techniques make it possible to assess the credibility of sources and predict disinformation[25].

This new reality has even given rise to a new form of algorithmic political communication[26]. That is, the intersection between scientific computing and social sciences, characterised by micro-segmentation, personalised diagnostics, and automated content generation[27]. The impact of this is that it encourages the creation of echo chambers in which people reaffirm their preconceived ideas and facts are altered. To this end, multiple AI tools have been created to counter misinformation on social media[28].

Algorithmic political communication is a very present problem in Latin America and the Caribbean. In fact, Bradshaw and Howard have found evidence of automated and human use of manipulation strategies in various countries in

23 Alonso-González, Marián & Sánchez Gonzales, Hada M. 2024. "Inteligencia artificial en la verificación de la información política. Herramientas y tipología". Revista Más Poder Local, 56, p. 30.

24 Bennet, Lance & Livingston, Steven. 2018. "The disinformation order: Disruptive communication and the decline of democratic institutions". European Journal of Communication, 32 (2), 122-139.

25 Sánchez Gonzales, Hada. 2022. "Transformación digital y audiencia. Tendencia y uso digital en medios verificadores". Ámbitos. Revista Internacional de Comunicación, (56), 9-20.

26 Campos-Domínguez, Eva & García-Orosa, Berta. 2018. "Comunicación algorítmica en los partidos políticos: automatización de producción y circulación de mensajes". Profesional de la Información, 27 (4), 769-777.

27 López-López, Paulo Carlos; Lagares Díez, Nieves & Puentes-Rivera, Iván. 2021. "La inteligencia artificial contra la desinformación: una visión desde la comunicación política". Razón y Palabra, 24 (111), 5-11.

28 Atanasova, Pepa; Simonsen, Jakob Grue; Lioma, Christina & Augenstein, Isabelle. 2020. "Generating fact checking explanations". In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7352–7364.

the region, mainly in Colombia, Ecuador, and Mexico[29]. In fact, cyborgs have been found in Mexico, which consist of bots or automated systems that can simulate human behaviour[30].

However, AI applied to journalism is positioned as a potential remedy to counteract disinformation, at least in three areas: 1) improving the state of professional journalism, 2) managing privacy, and 3) the ethical approach to technological development, which contributes to informed decisions that are free from bias[31]. Thus, AI has the potential to contribute to the development of what Adela Cortina called "media citizenship"[32].

In this regard, some academics have suggested a hybrid approach in which artificial intelligence tools and humans –including experts and non-experts recruited on crowdsourcing platforms– work together to combat the spread of misinformation on the internet[33]. According to these studies, combining these methods can not only result in more efficient and effective data verification processes, but can also lead to greater confidence in the results compared to purely AI-based methods. It can also leverage the integrated human dimension to increase the level of transparency around the classification of fake news.

The contemporary relationship between artificial intelligence, truth and disinformation represents a complex challenge for democratic societies, especially in Latin America and the Caribbean. While AI has amplified the risks of disinformation, it simultaneously offers promising tools to combat it. The Latin American region serves as a laboratory for understanding the impacts of disinformation, not only to ascertain its depth and scope, but also to identify possible measures to counteract it effectively.

29 Bradshaw, Samantha & Howard, Philip N. 2017. Troops, trolls and troublemakers:a global inventory of organized social media manipulation. Oxford: University of Oxford.

30 Barredo-Ibáñez, Daniel; De la Garza Montemayor, Daniel Javier; Torres Toukoumidis, Ángel; López-López, Paulo Carlos 2021. "Artificial intelligence, communication, and democracy in Latin America: a review of the cases of Colombia, Ecuador, and Mexico". Profesional de la Información, 30 (6), 1-17.

31 Manfredi Sánchez, Juan Luis & Ufarte Ruiz, María José. 2020. "Inteligencia artificial y periodismo: una herramienta contra la desinformación". Revista CIDOB d'Afers Internacionals, 124, 49-72.

32 Cortina, Adela. 2004. Ciudadanía mediática. El País. Disponible. https://elpais.com/diario/2004/11/24/opinion/1101250805_850215.html.

33 Demartini, Gianluca; Mizzaro, Stefano & Spina, Damiano. 2020. "Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities". IEEE, 65-75.

The potential of AI to counteract this phenomenon is multidimensional. Machine learning systems can analyse linguistic and contextual patterns more efficiently than human judgement to detect fake news. In addition, promising public-private initiatives are emerging. At the same time, forums such as the World Economic Forum are bringing together global experts to address the ethical challenges of artificial intelligence. A hybrid approach is emerging as an effective strategy: combining AI tools with human intervention, leveraging technological efficiency and the contextualisation skills of experts. This model could increase transparency in the classification of fake news and generate more robust verification processes.

On the other hand, the political dimension of disinformation is crucial. It is undeniable that disinformation not only distorts public perception, but also undermines democratic institutions, erodes public trust, and facilitates manipulation in contexts of high polarisation and competition for power. Therefore, the solution lies not only in technology, but in a comprehensive approach that includes digital education, the development of critical verification skills, and an ethical framework for the use of artificial intelligence.

In short, AI is both a problem and a solution in the fight against disinformation, requiring a sophisticated, interdisciplinary, and collaborative approach to protect the integrity of the democratic foundations of societies in Latin America and the Caribbean, where this scourge is particularly prevalent.

III. ARTIFICIAL INTELLIGENCE AND PUBLIC SAFETY IN LATIN AMERICA

The intersection between artificial intelligence (AI) and public security represents one of the most controversial fields in the implementation of emerging technologies, mainly due to the divergence between the motivations behind its use and the functionality and usability it can bring to the exercise of public security.

On the one hand, there is the aspiration of governments to incorporate technological devices for monitoring their citizens, which, although not new, has found fertile ground for expansion in the capabilities of AI. On the other hand, threats to public security are becoming increasingly dynamic and complex, and the integration of new technologies can contribute significantly to better combating and prosecuting such threats. This section discusses various aspects of how the relationship between these two variables has developed, first in general terms and then in the Latin American context.

PARADIGMS AND APPLICATIONS OF AI IN PUBLIC SECURITY

The integration of AI into the field of public security is manifested in various paradigms that are influencing, or even changing, some of the ways in which policing is conceived and carried out. At the first Global Meeting on the Opportunities and Risks of Artificial Intelligence and Robotics for Law Enforcement, organised by the INTERPOL Innovation Centre and the United Nations Interregional Crime and Justice Research Institute (UNICRI) in 2018, some of these paradigms were discussed, and we can identify at least four of them [34].

The first corresponds to machine learning and deep learning[35] for predictive crime analysis. This paradigm focuses on identifying criminal patterns by processing large volumes of historical data, allowing for the anticipation of possible incidents and the optimisation of police resource allocation.

The second paradigm encompasses intelligent video surveillance systems, which include facial, voice and object recognition capabilities. These systems function as "second eyes" in tasks that require high precision in observation, being able to identify faces even in suboptimal conditions or when they are partially covered. The COVID-19 pandemic significantly boosted the development of these technologies, mainly in India and China, and many of these innovations are being adopted by Latin American police forces.

34 "Artificial Intelligence and Robotics for Law Enforcement". Interpol y United Nations Interregional Crime and Justice Research Institute (UNICRI), 2019. Disponible en: https://unicri.it/sites/default/files/2019-11/Artificial%20Intelligence%20and%20Robotics%20for%20Law%20Enforcement.pdf.

35 Machine learning and deep learning are two types of intelligence classified as narrow or weak AI, the only type of AI that currently exists. Narrow AI specialises in specific tasks and, although it can outperform humans in its defined field, it cannot operate outside of it. Other types of AI, still theoretical, include Artificial General Intelligence (AGI or strong AI), which could match human cognitive ability in any task, and Artificial Superintelligence (ASI), which would exceed human cognitive abilities in all aspects. In terms of functionality current AI can be categorised into reactive AI, which operates only with present data without memory of the past (such as Deep Blue, IBM's chess system), and limited memory AI, which can use historical data to improve its decisions (such as autonomous driving systems). There are also theoretical categories such as AI with Theory of Mind, which would understand human emotions and thoughts, and self-aware AI, which would develop its own form of consciousness.

The third paradigm encompasses exploration technologies, from autonomous vehicles to geospatial analysis. These tools make it possible to cover large or difficult-to-access areas, providing crucial data for real-time decision-making. Geospatial analysis systems integrate multiple layers of information, from population movement patterns to crime rates, allowing for the identification of correlations that might go unnoticed using traditional methods. In the case of autonomous vehicles, mainly drones and exploration robots, these can access high-risk or difficult-to-access areas, collecting visual and sensory information that increases the safety of police personnel and improves the quality of information available for operational decision-making.

The fourth paradigm refers to integrations with communication systems infrastructure and architecture, which are essential for effective coordination between different units and jurisdictions. These systems facilitate not only the transmission of real-time information between different police forces and command centres, but also allow the integration of multiple data sources, from emergency calls to video surveillance feeds. The architecture of these systems incorporates distributed processing and real-time analysis capabilities, allowing critical information to reach response teams in a timely manner. In addition, these systems can be adapted to crisis situations where conventional communications infrastructure could be compromised, ensuring operational continuity in critical situations.

The specific applications of these paradigms are vast and diverse. In the forensic field, these technologies have revolutionised criminal investigation through advanced virtual autopsy tools that help determine causes of death[36], crime scene recreation systems[37], and automated DNA analysis[38]. Urban traffic management has been transformed with systems capable of identifying traffic violations, managing fines automatically and detecting stolen vehicles using computer vision software[39].

36 O'Sullivan, Shane, Andreas Holzinger, Kurt Zatloukal, Paulo Saldiva, Mohammed Imran Sajid, and Dominic Wichmann. "Machine learning enhanced virtual autopsy". Autopsy & case reports 7, no. 4 (2017): 3.

37 Sacco, M. A., P. Tarzia, L. Tarda, R. La Russa, F. Cordasco, and I. Aquila. "The artificial intelligence in autopsy and crime scene analysis". *La Clinica Terapeutica* 175, no. 4 (2024).

38 Alketbi, Salem K. "Emerging Technologies in Forensic DNA Analysis". *Available at SSRN 4956325* (2024).

39 Šusteková, D., and Marta Knutelska. "How is the artificial intelligence used in applications for traffic management". *Machines. Technologies. Materials*. 9, no. 10 (2015): 49-52.

In the field of criminal investigation, AI has proven particularly useful in detecting anomalous patterns in financial transactions, facilitating the identification of illegal activities[40]. Voice recognition and audio analysis systems enable the detection and location of gunshots, providing faster responses to violent incidents[41]. Vehicle licence plate reading and identification technology, one of the first applications of neural networks in security, has evolved to achieve previously unimaginable levels of accuracy[42].

A particularly relevant application is the development of tools for the protection of minors and the detection of online fraud. For example, systems such as PhotoDNA, developed by Microsoft and used by security agencies and digital platforms globally, employ cryptographic hashing and deep learning to identify, track and remove images of child exploitation even when they have been altered. These tools can process millions of images per second and have an accuracy rate of over 99.5% in identifying illegal material[43]. In the field of financial fraud in its many forms, one example is Feedzai, a Portuguese company that uses AI to analyse more than 100 billion transactions annually, detecting patterns of suspicious behaviour in real time and reducing fraud losses by up to 75%. This company is collaborating with Brazilian and Mexican banks, as well as their respective regulatory authorities, to combat this crime[44].

Cryptography-based packet tracing systems represent a significant advance in digital investigation, allowing law enforcement agencies to track illegal activities while preserving the privacy of legitimate users. For example, technologies such as homomorphic encryption allow encrypted data to be analysed without first having to decrypt it. Systems such as CryptoGraph use this technique to monitor network traffic patterns and detect malicious activity, while keeping the content of communications protected. These systems can identify indicators of criminal activity such as anomalous data traffic or suspicious communication patterns,

40 Stears, Chris, and Joshua Deeks. "The use of artificial intelligence in fighting financial crime, for better or worse?". *Journal of Money Laundering Control* 26, no. 3 (2023): 433-435.

41 https://www.soundthinking.com/spanish-lp/

42 Kaur, Parneet, Yogesh Kumar, and Surbhi Gupta. "Artificial intelligence techniques for the recognition of multi-plate multi-vehicle tracking systems: a systematic review". *Archives of Computational Methods in Engineering* 29, no. 7 (2022): 4897-4914.

43 https://news.microsoft.com/es-xl/features/como-photodna-for-video-es-utilizado-para-combatir-la-explotacion-infantil-en-linea/

44 https://www.feedzai.com/resource/how-latam-banks-can-measure-classify-and-prevent-scams/

maintaining a balance between security and privacy with false positive rates below 0.1%[45].

EXPERIENCES IN LATIN AMERICA

The implementation of AI systems in public safety in Latin America presents significant cases that illustrate both the scope and limitations of these technologies. The case of Uruguay (2013-2017) with Predpol is particularly illustrative. This crime prediction software produced detailed maps to optimise police deployment, based on four main variables: type of crime, location, date and time of commission. An evaluation of the system in Montevideo, where half of the police stations used Predpol and the other half used traditional statistical analysis methods, found no significant differences in effectiveness, leading to the suspension of the programme. However, the country continued its commitment to technology, and in 2020 its Senate approved the creation of a facial identification database for public security purposes[46].

Colombia presents two emblematic cases. On the one hand, PRISMA, which is designed to predict the risk of recidivism, represents an attempt to systematise decisions on preventive detention. The system collects comprehensive information on individuals under investigation, including previous arrests, pending proceedings, and prison records. However, its implementation has raised concerns similar to those observed in the United States, where similar systems have shown racial bias in their predictions[47].

On the other hand, the Command, Control, Communications and Computing Centre (C4) in Bogotá represents another notable case. This system integrates 6,000 video surveillance cameras with facial recognition and voice analysis capabilities, allowing criminals to be tracked by filtering specific characteristics in real

45 Wu Xu-dong y Li Xin. "Protect Cloud Computing's Data Using Fully Homomorphic Encryption". National Conference on Information Technology and Computer Science (CITCS 2012). Publicado por Atlantis Press, 2012. Disponible en: http://eprint.iacr.org/2009/571.

46 OECD/CAF. *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe.* Estudios de la OCDE sobre Gobernanza Pública. Paris: OECD Publishing, 2022. https://doi.org/10.1787/5b189cb4-es., 47.

47 Partnership on AI. "Artificial Intelligence Research and Ethics Community Calls for Standards in Criminal Justice Risk Assessment Tools". Publicado el 26 de abril de 2019 en https://partnershiponai.org/resource/artificial-intelligence-research-and-ethics-community-calls-for-standards-in-criminal-justice-risk-assessment-tools/

time. The project seeks to expand through an agreement with the National Registry to access national identification biometric data, although this has raised concerns about privacy and potential discrimination.

In Argentina, the 2019 implementation of a facial recognition system in Buenos Aires, integrated with the existing security camera infrastructure, marked a milestone in the region. The subsequent passage of a local law authorising exceptions for the use of these technologies in public safety matters set an important legislative precedent[48].

Ecuador developed the ambitious ECU 911 system, a national network that integrates 4,300 surveillance cameras and 16 regional centres, with more than 3,000 employees monitoring images and responding to calls. The system incorporates advanced technologies such as thermal cameras for volcanic monitoring, night-vision drones, and an automated platform for sending evidence to courts[49]. However, it has faced significant criticism, particularly after a New York Times investigation revealed that the videos were being shared with the country's national security agency[50].

Brazil has taken a more focused approach, as evidenced by the law passed in Brasilia in November 2020 authorising facial recognition for public security. The justification was based on its proven effectiveness in reducing crime rates during that year's carnival[51].

48 Asociación por los Derechos Civiles. "#ConMiCaraNo: Reconocimiento facial en la Ciudad de Buenos Aires". Publicado el 23 de mayo de 2019. https://adc.org.ar/2019/05/23/con-mi-cara-no-reconocimiento-facial-en-la-ciudad-de-buenos-aires/

49 OECD/CAF. *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe*. Estudios de la OCDE sobre Gobernanza Pública. Paris: OECD Publishing, 2022. https://doi.org/10.1787/5b189cb4-es., 49

50 Mozur, Paul, Jonah M. Kessel y Melissa Chan. "Made in China, Exported to the World: The Surveillance State". *The New York Times*, 24 de abril de 2019. https://www.nytimes.com/es/2019/04/24/espanol/america-latina/ecuador-vigilancia-seguridad-china.html

51 Interlira Reports. "The Implementation of Facial Recognition Cameras for Public Security in Brazil: Benefits and Controversies". Publicado el 4 de mayo de 2022. https://interlira-reports.com/es/news/the-implementation-of-facial-recognition-cameras-for-public-security-in-brazil-benefits-and-controversies/05/04/2022/

REGULATORY FRAMEWORK AND LEGAL CHALLENGES

The regulatory landscape in Latin America is heterogeneous and reflects the different approaches that countries in the region have taken to the challenge of regulating AI in the field of public security. Peru set a milestone by passing the first law on AI in the region in mid-2023, although with limited scope that was attempted to be remedied through regulations that still present significant technical deficiencies for effective implementation[52].

Argentina, Colombia, Brazil, and Mexico have bills under discussion, each with different approaches that highlight the diversity of interpretations of AI and its applications in security. It is noteworthy that only the Brazilian bill specifically addresses applications in public security, leaving a significant gap in the other legislative proposals[53].

The Chilean case deserves special attention due to its particular evolution. After a first failed attempt at regulation in Congress, the executive branch presented its own bill, strongly influenced by economic and public security factors. The proposal not only omits preventive measures to contain potential abuses in the use of automated facial recognition systems, but also explicitly rejects this need by granting exceptional status to the spheres of national and public security. As Chile's Minister of Science, Aisén Etcheverry, stated, it is recognised that although real-time biometric analysis and in public spaces carries high risks of violating fundamental rights, its use is justified in the context of security and the prosecution of organised crime, provided that it is carried out by competent authorities authorised by law.

Most legislative projects in the region follow the risk-based approach established by the European Union's Artificial Intelligence Act of 2023, which classifies AI systems according to their level of risk and imposes stricter regulations on applications considered high risk. However, there is a worrying trend towards the implementation of biometric surveillance systems in public spaces without the necessary transparency, a practice observed in Argentina, Brazil, Ecuador, Uruguay, Colombia, Bolivia, Paraguay, and Mexico.

52 Lara, J. Carlos. "Inteligencia Artificial en América Latina: Regulación no significa protección". *Derechos Digitales*, 7 de junio de 2024. https://www.derechosdigitales.org/23866/inteligencia-artificial-en-america-latina-regulacion-no-significa-proteccion/

53 Pablo Contreras, "Convergencia Internacional y Caminos Propios: Regulación de la Inteligencia Artificial en América Latina", *Actualidad Jurídica Iberoamericana* 21 (agosto 2024): 468-493.

This trend contrasts sharply with the restrictions implemented in other parts of the world. In the United States, cities such as San Francisco, Somerville, and Oakland have banned the use of facial recognition by public security forces. Even leading technology companies such as IBM, Amazon, and Microsoft have called for stricter regulation, recognising the risks inherent in these technologies.

RISKS AND CONCERNS

The implementation of AI in public security faces significant challenges that go beyond technical considerations. A 2024 Fair Trials report on seven Latin American countries documents violations of fundamental rights such as privacy, equality, non-discrimination, presumption of innocence, and due process. These concerns are magnified by the lack of government transparency in initiatives that integrate these systems into criminal justice systems.

A particular phenomenon that characterises the implementation of AI in the region is "technological solutionism", the tendency to try to solve structural problems by implementing technological tools without addressing the underlying causes. This approach has proved particularly problematic in the Latin American context, where deep social inequalities require more holistic approaches. Paradoxically, one of the least explored areas from a public policy perspective is the incorporation of AI solutions for the study of social factors such as poverty, education, and health from a public safety perspective.

Facial recognition systems raise specific concerns that deserve detailed attention. Experts distinguish between different techniques: authentication (one-to-one comparison, as in access control), identification (one-to-many comparison, as in searching for suspects), and classification (analysis of faces for categorisation according to criteria such as gender or ethnicity). Each of these techniques presents particular ethical and technical challenges.

Facial recognition systems raise specific concerns that deserve detailed attention. Experts distinguish between different techniques, each with its own ethical and technical challenges:

Authentication (one-to-one comparison) is mainly used in access control and identity verification. In technical terms, these systems face challenges with image capture quality and consistency, variations in lighting, and the natural ageing of faces. Ethical challenges include the storage and protection of biometric data, informed consent for its collection, and the right to be forgotten. For example, when an employee leaves an organisation, the question arises as to whether to retain or delete their biometric data.

Identification (one-to-many comparison) presents more complex technical challenges due to the need to compare a face against extensive databases. Current systems show significantly higher false positive rates when operating at this scale. A NIST study revealed that algorithms can have error rates up to 100 times higher in certain demographic groups, particularly people of colour and women. Ethical challenges are particularly concerning: the use of these technologies for mass surveillance, potential racial and ethnic discrimination, and the risk of misidentification leading to unjustified arrests.

Classification, which analyses faces for categorisation according to criteria such as gender or ethnicity, presents the most controversial ethical challenges. Technically, these systems can perpetuate and amplify existing biases, especially when trained with datasets that are not representative of the racial and ethnic diversity of Latin America. The accuracy of these systems drops significantly when analysing faces from groups that are underrepresented in the training data. Ethical challenges include the reinforcement of stereotypes, the legitimisation of racial profiling, and the potential use for systematic discrimination.

TOWARDS RESPONSIBLE IMPLEMENTATION

To ensure responsible implementation of AI in public safety, it is essential to establish a framework of robust and enforceable guiding principles. Equity must be a fundamental pillar, ensuring that systems respect basic rights such as due process, presumption of innocence, and freedom of expression. This involves not only the technical design of the systems, but also their implementation and ongoing monitoring.

Accountability must be institutionalised at the organisational level, establishing clear oversight and accountability mechanisms. This requires the creation of specific governance structures for the management of AI systems, with clearly defined roles and responsibilities.

Transparency emerges as a critical principle, especially in the context of public safety. The decision-making processes of systems should not be "black boxes", but rather that allow for audit and scrutiny. This is particularly relevant when decisions directly affect the rights and freedoms of citizens.

The "explainability" of systems must ensure that actions and decisions are understandable not only to technical users, but also to judicial system operators and affected citizens. This involves developing interfaces and documentation that allow for understanding the reasoning behind automated decisions.

Considering the above, for effective implementation, it is recommended to:

1. Establishing human rights impact assessment frameworks prior to the implementation of AI systems in public security.
2. Develop specific protocols for the use of facial recognition and biometric technologies.
3. Implement regular audit systems to evaluate performance and potential biases.
4. Create mechanisms for citizen participation in the oversight and evaluation of these systems.
5. Establish clear limits on the use of personal and biometric data.

ARTIFICIAL INTELLIGENCE AND TECHNOLOGICAL SOLUTIOSM: FINAL CONSIDERATIONS

The integration of AI into public security in Latin America represents a critical crossroads between technological innovation and the protection of fundamental rights. While these technologies offer significant opportunities to improve the efficiency and effectiveness of security forces, their implementation must be carefully balanced with robust safeguards that protect individual rights and prevent abuse.

The experience accumulated in the region demonstrates both the potential and the risks inherent in these technologies. Cases such as Predpol in Uruguay illustrate that the implementation of technological solutions does not always produce the expected results, while experiences such as ECU 911 in Ecuador and C4 in Colombia highlight the need for greater transparency and oversight in the handling of sensitive data.

The region faces the particular challenge of developing regulatory frameworks that allow it to reap the benefits of AI while protecting fundamental rights in a context of pronounced social inequalities. Success in this task will depend on the ability of governments to establish public policies that promote transparency, accountability, and citizen participation in the development and deployment of these technologies.

It is essential to remember that any interference with individual rights through the use of AI must be duly justified, legal, necessary and proportionate. This requires not only adequate regulatory frameworks, but also effective oversight mechanisms and the political will to implement them.

A particular phenomenon that characterises the implementation of AI in the region is "technological solutionism", the tendency to try to solve structural

problems by implementing technological tools without addressing the underlying causes. This approach, coined by Evgeny Morozov, manifests itself when authorities prioritise immediate technological solutions over long-term structural reforms. For example, while cities such as Rio de Janeiro have invested millions in AI-powered video surveillance systems to monitor high-risk areas, social prevention and community development programmes in these same areas receive a fraction of the budget.

Technological solutionism in public safety is evident in three main patterns: First, the implementation of mass surveillance systems without an adequate regulatory framework. For example, the deployment of facial recognition cameras in Buenos Aires was carried out before comprehensive protocols on biometric data management were in place.

The second pattern concerns the adoption of predictive technologies without considering pre-existing social biases. The CompStat 2.0 system, implemented in several Latin American cities, has been criticised for perpetuating patterns of excessive surveillance in marginalised communities, based on historical data that reflects systematic biases in law enforcement.

The third pattern is the lack of integration between technological systems and social programmes. While significant resources are invested in AI tools to detect criminal patterns, there is a gap in the use of these same technologies to identify and address social risk factors.

This approach has been particularly problematic in the Latin American context, where deep social inequalities require more holistic approaches. The region has the highest inequality rates in the world. This reality is reflected in crime patterns closely linked to socioeconomic factors.

Paradoxically, one of the least explored areas from a public policy perspective is the incorporation of AI solutions for the study of social factors such as poverty, education, and health from a public safety perspective. There are promising examples globally that could be adapted to the Latin American context. For example, the Risk Terrain Modelling programme, implemented in cities such as Newark and Minneapolis, uses AI to analyse not only crime patterns, but also environmental and socio-economic factors that contribute to insecurity. This approach has reduced crime in specific areas by up to 35% through targeted interventions that combine surveillance with improvements in infrastructure and social programmes.

Some emerging initiatives in the region are beginning to address this gap. A project carried out by the University of the Amazon is developing an AI system that integrates crime data with social indicators such as school dropout rates,

youth unemployment and access to basic services. This approach makes it possible to identify priority areas for intervention where improvements in social conditions could have the greatest impact on reducing crime.

Overcoming technological solutionism requires a paradigm shift in the implementation of AI in public safety. It is necessary to develop evaluation frameworks that consider not only the technical effectiveness of solutions, but also their social impact and their ability to address structural causes. This involves creating multidisciplinary teams that include not only technology and security experts, but also social development specialists, anthropologists, and community representatives.

IV. CONCLUSION AND FINDINGS

The analysis developed throughout this chapter reveals significant patterns in the implementation of artificial intelligence to address disinformation and public security in Latin America and the Caribbean. Far from offering unambiguous solutions, the evidence suggests that the success of these implementations depends crucially on the institutional capacity, regulatory framework, and sociopolitical context of each country.

A fundamental lesson emerges from the study: the need to overcome the false dilemma between technological innovation and the protection of fundamental rights. The most successful experiences in the region demonstrate that these objectives are not only compatible but mutually reinforcing. The implementation cases analysed indicate that when technology is deployed with adequate safeguards and democratic oversight, it can simultaneously strengthen citizen security and trust in institutions.

The regional landscape outlined in the chapter also reveals an instructive paradox: while some countries have made significant progress in the technical implementation of AI systems, they have often neglected crucial aspects such as citizen participation and transparency. This disparity has resulted in a growing gap between the technological sophistication and social legitimacy of these tools, underscoring the importance of developing more comprehensive approaches.

A particularly relevant finding is the interconnection between the challenges of disinformation and public safety. The Latin American experience shows that these phenomena do not operate in isolation: disinformation can exacerbate security crises, while insecurity can create conditions conducive to the proliferation of false narratives. This interrelationship suggests the need for coordinated strategies that address both issues in a coherent manner.

The analysis carried out in this chapter also reveals a worrying trend: the temptation to import technological solutions without adapting them adequately to the regional context and. The limited effectiveness of some predictive and facial recognition systems suggests that solutions developed for other contexts do not necessarily respond to the social, cultural, and economic realities of Latin America. This observation highlights the importance of developing local capacities for innovation and technological adaptation.

Looking ahead, three fundamental imperatives emerge for the region. First, the need to develop regulatory frameworks that promote innovation while effectively protecting citizens' rights. Second, the importance of strengthening institutional capacities to implement and oversee AI systems. Third, the urgency of fostering an innovation ecosystem that responds to the specific needs of the region.

The evidence presented also suggests the need to rethink the role of civil society, academia, and the private sector in these processes. The most successful experiences have actively involved these actors not only in oversight, but also in the design and implementation of solutions. This participatory approach not only enhances the legitimacy of technological interventions, but also contributes to their effectiveness and sustainability.

Finally, the analysis emphasises that the implementation of AI in these critical areas cannot be separated from broader objectives of social development and democratic strengthening. Technology, however sophisticated, must be conceived as a tool at the service of these fundamental objectives, not as an end in itself. Only through this broader understanding e Latin America truly harness the transformative potential of AI to build safer, more informed and democratic societies.

AUTHORS

> ALAM IQBAL, BADAR

Researcher in the Department of Economics at Stellenbosch University, South Africa. He is a renowned scholar with more than four decades of experience in the fields of international economics and business. He began his career in 1978 as an Assistant Professor in the Department of Commerce at Aligarh Muslim University (AMU), India, where he was promoted to Associate Professor in 1986 and Full Professor in 1998. He served as Dean of the Faculty of Commerce at AMU from 2006 to 2008 and was Acting Vice-Chancellor on two occasions. He has also been a member of the University Court, the Executive Council, and the Academic Council of AMU.

Throughout his career, he has worked as a Visiting Professor at various international institutions, including Claflin University and Kentucky State University in the United States (as a Fulbright Scholar-in-Residence), the University of Rennes 1 in France, the Institute of Developing Economies in Tokyo, the University of South Africa, Vaal University of Technology in South Africa, Maringá University in Brazil, the Beijing Institute of Technology in China, and the University of Warsaw in Poland, among others. He has been a DAAD Fellow in Germany on three occasions, a Ford Foundation Fellow in the United States, and a Fulbright Professor in two separate periods. Additionally, he has served as Adjunct Professor of Economics and Finance at the Monarch Business School in Switzerland.

His academic contributions include the publication of 12 books and more than 138 articles in international journals in countries such as Switzerland, Canada, the United Kingdom, Mexico, and Brazil, as well as 81 articles in highly regarded Indian publications. He has participated in international conferences and colloquia, with particular emphasis on issues related to the BRICS countries. His research areas include international economics, international business, and development economics.

> BERASALUCE IZA, JULEN

Research Professor at the Center for Economic Studies of El Colegio de México (COLMEX) and Advisor to the Institute for Industrial Development and Digital Transformation (INADI). He holds a Bachelor's degree in Economics from the University of the Basque Country, an Erasmus Mundus Master's in Quantitative Economics from Université Paris 1 Panthéon-Sorbonne and Universitat Autònoma de Barcelona, and a PhD in Economics from Universitat Autònoma de Barcelona. He is currently a Research Professor at the Center for Economic Studies of El Colegio de México and is a member of the National System of Researchers. His work focuses on industrial policy and development experiences in Asian countries.

He is the author and editor of several works, including: *Hacia una nación digital. La importancia de la sostenibilidad y la tecnología 5G en el desarrollo industrial*; *De la revolución industrial a la revolución digital. Hacia una agenda digital para México*; *Estado desarrollador. Casos exitosos y lecciones para México*; and *Corea y México. Dos estrategias de crecimiento con resultados dispares*.

> BETANZOS TORRES, EBER OMAR

Lawyer, philosopher, economist, theologian, and information scientist. He holds a PhD in Human Rights and completed a postdoctoral fellowship in Public Governance. He is a specialist in constitutional justice, judicial protection of human rights, international criminal justice, Indigenous peoples' rights, and mechanisms of international cooperation.

He is currently the Coordinator of the Postdoctoral Program in Artificial Intelligence for Public Administration at the National Institute of Public Administration (A.C.). He has held prominent positions in the Mexican public sector, including Secretary of Study and Account at the Supreme Court of Justice of the Nation; Deputy Prosecutor for Human Rights at the Office of the Attorney General; and Undersecretary at the Ministry of Public Administration. He currently serves as Head of the Technical Unit of the Federal Audit Office, where he leads initiatives on government oversight, public integrity, and the strengthening of the rule of law.

He is a member of the National System of Researchers. He has authored books, articles, and studies on public auditing, governmental ethics, human rights, transparency, internal control, and digital governance. His work has been recognized for its contributions to social research and the enhancement of institutional capacities.

> CARBAJAL GLASS, FAUSTO

He is a Senior Policy Advisor at Miranda Partners and a founding partner of Delphi Solutions and Strategic Services, a strategic consulting and scientific research firm focused on political risk and security dynamics in Latin America. He has prior experience in Mexico's public sector, specifically at the Ministry of Foreign Affairs and the Ministry of the Interior. He holds a degree in International Relations from Universidad Iberoamericana, Mexico City, and a Master's in War Studies from King's College London. He completed the courses "Defense Strategy and Policy" at the William J. Perry Center for Hemispheric Defense Studies and "Terrorism and Counterterrorism" at Leiden University, Netherlands.

He is currently pursuing a PhD in Development and National Security at the Mexican Institute for Strategic Studies on Security and National Defense (IMEESDN). He is a fellow of the International Strategy Forum, an initiative of the Eric & Wendy Schmidt Fund for Strategic Innovation. He is also an associate and senior fellow of the Mexican Council on Foreign Relations (COMEXI), a senior research fellow at the Center for Research on Peace Mexico (CIPMEX), as well as a member of the Strategic Hub for Organised Crime Research (SHOC) at the Royal United Services Institute (RUSI), the European Consortium for Political Research (ECPR), and the North American Studies Network of the Center for Research on North America (UNAM).

> CASTAÑÓN IBARRA, ROSARIO

National Researcher and former Director of Technology Transfer at the National Autonomous University of Mexico (UNAM). She is a Chemical Engineer, holds a Master's degree in Engineering (Planning), and a PhD in Organizational Management; all her academic degrees were obtained at UNAM. She is a faculty member at the Institute of Applied Sciences and Technology (ICAT) at UNAM, where she has taught courses related to technological innovation management at the master's and doctoral levels, both at UNAM and at other universities. She has also taught diploma programs and short courses organized by various institutions nationwide.

Her research has focused on topics such as competitive technological intelligence, university-industry linkages, technology transfer, knowledge management, and intellectual property management. In the latter area, from 2007 to 2010, she served as Coordinator of the Ibero-American Network for Strengthening the Culture of Intellectual Property. From August 2008 to February 2012, she was the Director of Technology Transfer at the National Autonomous University of Mexico.

> COELLO COELLO, CARLOS

Member of The National College (El Colegio Nacional) and Researcher at the Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN). Visiting Professor at the Basque Center for Applied Mathematics in Spain. He has taught courses on evolutionary computation, multi-objective evolutionary optimization, programming languages, and engineering optimization at CINVESTAV-IPN at both the master's and doctoral levels. He has also taught courses in Spain, England, Argentina, Chile, India, Bolivia, Colombia, Slovenia, and the United States.

He has presented scientific papers at major conferences specializing in evolutionary computation, such as the IEEE Congress on Evolutionary Computation (CEC), the Genetic and Evolutionary Computation Conference (GECCO), and Parallel Problem Solving from Nature (PPSN), among others. He has extensive experience as an expert consultant in his field. He is a Member of the Foundation Advisory Board of The International AIQT Foundation, which aims to establish an internationally competitive center in the areas of artificial intelligence and quantum technology. In addition, he serves as Scientific Advisor to Complexica in Australia, and was Senior Advisor to the Hunan Zixing AI Research Institute in China (2017-2020), among other roles.

> FIERRO GONZÁLEZ, AMAURY

Specialist in intelligence and geopolitics. He is currently an instructor at Rabdan Academy, the leading educational institution in defense and security in the Middle East, and Director of Learning at Tech Corp. He holds a Master's degree in Intelligence from King's College London and a Bachelor's degree in International Relations from Universidad de las Américas Puebla, with a specialization in counterterrorism from the International Institute of Management in Galilee, Israel.

He has served as a trainer for various armed forces and police agencies across Latin America, as well as a border intelligence analyst in the United States, a military intelligence analyst in the United Kingdom, and an intelligence manager in the private sector in Mexico. He is an associate of the Mexican Academy of International Law, the National Institute of Public Administration, and the Teaching Network of the Institute for Strategic Studies of Guatemala.

> FLORES JIMÉNEZ, DIEGO

Head of the Electronics and Digital Industry Sector at the Secretariat of Economy. He holds a Master's degree in Government and Public Policy from Universidad Panamericana and a Bachelor's degree in International Relations from the Instituto Tecnológico Autónomo de México (ITAM). He specializes in national and regional security, having completed specialized studies at the School of Intelligence for National Security of the National Intelligence Center

> GARCÍA MONTES, MARIEL

Researcher and PhD candidate in Science, Technology, and Society at MIT. She is a specialist in the intersection of digital media, society, and activism, with more than 15 years of experience in the field. She is currently a doctoral student in MIT's Program in Science, Technology, and Society, affiliated with the Berkman Klein Center for Internet and Society at Harvard and the MIT Data + Feminism Lab. Before starting her PhD, she earned a degree in Comparative Media Studies at MIT and served as a research assistant at the Center for Civic Media.

Her interests include youth and media, digital literacy, privacy and surveillance, civic and moral education, participatory processes, and feminism. She is currently researching mutual aid networks and online community platforms. Previously, she worked on strengthening technological capacities as a Junior Program Officer at the Wikimedia Foundation; as a consultant for the Open Knowledge Foundation and Internews; as a mentor at UNICEF; and as a member of SocialTIC, where she contributed to research, training, and community initiatives such as Escuela de Datos. She also worked in communication strategy at Nova México and was an intern at the Youth and Media Lab at the Berkman Center.

Mariel studied Philosophy at UNAM and seeks to generate impact at the intersection of culture, technology, and philosophy, prioritizing action over theory.

> GUILLÉN, DAVID

Consulting and Project Management Manager at Cambiotec, A.C. He holds a degree in Mechanical Engineering and a Master's in Industrial Administration from UNAM. He is a specialist in technological innovation management and an expert in competitive technological intelligence studies. He serves as an evaluator for the Technology Transfer Network of the Pacific Alliance.

He has led and collaborated on consulting projects for international organizations such as the Inter-American Development Bank (IDB), the World Intellectual Property Organization (WIPO), and the Ministry of Industry and Commerce of Panama; and has managed technology valuation and technology transfer processes for institutions such as UNIBE (Dominican Republic) and CICESE in Mexico.

He has developed technology intelligence studies for companies such as Huawei (cybersecurity), Rotoplas (wastewater treatment), and CNH-Industrial (tractor 4.0), as well as for educational institutions including UNAM, CICY, and IPN. He has led teams that have won the ADIAT Innovation Award (2011), Green Tech (2012), the BIT Award (2014), and the National Technology and Innovation Award. He is co-author of the book De emprendedor a Empresario (2014) and of the Technology Valuation Manual for the Network of Technology Transfer Offices (2019).

> JUÁREZ AGUILAR, BEATRIZ

She holds a PhD in Law (UNAM) and a Master's degree in China Studies (El Colegio de México), as well as a Fellowship at Renmin University of China. She has taught at Shanxi University (Shanxi, China) and in the Graduate Law Program at UNAM. She has more than 15 years of experience in multidisciplinary research on China as a normative center and is well acquainted with the Chinese legal, political, and academic systems. She has published articles, book chapters, and presented papers at national and international conferences; she is currently preparing the book Introduction to Law in China. She has an upper-intermediate level of Chinese (HSK4), enabling her to analyze sources in Chinese.

> LINARES SALGADO, JORGE ENRIQUE

He holds a Bachelor's, Master's, and PhD in Philosophy from the Faculty of Philosophy and Letters at the National Autonomous University of Mexico (UNAM). He served as Director of the Faculty of Philosophy and Letters (FFyL), UNAM, from 2017 to 2021. He is a full-time tenured professor in the Department of Philosophy at the Faculty of Philosophy and Letters and a Level II member of Mexico's National System of Researchers.

His research areas include the ethics of science and technology, bioethics, and contemporary ethics. He has received distinctions such as the National Philosophy Prize from the Mexican Philosophical Association, the Mexican Academy of Sciences Award for the best doctoral thesis in the humanities, and the UNAM Distinguished Young Scholar Award in the field of Humanities Research.

> LÓPEZ-PORTILLO ROMANO, JOSÉ RAMÓN

Professor at Oxford, co-founder of the Center for Mexican Studies, and member of the International Science Council. He is a Mexican academic, diplomat, entrepreneur, and public servant. He has served as Deputy Secretary of State in Mexico, Permanent Representative, and Independent Chair of the FAO Council. An expert on the socioeconomic impact of science, technology, and innovation, he has advised the Government of Mexico on these issues, leading efforts behind the UN resolutions on the impact of technological change in the 2030 Agenda.

As a member of the "Group of Ten Experts on the Technology Facilitation Mechanism", he has advised the UN Secretary-General and contributed extensively to the Guidebook and Roadmap toward the SDGs. He is a member of the International Science Council. Most recently, he contributed to the AI Safety Summit at Bletchley Park and to its expert advisory panel. He also leads LobsterTel, a telecommunications company operating in the United States and Mexico.

> LOUISE SMITH, HANNAH

Postdoctoral Researcher at the Center for Advanced Studies in Biomedical Innovation Law (CeBIL) at the Faculty of Law, University of Copenhagen. Dr. Smith focuses on the regulation of new and emerging technologies, including advanced medical informatics, artificial intelligence, and quantum technologies. She holds a DPhil in Socio-Legal Studies from the University of Oxford, as well as an M.St in Socio-Legal Research, a BCL in Law, and a BA in Jurisprudence from the same institution.

With her background in socio-legal studies, Hannah has a particular interest in how laws are shaped by social norms, the role of power in the creation and implementation of law, and the opportunities to leverage the dynamic nature of law to anticipate and adapt to regulatory challenges. Her previous positions include a research fellowship at the University of Western Australia, a stipendiary lectureship at Magdalen College, University of Oxford, and an internship at the UK Cabinet Office. Hannah has published in several prestigious journals, including The Law Quarterly Review, European Data Protection Law Review, and Computer Law and Security Review.

> MINSSEN, TIMO

Professor of Law at the University of Copenhagen (UCPH) and Founding Director of the Center for Advanced Studies in Biomedical Innovation Law (CeBIL). Affiliated Researcher at LML at the University of Cambridge (United Kingdom) and Inter-CeBIL Research Affiliate at the Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics at Harvard Law School (USA). His research, supervision, teaching, and part-time advisory practice focus on Intellectual Property, Competition Law, and Regulation, with a particular emphasis on new technologies, big data, and artificial intelligence in the field of health and the life sciences.

Timo holds a law degree from the University of Göttingen (Germany), as well as postgraduate degrees in biotechnology and intellectual property (LL.M., M.I.C.L., LL.Lic., and LL.D.) from Lund University and Uppsala University. He has been a Global Visiting Professor at the Technical University of Munich (TUM) and a Research Fellow at the universities of Cambridge and Oxford, Harvard Law School, Chicago-Kent College of Law, and the Max Planck Institute for Innovation and Competition. He has also received training within the German judicial system and at the European Patent Office.

Drawing on his academic and practical experience in law firms, courts, and life-science start-ups, Timo serves on several international committees and acts as an advisor to the WHO, WIPO, the European Commission, and various organizations, companies, national governments, and law firms.

> OROPEZA GARCÍA, ARTURO

PhD in Law from the National Autonomous University of Mexico (UNAM). Permanent Researcher at the Institute of Legal Research of UNAM. Current President of the Institute for Industrial Development and Digital Transformation (INADI). Member of Mexico's National System of Researchers (SNI). Founding member of the China-Mexico Studies Center (CECHIMEX) and of the Program of Asian and African Studies (PUEAA) at UNAM. Member of the Editorial Board of China Global Review. Arbitrator of the MERCOSUR Dispute Settlement System appointed by Brazil under the Olivos Protocol.

He has been a lecturer and visiting professor at numerous universities and academic institutions around the world, including: the Chinese Academy of Social Sciences in Beijing, China; the National Law School of India University in Bangalore, India; Chuo University, Japan; Soongsil University, South Korea; Universidade Autónoma de Lisboa, Portugal; Université de Montréal, Canada; University of Johannesburg, South Africa; the Ibero-American General Secretariat (SEGIB) in Spain; the Organization of American States (OAS); the Latin American Integration Association (ALADI) in Uruguay; and the Economic Commission for Latin America and the Caribbean (ECLAC) in Chile, among others.

He is the author and editor of more than 40 books on globalization, specializing in China and Southeast Asia. His most recent publications include: From the Industrial Revolution to the Digital Revolution; Towards a New Era of Digital Industrial Development for Mexico; Nearshoring: The Opportunity for a New Economic and Social Development for Mexico; The USMCA in the Context of the China-United States Confrontation; and Artificial Intelligence: Towards a New Era in the History of Humanity, among others.

> PANG, ZHONGYING

Academic in the humanities and social sciences, teaching Global International Relations, International Political Economy, and Global Economic History at Sichuan University in Chengdu, China. He holds a PhD in Law from Peking University. He is a recognized expert in international politics and currently serves as Professor, Doctoral Supervisor, and Chair of the Institute of Maritime Development at the Ocean University of China. He is also an Associate Professor at the School of Humanities and Social Sciences at Tsinghua University, as well as Professor, Doctoral Supervisor, and Director of the Global Institute in the Department of International Relations at Nankai University.

Previously, he held various positions including Assistant and Professor in the Department of International Economics at Renmin University of China; Assistant Researcher and Deputy Director of the Division of International Economic Relations Research at the Institute of World Economics and Politics of the Chinese Academy of Social Sciences; Second Secretary at the Embassy of China in Indonesia; and member of the boards of organizations such as the United Nations Association of China and the China Reform Forum.

His main areas of research and teaching include global issues and global governance, Asian and African studies, international politics and economics, and diplomatic policy analysis.

> PORSDAM MANN, SEBASTIAN

Assistant Professor at the Center for Advanced Studies in Biomedical Innovation Law. With an academic background in philosophy, neuroscience, and applied ethics (BA and PhD, University of Cambridge) and ongoing doctoral research in international human rights law (DPhil in progress, University of Oxford), he brings an interdisciplinary perspective to his work. His postdoctoral appointments include positions at Harvard Medical School, the University of Oxford, and the University of Copenhagen, supported by a personal fellowship from the Carlsberg Foundation.

His research focuses on exploring the practical potential and regulatory conditions necessary for new technologies—such as blockchain and generative artificial intelligence—to advance scientific progress and support normative goals, particularly in relation to the human right to science. Porsdam Mann is co-author/editor of The Right to Science: Then and Now (Cambridge University Press, 2021) and Scientific Freedom: The Heart of the Right to Science (Rowman & Littlefield, 2023). His academic contributions have appeared in leading journals such as Nature Machine Intelligence, npj Digital Medicine, NEJM AI, PNAS, Philosophical Transactions of the Royal Society, Human Rights Quarterly, Molecular & Cellular Proteomics, and the American Journal of Bioethics. His work has a field-weighted citation impact of 5.55 (SciVal, 2021-2025).

> RUÍZ DURÁN, CLEMENTE

Former Coordinator of the Graduate Program in Economics at UNAM and Advisor to the Institute for Industrial Development and Digital Transformation (INADI). He holds a PhD in Economics from the University of Pittsburgh, Pennsylvania, United States (1975), and a specialization in welfare institutions from the Autonomous University of Baja California (2004). He is a professor at the Faculty of Economics and a master's and doctoral-level lecturer in the Graduate Program of the Faculty of Economics at the National Autonomous University of Mexico (UNAM). He has served as Coordinator of the Urban and Regional Development Field within the Graduate Program in Economics and is a member of the National System of Researchers of the National Council of Humanities, Science and Technology (CONAHCYT).

Throughout his teaching career at the undergraduate, master's, and doctoral levels, he has taught courses such as Mexican Economy I and II; Economic Research and Analysis III and IV (1993); Urban and Regional Economics (2012-present); Doctoral Seminar in Urban and Regional Economics (2004-present); and Industrial Policy Seminar (1998-2003), among many other courses and seminars taught since 1987. Additionally, he has supervised and examined more than 200 theses throughout his career, from undergraduate to doctoral level.

> SAVAGE, SAIPH

Assistant Professor at Northeastern University's Khoury College of Computer Sciences, where she leads the Civic A.I. Lab. Her research focuses on designing, building, and studying public-interest artificial intelligence technologies that empower workers, governments, industry leaders, and NGOs. She has secured more than $1.52 million in funding from institutions such as the National Science Foundation (including the prestigious NSF CAREER Award), Google, Meta, the United Nations, and NATO.

Her work has received global recognition: she was named one of MIT Technology Review's "35 Innovators Under 35" and one of Forbes' "20 AI Leaders in Mexico". UNESCO highlighted her research as among the most influential in AI worldwide. She has received awards at leading human-computer interaction conferences, including ACM CHI, CSCW, and The Web Conference. Her work has been featured in major international media outlets such as BBC, The New York Times, The Economist, Vice, Wired, and Forbes, and appeared on the cover of WIRED en español.

In the realm of public policy and leadership, she advises governments and senators in the United States and Latin America on key AI issues. She is

an expert within the OECD's Global Partnership on AI (GPAI) and has served as a researcher at the Center for Democracy and Technology (CDT) and the Federation of American Scientists (FAS). She is also Co-Chair of the International Panel on the Information Environment (IPIE), an initiative presented at the Nobel Prize Summit.

Dr. Savage earned her master's and PhD in Computer Science from the University of California, Santa Barbara (UCSB), and studied Computer Engineering at UNAM. Previously, she worked at Intel Labs, Microsoft Bing, and Stanford, and collaborated with Carnegie Mellon, the University of Washington, and West Virginia University, designing technology for rural regions.

> SOLLEIRO REBOLLEDO, JOSÉ LUIS

Senior Researcher at the Institute of Applied Sciences and Technology (ICAT) at UNAM. His academic output includes more than 200 articles, book chapters, and conference papers published in international proceedings. He is a member of the National System of Researchers. He served as General Director of Outreach at the National Autonomous University of Mexico (UNAM) from 2008 to 2012.
He has taught undergraduate, master's, doctoral, and continuing education courses at numerous institutions in 17 countries. He has advised private and public companies, universities, international organizations, and business associations on various issues related to innovation management. He coordinated the project on the State Innovation System of the State of Mexico, sponsored by the Inter-American Development Bank and the Council for Science and Technology of the State of Mexico.

He was President of LES Mexico, A.C., the Mexican chapter of the Licensing Executives Society International (LESI). Among the recognitions he has received are the Jesús Silva Herzog Award for Economic Research and the Ernest Feder Research Award. He was awarded an Honorary Doctorate by the Ibero-American Council in Honor of Educational Excellence. In April 2017, he received the distinction "Outstanding Mexican Career" as part of the Innovation Match 2016-2017 Awards for Mexican Talent in Innovation.

> SONG, JIN

Director of the Development Research Center at the Institute of World Economics and Politics (IWEP) and Researcher at the Chinese Academy of Social Sciences (CASS). She holds a PhD in Economics and serves as a Senior Researcher and Director of the Development Research Center at IWEP, CASS, which she joined in 2011. Her research areas include development economics and labor economics, with a focus on the labor market, income inequality, and the transition of industrial structure.

She has published numerous articles in leading journals such as China Agricultural Economic Review, Journal of World Economy, Journal of Quantitative & Technical Economics, Management World, and Studies in Labor Economics. She has also contributed to several books published by Oxford University Press, Cambridge Press, and World Scientific.

She is an economic consultant for the World Bank, collaborating on major research projects including China 2030, Urbanization in China, and Options for Elderly Care in China.

> SUDOH, OSAMU

Professor at Chuo University and Professor at the Graduate School of Interdisciplinary Information Studies (GSII) at the University of Tokyo. He earned his PhD in Economics from the same institution. Throughout his career, he has held significant academic and research positions, including his appointment as Professor at the Stockholm School of Economics, Sweden (1995); Research Professor at the NTT Research Institute (1997-1999); and Dean of the GSII at the University of Tokyo (2012-2015).

His research areas include social informatics, medical informatics, and applied economics. Since 2018, he has represented Japan on the OECD's AI expert group (AIGO). He has also been invited to serve on various governmental committees and advisory councils focused on public policy development. Since 2016, he has chaired the "Conference on a Networked Society with AI", organized by Japan's Ministry of Internal Affairs and Communications.

> TSURU, KIYOSHI

Leading attorney in the field of intellectual property, specializing in innovation and cyberspace. He is the founder and director of TMI Law, where he advises companies and organizations on the protection and management of intangible assets in the digital environment. His expertise includes domain name dispute resolution, copyright protection in digital contexts, software and patent protection, geographical indications, electronic transactions related to intellectual property rights, and arbitration.

He also serves as Country Manager for BSA | The Software Alliance, representing the organization in Mexico and promoting policies and strategies for software protection and intellectual property rights enforcement. Additionally, he is the Mexico representative for the Entertainment Software Association (ESA), acting as a liaison with the digital entertainment industry, ensuring regulatory compliance and promoting intellectual property protection initiatives within the sector.

His academic background includes an LL.M. in Intellectual Property Law from George Washington University; the Internet Law Program at Harvard Law School's Berkman Center for Internet & Society; the CIDE-Yale U.S.-Mexico Leadership Program; a specialization in Obligations and Contracts from the Escuela Libre de Derecho (ELD); and a Bachelor's degree in Law from Universidad Iberoamericana. His areas of specialization include intellectual property and digital law, domain name dispute resolution, digital copyright protection, software and patent protection, geographical indications and trademarks, electronic transactions and digital contracts, as well as arbitration and mediation in intellectual property.

> YADAV, ARTI

Assistant Professor in the Department of Commerce and Management Studies at Ramanujan College, University of Delhi. She completed her Postdoctoral Fellowship at the Indian Council of Social Science Research (2019-2021), her PhD in Commerce at Aligarh Muslim University (2014-2017), her Master's in Commerce at the University of Delhi (2013), and her Bachelor's in Commerce at Daulat Ram College, University of Delhi (2011).

Her teaching areas include International Business, Economics, and Corporate Law, while her research focuses on international trade, foreign direct investment, macroeconomic variables, and economic growth and development.

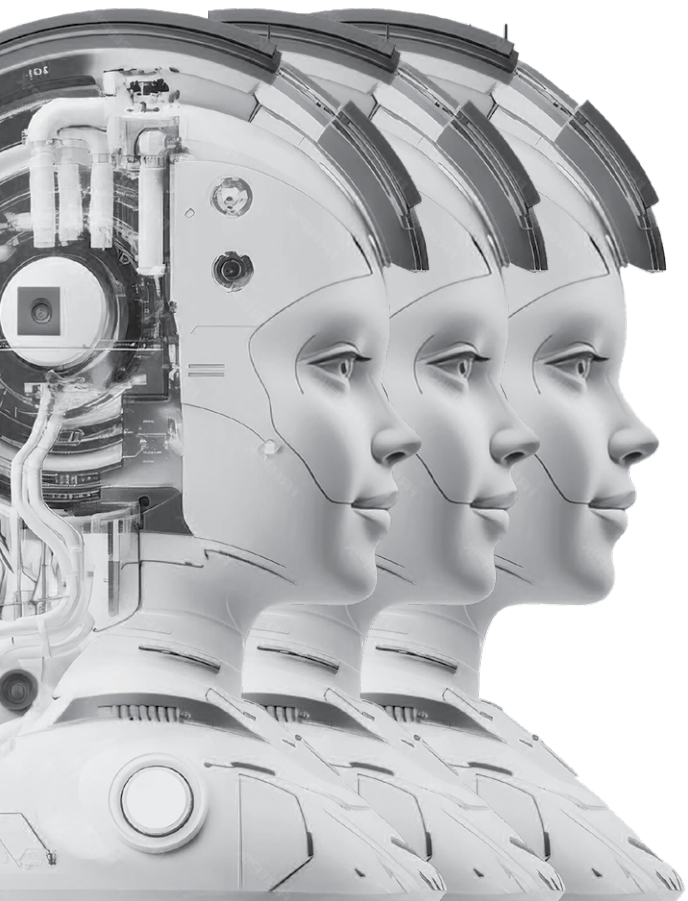
> ZHOU, HUI

Deputy Director and Associate Professor, Department of Cyber and Information Law, Institute of Law at the Chinese Academy of Social Sciences (CASS). His academic work focuses on the intersection of law and information technologies, contributing to the development of policies and regulations in the cyber domain in China.

> ZREIK, MOHAMAD

Postdoctoral Researcher at the School of International Studies, Sun Yat-sen University. He holds a PhD in International Relations from Central China Normal University (2021). His research areas include China's foreign policy, the Belt and Road Initiative, Middle Eastern studies, China-Arab relations, East Asian affairs, Eurasian geopolitics, and political economy. He serves as a reviewer for several academic journals and is a member of the Board of Directors of the Center for Asian and Chinese Studies.

He served as the leader of the Lebanese community in Wuhan (October 2020-October 2021) and was a member of the International Federation of Arab Journalists and Writers, Friends of China. His research and articles have been published in high-impact journals and renowned international media outlets. His writings have been translated into multiple languages, including French, Arabic, Spanish, German, Albanian, Russian, Bosnian, and Bulgarian.

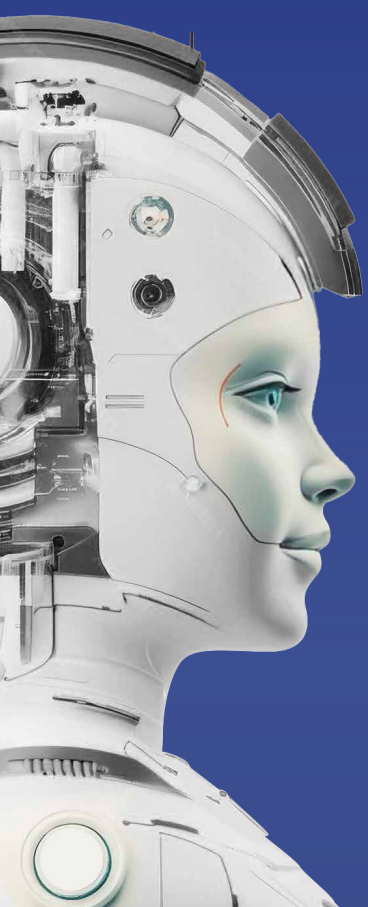


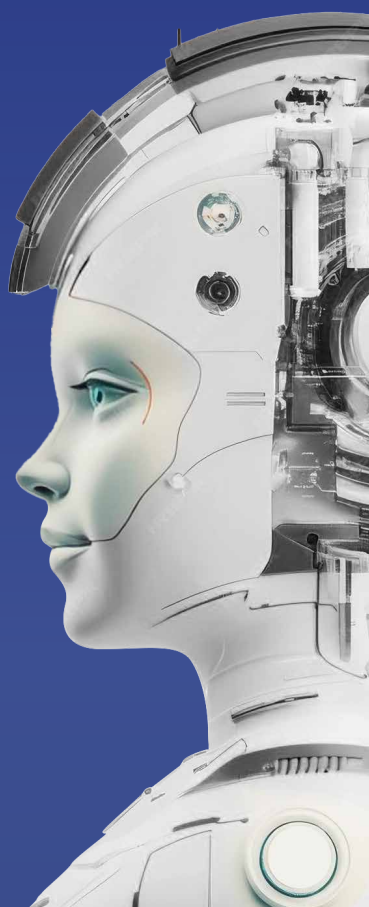
Artificial Intelligence:

Towards a new era of human civilization

Artificial Intelligence has become ubiquitous. Its reach knows no bounds, advancing at such speed that it now threatens to outpace human understanding itself. The vast benefits it promises are only beginning to come into focus, and the world will not renounce them in the pursuit of a better future. However, it is equally clear that Artificial Intelligence poses serious risks to equitable and sustainable development and that, left unchecked, it can significantly erode human liberty.

The future shaped by Artificial Intelligence should not be reduced to a debate between Technophiles and Luddites; it is, and it must be, far more than that. It is an opportunity to ensure the responsible use of the most powerful tool humanity has yet created, in service of building fairer and more sustainable societies as they advance further into the twenty-first century. How we meet this moment will define not only the trajectory of technology, but the character of the world it leaves behind.





ISBN: 978-607-99893-3-0

9 786079 989330