HARVARD Kennedy School
CARR CENTER
for Human Rights Policy

FALL 2024
ISSUE 2024-06

CARR CENTER FOR HUMAN RIGHTS POLICY
HARVARD KENNEDY SCHOOL

Artificial Intelligence: Promises, Risks, and Regulation: Something New Under the Sun



Luís Roberto Barroso

Carr Center
Discussion Paper

ARTIFICIAL INTELLIGENCE: PROMISES, RISKS, AND REGULATION: SOMETHING NEW UNDER THE SUN

Carr Center for Human Rights Policy
Harvard Kennedy School, Harvard University
December 16, 2024 | Issue 2024-06
Publication design by Kyle Faneuff

Luís Roberto Barroso
Chief Justice, Brazil Supreme Court; Senior Fellow, Carr Center

The views expressed in the Carr Center Discussion Paper Series are those of the author(s) and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. Faculty Research Working Papers have not undergone formal review and approval. Such papers are included in this series to elicit feedback and to encourage debate on important public policy challenges. Copyright belongs to the author(s). These papers may be downloaded for personal use only.

I. Author's Note

There is something new under the sun.[1] Many of our beliefs and certainties may be coming to an end. Like the ancient navigators contemplating the immensity of the oceans, full of promises, mysteries, and dangers, we are once again facing the unknown. There is a feeling in the air that a major transformation is about to take place. A revolution, perhaps. Something grand, like the invention of the printing press, which expanded human knowledge exponentially, or the Enlightenment, which reshaped social life, culture, and politics.[2] The future has never felt so imminent and unpredictable.[3]

Given the seemingly infinite possibilities of technology, there is only one reliable nautical chart: the values that have long guided the progress of civilization and the evolution of the human condition on Earth. Whether secular or mythical, they originate from Greece and extend through the Torah, the Gospels, Buddha, Thomas Aquinas, Kant, and many others who have shaped the ethical heritage of humanity. But there is a dramatic point here: the dizzying scientific progress that we have witnessed over the centuries has not been matched by a corresponding ethical—and even spiritual—evolution in the human condition. Goodness, true justice and solidarity are often neglected in a world plagued by extreme poverty, unjust inequality, wars, and a domestic and international order in which some win all and others always lose. It is against this backdrop that we face the topic of Artificial Intelligence (AI) and its potential to make the world better, worse, or even to lead to its annihilation.[4]

Perhaps no topic in the history of civilization has sparked such widespread and simultaneous reflection. In the media, in pubs, in universities, at major international events, and in expert meetings, one discussion subject has become omnipresent: AI. No aspect of its implications has gone unexamined, engaging both the brightest minds and the most ordinary citizens. This



World leaders at the 2023 AI Safety Summit held in Bletchley Park, Milton Keynes, United Kingdom.

text is part of a wealth of writings that aim to capture the spirit of the times, chart a course, and steer history in the right direction. It seeks to avoid the pitfalls that could endanger, if not our very lives, then at least humanity as we know it. Faith in science, like every belief, must not lead to fanaticism. We need to set courses and limits. Here is yet another attempt to do so.

This article is structured as follows. The introduction provides basic notions about the topic. Part I explores the positive potential of AI. Part II catalogs the main risks involved. Part III identifies principles that should govern AI regulation. Finally, the conclusion aims to alleviate anxieties about the future.

[1] There is a well-known biblical passage, which reads: "What has been, will be again, what has been done will be done again. There is nothing new under the sun." Ecclesiastes 1:9. The meaning of this phrase is that ages pass and human beings struggle with the same existential questions. Perhaps some new questions are emerging instead.

[2] See Henry Kissinger, Eric Schmidt, & Daniel Huttenlocher, ChatGPT Heralds an Intellectual Revolution, WALL ST. J. (Feb. 24, 2023), https://www.wsj.com/articles/chatgpt-heralds-an-intellectual-revolution-enlightenment-artificial-intelligence-homo-technicus-technology-cognition-morality-philosophy-774331c6.

[3] This unpredictability can be explained, on one hand, by the progressive autonomy acquired by solutions based on machine learning and, on the other, by the accelerated nature of innovations in this field, which follow one another (or renew themselves) at a pace difficult to fully understand. Nina Schick observes that there were four centuries between the invention of the printing press and the development of photography, for example, but that in just three decades we have gone from the advent of the internet to smartphones, and from there to the platformization of lives on social media, with serious implications for the information landscape. According to Schick, such rapid changes in vital segments entail a high level of uncertainty, which must be taken into account by society as a whole. NINA SCHICK, DEEP FAKES AND THE INFOCALYPSE: WHAT YOU URGENTLY NEED TO KNOW 11 (2020).

[4] AI research avoids alarmism, which does not mean that potentially catastrophic developments are not considered as serious hypotheses. See generally, Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control (2019) (discussing at length the threats and dangers of AI technologies). The most relevant concerns revolve around artificial general intelligence, also referred to as artificial superintelligence. Artificial general intelligence is a state in which computers would surpass human capabilities to a prominent extent, leading, according to Kai-Fu Lee, to "control problems" and "alignment problems." See KAI-FU LEE, INTELIGÊNCIA ARTIFICIAL: COMO OS ROBÔS ESTÃO MUDANDO O MUNDO, A FORMA COMO AMAMOS, NOS RELACIONAMOS, TRABALHAMOS E VIVEMOS 159–173 (2019).

"Currently, AI lacks self-awareness, discernment of right or wrong, and emotions, feelings, morality, or even common sense."

INTRODUCTION: THE DAWN OF THE FOURTH INDUSTRIAL REVOLUTION

1. A BRAVE NEW WORLD

A new Industrial Revolution looms on the horizon. The first took place in the mid-18th century and was characterized by the use of steam as a source of energy. The second Industrial Revolution, spanning the late 19th and early 20th centuries, was defined by electricity and the internal combustion engine. The third, in the late 20th century, saw digital technology replacing analog, ushering in the Technological or Digital Revolution, marked by the universal use of personal computers, smartphones, and the Internet, connecting billions worldwide.[5] Now, the fourth Industrial Revolution is beginning to permeate our lives, combining AI, biotechnology, and the pervasive use of the Internet. This creates an interconnected ecosystem that encompasses people, objects, and even pets, forming an Internet of Things and Senses.[6]

In this unraveling world, new technologies can perform our simplest day-to-day activities as well as highly complex tasks. They can clean rooms, regulate temperatures, and soon may drive cars autonomously.[7] These technologies promise to reenable lost bodily movements,[8] provide more accurate medical diagnoses,[9] overcome neurological deficiencies, expand cognitive abilities, [10]create a person's "virtual twin,"[11] and even reproduce a deceased individual[12] and allow people to reunite with loved ones who have passed away.[13] They offer care for the elderly,[14] help find friends or ideal romantic partners,[15] write texts in diverse languages,[16] distribute welfare aid to the most vulnerable, and deliv-

[5] See Luís Roberto Barroso, Revolução Tecnológica, Crise Da Democracia e Mudança Climática: Limites Do Direito Num Mundo Em Transformação, 5 REVISTA DE ESTUDOS INSTITUCIONAIS 1234, 1262 (2019) (Braz.).

[6] Id. at 1277.

[7] James Manyika, Getting AI Right: Introductory Notes on AI & Society, 151 DAEDALUS 5, 12 (2022).

[8] Jamie Ducharme, New Technique Restores Motion and Feeling to Paralyzed Man, TIME (July 28, 2023), https://time.com/6298543/paralysis-reversal-keith-thomas/.

[9] Steven E. Dilsizian & Eliot L. Siegel, Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment, 16 CURR. CARDIOL. REP. 1, 1–8.

[10] Albrecht Schmidt, Augmenting Human Intellect and Amplifying Perception and Cognition, 16 IEEE PERVASIVE COMPUT. 6, 6–10 (2017).

[11] Mindbank.ai, (2024), https://www.mindbank.ai/ (last visited Jan. 17, 2024). The site invites users to virtually duplicate themselves in order to augment "productivity, mental wellness, personal development, and longevity."

[12] Vanessa Bates Ramirez, Grief Tech Uses AI to Give You (and Your Loved Ones) Digital Immorality, SINGULARITY HUB (Aug. 16, 2023), https://singularityhub.com/2023/08/16/grief-tech-uses-ai-to-give-you-and-your-loved-ones-digital-immortality/. The aim of the system is to replicate users' personalities, their way of thinking, speaking, and other characteristics, so that they can even interact with loved ones after the duplicated person has died.

[13] See Here.After, https://www.hereafter.ai/ (last visited Jan. 17, 2024) ("Your stories and voice. Forever. Preserve memories with an app that interviews you about your life. Then, let loved ones hear meaningful stories by chatting with the virtual you.").

[14] Jason Horowitz, Who Will Take Care of Italy's Older People? Robots, Maybe, N.Y. TIMES (Mar. 27, 2023), https://www.nytimes.com/2023/03/25/world/europe/who-will-take-care-of-italys-older-people-robots-maybe.html.

[15] See, e.g., InnerCircle, https://theinnercircle.co/ (last visited Jan. 17, 2024); Tinder, https://tinder.com/ (last visited Jan. 17, 2024).

[16] ChatGPT, https://chat.openai.com/auth/login (last visited Jan. 17, 2024).

er essential public services to underprivileged areas.[17] Additionally, they can predict crime and recidivism,[18] improve environmental monitoring, help plan smart cities,[19] estimate job applicants' performance, assess the likelihood of loan defaults, and foresee the development of serious illnesses,[20] among other issues.[21]

Further, it is estimated that these technologies can: reveal a person's sexual orientation;[22] foresee and report the intention to have an abortion;[23] replace hundreds of extras and actors in Hollywood;[24] create or eliminate thousands of mechanical or creative jobs;[25] manipulate or falsify information, sounds, images, beliefs, and desires;[26] generate addictions;[27] interfere with consumer behavior;[28] influence the outcome of electoral processes;[29] provoke violent behavior;[30] strengthen extremist agendas;[31] aggravate in-

[17] Sonia K. Katyal, Democracy & Distrust in an Era of Artificial Intelligence, 151 DAEDALUS 322, 327, 322–334 (2022); see also René Urueña, Regulating the Algorithmic Welfare State in Latin America, 2023-27 MAX PLANCK INSTITUTE RESEARCH PAPER SERIES 1 (2023) (discussing what the author describes as "the rise of the algorithmic welfare state in Latin America").

[18] Virgínia Eubanks, The Allegheny Algorithm, in AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 272 (2018); Pranshu Verma, The Never-Ending Quest to Predict Crime Using AI, WASH. POST (July 15, 2022), https://www.washingtonpost.com/technology/2022/07/15/predictive-policing-algorithms-fail/; Katyal, supra note 20 at 327; Richard M. Re, et al., Developing Artificially Intelligent Justice, 22 STAN. TECH. L. REV. 242, 243–244 (2019).

[19] Victor Galaz, et al., Artificial Intelligence, Systemic Risks, and Sustainability, 67 TECH. SOC. 1, 2 (2021), https://doi.org/10.1016/j.techsoc.2021.101741.

[20] Jake Silberg & James Manyika, Notes from the AI Frontier: Tackling Bias in AI (and in Humans), McKinsey Global Institute (June 6, 2019), https://www.mckinsey.com/featuredinsights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans#.

[21] The positive outcomes are indeed impressive, leading some to consider the use of new technologies to transform governance mechanisms in favor of establishing a supposedly neutral and effective "algorithmic democracy." However, there is no such thing as algorithmic neutrality, and democratic legitimacy is necessarily linked to representation based on popular will. Along these lines, the Parliamentary Assembly of the Council of Europe believes that the definition of political and social objectives cannot be relegated to algorithms. On the contrary, it must remain in the hands of human beings who submit to a system of political and legal accountability. See European Union Council of Europe, Artificial Intelligence and Electoral Integrity, EUROPEAN CONFERENCES OF ELECTORAL MANAGEMENT BODIES (2022), https://www.coe.int/en/web/electoral-management-bodies-conference/concept-paper-2022.

[22] Sara Morrison, This Outed Priest's Story Is a Warning for Everyone About the Need for Data Privacy Laws, VOX (July 21, 2021), https://www.vox.com/recode/22587248/grindr-app-location-data-outed-priest-jeffrey-burrill-pillar-data-harvesting.

[23] Joseph Cox, Data Broker Is Selling Location Data of People Who Visit Abortion Clinics, VICE (May 3, 2022), https://www.vice.com/en/article/m7vzjb/location-data-abortion-clinicssafegraph-planned-parenthood; Zeynep Tufekci, We Need to Take Our Privacy Back, N.Y. TIMES (May 19, 2022), https://www.nytimes.com/2022/05/19/opinion/privacy-technology-data.html.

[24] Lois Beckett & Kari Paul, "Bargaining for Our Very Existence": Why the Battle over AI Is Being Fought in Hollywood, THE GUARDIAN (July 22, 2023), https://www.theguardian.com/technology/2023/jul/22/sag-aftra-wga-strike-artificial-intelligence.

[25] Manyika, supra note 10 at 20.

[26] Philipp Hacker, Andreas Engel, & Marco Mauer, Regulating ChatGPT and Other Large Generative AI Models, 2023 ACM INT'L CONF. PROC. SERIES 1–2.

[27] Shabina Mohammad, Raghad A. Jan, & Saba L. Alsaedi, Symptoms, Mechanisms, and Treatments of Video Game Addiction, 15 CUREUS (2023); Beckett and Paul, supra note 27.

[28] Ali Makhdoumi, How AI Could Potentially Manipulate Consumers, DUKE FUQUA SCHOOL OF BUSINESS (Jan. 10, 2024), https://www.fuqua.duke.edu/duke-fuqua-insights/how-ai-could-potentially-manipulate-consumers.

[29] Jonathan Heawood, Pseudo-Public Political Speech: Democratic Implications of the Cambridge Analytica Scandal, 23 INF. POLITY 429, 84–89 (2018).

[30] Eleanore Pauwels, Artificial Intelligence and Data Capture Technologies in Violence and Conflict Prevention, GLOBAL CENTER ON COOPERATIVE SECURITY 1 (2020), https://www.globalcenter.org/wp-content/uploads/GCCS_AIData_PB_H-1.pdf.

[31] SCOTT H VLACHOS, The Link Between Mis-, Dis-, and Malinformation and Domestic Extremism: A Comparative Analysis of the United States and Germany (2022).

equality and discrimination against minority groups;[32] alter and acquire self-will;[33] activate weapons of mass destruction; and put people's lives, health, and safety at risk.[34]

The list is endless and can lead us to both the sublime and the horrific, to freedom or slavery. It can result in the broad affirmation of human rights or their suppression. Intuitively, the problem lies not with the technology itself, but in how we use it and, more importantly, in how we intend to distribute the benefits it will generate. Hence, the challenge ahead is to create an institutional design that promotes the beneficial uses of AI and limits its abuses, preventing the automated production of injustice[35] and the multiplication of existing risks.[36]

2. WHAT IS ARTIFICIAL INTELLIGENCE?

Simply defined, Artificial Intelligence consists of software that transfers human capabilities to computers. These capabilities involve cognitive tasks and decision-making, usually based on the data, instructions, and goals with which they are provided.[37] However, there is no consensus on the specific technical concept of AI and its scope.[38] Numerous organizations and institutions, such as the Organisation for Economic Co-operation and Development (OECD)[39] and the United Nations Educational, Scientific, and Cultural Organization (UNESCO),[40] are attempting to define its boundaries. Amid these efforts, some common features can



The Place de Fontenoy in Paris, home to UNESCO's Headquarters.

be identified. AI systems process data and information similarly to human intelligence, including learning, reasoning, perception, and communication through language. Once requested, ChatGPT4 provided the following definition:

> "Artificial intelligence (AI) is a field of computer science focused on creating systems and machines capable of performing tasks that would typically require human intelligence. These tasks include learning (the ability to improve performance with experience), reasoning (the ability to solve problems using logical methods), perception (the ability to interpret sensory data to understand aspects of the world), and linguistic interaction (the ability to understand and produce natural language)."

[32] Julia Angwin, et al., Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased against Blacks, PR PUBLICA (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Eubanks, supra note 21.

[33] Some fear that AI's capacity for autonomous learning could lead it to acquire superhuman intelligence, making it uncontrollable. This phenomenon is called "singularity." Matthew Hutson, Can We Stop Runway A.I.? Technologists Warn About the Dangers of the So-Called Singularity. But Can Anything Actually Be Done to Prevent It?, THE NEW YORKER, 2023, https://www.newyorker.com/science/annals-of-artificial-intelligence/can-we-stop-the-singularity.

[34] Manyika, supra note 10 at 21.

[35] SARA DEGLI-ESPOSTI, LA ÉTICA DE LA INTELIGENCIA ARTIFICIAL 10 (2023).

[36] MARK COECKELBERGH, ÉTICA NA INTELIGÊNCIA ARTIFICIAL 167 (2024).

[37] J. McCarthy, et al., Dartmouth Summer Research Project on Artificial Intelligence, A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE 1 (1956), http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf; Manyika, supra note 10 at 15.

[38] Organizations representing AI companies posit the formulation of a more restrictive concept of AI, while human rights advocacy organizations support extending the concept to other technologies, which can also have adverse effects on human rights. In such a context, the very scope of the AI concept depends, in part, on how much it should regulate. See, e.g., Tambiana Madiega, Brief Legislation in Progress, EPRS – EUROPEAN PARLIAMENTARY RESEARCH SERVICE 6–8 (2023), https://epthinktank.eu/eu-legislation-in-progress/ (last visited July 29, 2024).

[39] "An AI system is a machine-based system that, for explicit or implicit purposes, infers, from the information it receives, how to generate results such as predictions, contents, recommendations or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptability after deployment." OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 (2019), https://oecd.ai/en/wonk/documents/g20-ai-principles.; Stuart Russell, Karine Perset, & Marko Grobelnik, Updates to the OECD's Definition of an AI System Explained, OECD AI: POLICY OBSERVATORY (2023), https://oecd.ai/en/wonk/ai-system-definition-update (last visited Jan. 17, 2024).

[40] "Therefore, this recommendation approaches AI systems as systems which have the capacity to process data and information in a way that resembles intelligent behavior, and typically includes aspects of reasoning, learning, perception, prediction, planning or control." UNESCO, Recommendation on the Ethics of AI, UNESCO 1 (2021), https://unesdoc.unesco.org/ark:/48223/pf0000380455.

"Intuitively, the problem lies not with the technology itself, but in how we use it and, more importantly, in how we intend to distribute the benefits it will generate. Hence, the challenge ahead is to create an institutional design that promotes the beneficial uses of AI and limits its abuses, preventing the automated production of injustice and the multiplication of existing risks."

—

Luís Roberto Barroso
Senior Fellow at Harvard Kennedy School.

Currently,[41] AI lacks self-awareness, discernment of right or wrong, and emotions, feelings, morality, or even common sense. In other words, it is entirely dependent on human intelligence to guide it, including ethical values.

Computers have no will of their own.[42] Although this is the conventional understanding, some experiments reveal a surprising capacity for learning, giving rise to new concerns. An example is Alpha Zero, an AI program developed by Google that defeated Stockfish, the world's most powerful chess software at the time.[43] Unlike earlier programs, Alpha Zero was not fed with human-conceived moves.[44] Instead, it was only given the rules of the game and it trained by playing against itself, developing original and unorthodox moves and strategies with its own logic.[45]

Two strands of thought have disputed for primacy in AI research over the years. The first was inspired by *the way the human mind works*, aiming to mimic how we formulate questions and develop logical reasoning. This strand dominated AI experiments until the 1980s. The second was inspired by *the way the structures of the human brain work*. It connected information units similar to neurons to simulate brain function.[46] The latter approach, known as the "connectionist approach," has become the dominant one. It does not attempt to reproduce human thinking, but instead seeks to establish correlations and patterns among thousands of data points and outcomes, primarily modeled on statistics and neuroscience.[47]

AI systems are based on data and algorithms. The larger the data set they have access to, the more correlations can be confirmed and discarded, leading to more accurate results.[48] A given set of data or correlated characteristics allows an AI model to recognize a dog or a cat, a good or bad debtor, a person with depressive tendencies, or a child at risk. Establishing correlations between these elements may seem random or irrational to the human mind. However, the model is based on statistics, not logic.

"Computers have no will of their own. Although this is the conventional understanding, some experiments reveal a surprising capacity for learning, giving rise to new concerns."

An algorithm is a fundamental concept in computer science, referring to a set of instructions, rules, and parameters that guide computers in performing assigned tasks. These formulas, codes, and scripts select, process, and store data to achieve specific results. The selected data (inputs) and their process lead to the desired results (outputs), which can vary widely. For example, if the result differentiates objects and living beings, it is a discriminative AI; if it predicts consumer, financial, or political behavior, it is a predictive AI; if it generates content, such as texts, images, or sounds, it is a generative AI.[49]

[41] This caveat is necessary considering that future AI could potentially grant machines significant levels of autonomy and consciousness, creating a scenario where intelligent applications develop their own rationality and pursue unforeseen objectives. See DEGLI-ESPOSTI, supra note 38 at 10; see also LUCRECIO REBOLLO DELGADO, INTELIGENCIA ARTIFICIAL Y DERECHOS FUNDAMENTALES 24 (2023).

[42] Patrick Henry Winston, Artificial Intelligence Desmystified (2018). (mimeographed). Draft of September 30, 2018, kindly sent by the author, p. 2: "They (the programs) do not perceive like us and do not think like us; in fact, they think nothing." See Mariana Lenharo, AI Consciousness: Scientists Say We Urgently Need Answers., NATURE (2023), ature.com/articles/d41586-023-04047-6.

[43] HENRY KISSINGER, ERIC SCHMIDT, & DANIEL HUTTENLOCHER, THE AGE OF AI AND OUR HUMAN FUTURE 7–26 (2021).

[44] *Id.*

[45] *Id.*

[46] Hubert L. Dreyfus & Stuart E. Dreyfus, Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint, 1 DÆDALUS 15 (1988).

[47] *Id.*

[48] *Id.*

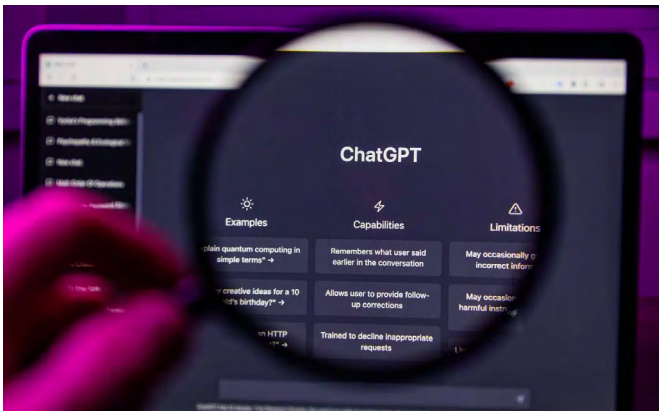
[49] HACKER, ENGEL, & MAUER, supra note 29 at 1–3.

3. MACHINE LEARNING, FOUNDATION MODEL, AND OTHER RELEVANT CONCEPTS

The most advanced AI systems today are those capable of developing machine learning. Machine learning refers to a model's ability to acquire knowledge autonomously, without prior explicit coding, based on identifying correlations within large amounts of data. This capacity for machine learning is what differentiates AI from mere automation in narrower concepts of AI.[50] Machine learning underpins most of the AI services we use today, such as content recommendation systems on platforms like Netflix, YouTube, and Spotify; models for selecting and ranking results on search engines like Google, Bing, and Baidu; and feeds and contact recommendation systems on social media like Facebook and X (formerly Twitter).[51]

Machine learning systems use algorithms and artificial neural networks, which are inspired by networks of human neurons. These mathematical models mimic our nervous system,[52] allowing different data processors to work together to process information. Deep learning, a technique that enhances machine learning, establishes multiple layers of artificial neural networks. This simulates the complex functioning of the human brain, enabling these networks and processors to manage information and establish correlations simultaneously.[53] Deep learning is used in image and speech recognition, machine translation, and text processing. There are three types of machine learning: supervised, unsupervised, and reinforcement. [54]

AI systems can be classified in various, sometimes overlapping, ways based on their use or purpose. Foundation models are trained with large amounts of data and are designed to adapt to multiple tasks. ChatGPT (Chat Generative Pre-trained Transformer), a generative AI, is an example of a foundation and large language model capable of generating art, images, texts, and sounds.[55] Generative AI programs can create new content, not just analyze or classify existing content. For instance, if someone searches how an electric car works on Google, the search engine will provide a link to a third-party website's answer. In contrast, ChatGPT will explain how an electric car works in its own words.[56]



The landing page of the popular AI chatbot *ChatGPT*.

Foundation models are considered a step towards so-called general-purpose AI, which has not yet been fully achieved but is already capable of performing a wide range of tasks and is not limited to specific purposes.[57] In contrast, fixed-purpose AI systems

[50] Dierle José Coelho Nunes & Otávio Morato de Andrade, O Uso Da Inteligência Artificial Explicável Enquanto Ferramenta Para Compreender Decisões Automatizadas: Possível Caminho Para Aumentar a Legitimidade e Confiabilidade Dos Modelos Algorítmicos?, 18 REV. ELETRÔNICA DO CURSO DIREITO DA UFSM 1, 4 (2023), https://periodicos.ufsm.br/revistadireito/article/view/69329; Sara Brown, Machine Learning, Explained, MIT MANAGEMENT SLOAN SCHOOL (2021), https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (last visited Jan 18, 2024).

[51] Karen Hao, What Is Machine Learning?, MIT TECHNOLOGY REVIEW (2018), https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/ (last visited Jan 18, 2024); Nunes and Andrade, supra note 54 at 5.

[52] See FÁBIO RIBEIRO PORTO, WALTER SHUENQUENER DE ARAÚJO, & ANDERSON DE PAIVA GABRIEL, INTELIGÊNCIA ARTIFICIAL GENERATIVA NO DIREITO 37 (2024). (mimeographed).

[53] Hao, supra note 55; Re, et al., supra note 21 at 244–246.

[54] Supervised learning, which is the main type currently used, is one in which it is possible to define the appropriate output. This is the case of AI systems aimed at categorizations (in cats, dogs, means of transport, etc.). Unsupervised learning (unsupervised), less common, is one that seeks to identify patterns, correlations, and groupings without prior definition of the appropriate result. Finally, reinforcement learning aims to train and improve machines through a trial-and-error system, in order to teach them to make the best decisions. Brown, supra note 54.

[55] Not all generative AI constitutes a foundational model. It can be modeled for very specific purposes. Generative capabilities may include manipulation and analysis of text, image, video, and speech production. Generative apps include chatbots, photo and video filters, and virtual assistants. Elliot Jones, Explainer: What Is a Foundation Model?, ADA LOVELACE INSTITUTE (2023), https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/#table1.

[56] Hasala Ariyaratne, ChatGPT and Intermediary Liability: Why Section 230 Does Not and Should Not Protect Generative Algorithms, SSRN 4, 4–5 (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4422583#:~:text=Section%20230%20was%20designed%20to,content%20in%20part%20by%20itself..

[57] RISHI BOMMASANI, ET AL., ON THE OPPORTUNITIES AND TASKS OF FOUNDATION MODELS 4–12 (2021), https://crfm.stanford.edu/report.html; RISTO UUK, CARLOS IGNACIO GUTIERREZ, & ALEX TAMKIN, OPERATIONALISING THE DEFINITION OF GENERAL PURPOSE AI SYSTEMS: ASSESSING

(or "narrow AI") are designed for specific tasks, in a more restrictive manner, and therefore trained with a more targeted database. This category includes most of the AI systems currently in use, such as voice assistants like Siri, Alexa, and Google Assistant, which are designed to understand and fulfill voice commands. It also includes content recommendation systems on streaming platforms, spam filters, weather prediction systems, and facial recognition software, among others. These systems are all intended for very specific purposes.[58]

Finally, the term strong AI, also referred to as Artificial General Intelligence (AGI), describes systems capable of understanding, learning, and applying knowledge similarly to human beings. Such systems would be able to reason, solve problems, and make their own decisions. This type of AI is a theoretical concept that has not yet been achieved, but some researchers believe it could be developed in the coming years.[59]

> "The truth is that analog life is being left behind."

There are many ways of exploring the technicalities of the subject. For example, there is the AI value chain, which includes various phases: design, development, implementation, and maintenance of a system.[60] Each stage presents unique challenges and risks and involves different stakeholders such as developers,[61] deployers,[62] users,[63] and final recipients.[64] These participants hold distinct expertise and skills, contributing differently and bearing varying degrees of responsibility and potential for damage.[65] This diversity of roles increases the difficulty of regulating the matter.

It is time, however, to move on, exploring the implications of AI that transcend strictly technical issues.

PART I: ARTIFICIAL INTELLIGENCE AND ITS BENEFITS

AI is increasingly becoming part of our daily lives, often so seamlessly that we don't even associate certain utilities with it. The truth is that analog life is being left behind. Certainly, there will always be those who prefer objects manufactured in the old style, such as high-end brand watches that celebrate an artisanal past, despite being less accurate and functional than their digital counterparts. These products continue to attract consumers, proving that the human species is not entirely driven by reason and pragmatism. However, aside from extravagances and idiosyncrasies, the fact remains that today we use search algorithms to research any topic; recommendation algorithms to choose products, books, trips, and accommodations; and intelligent traffic management and temperature measurement to optimize our commutes or decide what to wear. In short, AI brings many positive improvements, making our lives better and easier.

Considering the large-scale impact of AI, its uses and potentialities are so vast that listing and organizing them is challenging. Some notable examples are described below.

I. IMPROVED DECISION-MAKING CAPACITY IN MANY AREAS

In many areas, AI will be better at making decisions than humans for different reasons. Firstly, it can store a much larger amount of information than the human brain. Secondly, it can process this information much faster. Thus, AI can make correlations within a massive volume of data, beyond the capabilities of an individual or even a team. These correlations can reveal associations between factors that are too complex or nuanced for humans to detect. However, as previously noted, the efficiency of AI depends on the quantity and quality of the data it is fed. Moreover, in its current state, generative AI tools can sometimes produce absurd information, a phenomenon

FOUR APPROACHES, STANFORD UNIVERSITY (2023).

[58] Abid All Awan, What is Narrow AI?, DATACAMP (June 2023), https://www.datacamp.com/blog/what-is-narrow-ai.

[59] TOM TAULLI, INTRODUÇÃO À INTELIGÊNCIA ARTIFICIAL: UMA ABORDAGEM NÃO TÉCNICA 218 (2020).

[60] Hacker, Engel, & Mauer, supra note 29 at 8–11.

[61] This would be the case for OpenAI regarding GPT-4 (foundational model).

[62] In the case of ChatGPT (specific purpose model), OpenAI is both the developer and implementer. Be My Eyes, in turn, is only an implementer of Virtual Volunteer (also for specific purposes), adapted from GPT-4.

[63] In the case of the products referred to in the previous note, the user is the one who generates the text with ChatGPT or consults the voice application, via Virtual Volunteer.

64 The ones who use the text and voice directions. Note that the news produced may or may not be true. The voice can be good guidance or a deep fake, produced to deceive the recipient.

[65] Hacker, Engel, & Mauer, supra note 29 at 8–11.

known as hallucinations.[66] It is also important to note that, in areas requiring emotional intelligence, ethical values, or understanding the subtleties of human behavior, human intervention is indispensable and human decision-making remains superior.

II. AUTOMATION

AI allows the automation of various tasks, both ordinary and complex, significantly increasing productivity and efficiency across many fields. Machines used on industrial production lines can perform repetitive, stressful, or strenuous tasks that would be challenging for humans. Additionally, AI reduces the margin for error and can eliminate risks in hazardous jobs such as mining, bomb disposal, underwater cable repairs, or space travel. Furthermore, AI can operate continuously 24/7, producing on a larger scale with greater precision, and at a lower cost. Unlike humans, AI does not get tired, fall ill, change moods, or pose the risk of labor complaints. The potential negative impact on the labor market will be examined further along in this paper.

III. LANGUAGE

The impact of AI on the field of language has been profound and versatile, especially due to the use of natural language processing. In the last years, the quality of translations made by Google Translate, ChatGPT, and DeepL, to name a few, has improved significantly, making them very accurate and proficient. This breaks down many language barriers in human communication. Tools like Siri, Alexa, and Google Assistant respond to voice commands, whereas others turn text into speech. Chatbots help solve customer questions and problems, and generative AI, which has been leaving the world amazed, communicates with the user through text, sounds, and images. The advances in this area are extraordinary.

IV. RESEARCH AND INNOVATION

AI has expanded research and innovation in almost all areas of human activity, from physics and chemistry to automotive and space industries. The volume of science produced based on AI has grown significantly. AI can simplify and expedite clinical research and trials of new medicines, materials, and products. By analyzing large amounts of data, AI accelerates the process of scientific discovery. This leads to significant cost and time reductions in the development of new drugs and autonomous cars, with the latter promising to reduce accidents. AI holds the potential to help humanity overcome numerous challenges, such as addressing climate change, combating hunger, controlling pandemics, promoting urban sustainability, and preventing diseases such as cancer and Alzheimer's.[67] The movement advocating for the beneficial use of AI's potential is known as "data for good."[68]

V. APPLICATIONS IN MEDICINE

Medicine is one of the areas where AI will significantly impact people's lives and health. Technologies such as machine learning and natural language processing are poised to improve the quality of patient care and reduce costs. Some potential benefits include enhanced diagnostics, image analysis, robotic surgeries, personalized treatment planning, telemedicine, disease prediction, and patient data management. While AI will not render doctors' work dispensable, it could shift some of the roles, emphasizing the human aspects of empathy and motivation over technical tasks. However, there will also be ethical and legal implications, such as accountability for errors made by AI equipment.[69]

VI. APPLICATIONS IN THE JUSTICE SYSTEM

AI offers the prospect of radical changes to the practice of law and jurisdictional services. In an environment where reliance on precedents is paramount, efficient research of case law is essential. AI can simplify and expedite this process by enabling lawyers to draft papers, prosecutors to prepare reports, and judges to issue opinions based on research and summaries generated by it. However, professionals must closely supervise the generated output, as ultimate responsibility for AI use still falls on them. In courts, AI programs can summarize extensive cases and group them by subject, optimizing judges' time and energy. Additionally, the digitization of proceedings, the automation of certain procedures, and online dispute resolutions can make justice more agile and efficient. In various Brazilian courts, over a hundred projects use AI in jurisdictional services.

[66] It should be noted that AI hallucinations could have brutal repercussions on relevant social processes. For example, according to a report by AI Forensics, Microsoft's Bing chatbot provided misleading information in 30 percent of basic queries on election-related topics in Germany and Switzerland. The problem also grew when questions were asked in languages other than English. Carlos Guadián, Cómo Va a Afectar La Inteligencia Artificial Las Elecciones En 2024, CLUDPAD (2024), https://carlosguadian.substack.com/p/como-va-a-afectar-la-inteligencia?utm_source=post-email-title&publication_id=259698&post_id=141029307&utm_campaign=email-post-title&isFreemail=true&r=4n36v&utm_medium=email.

[67] See European Commission, The Impact of AI on R&I, QUARTERLY R&I LITERATURE REVIEW 2023/Q2 (2023), https://research-and-innovation.ec.europa.eu/document/download/63eb38a0-6632-4ab0-bb27-b898f553858f_en?filename=ec_rtd_quarterly-ri-review_022023.pdf.

[68] JOSÉ MANUEL MUÑOZ VELA, RETOS, RIESGOS, RESPONSABILIDAD Y REGULACIÓN DE LA INTELIGENCIA ARTIFICIAL: UN ENFOQUE DE SEGURIDAD FÍSICA, LÓGICA, MORAL Y JURÍDICA 65 (2022).

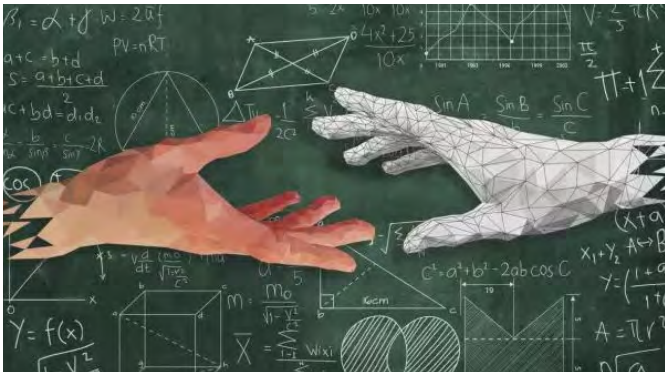
[69] For further information on this subject, see Thomas Davenport & Ravi Kalakota, The Potential for Artificial Intelligence in Healthcare, 6 FUTUR. HEALTHC. J. 94 (2019).

There is a controversial and particularly interesting point here: the use of AI to support the drafting of judicial decisions. Many fear, not without reason, the risks of prejudice, bias, and lack of transparency and explainability, as well as a lack of social sensitivity, empathy, and compassion. However, it is important to remember that human judges are subject to these same problems. This brings us to an opposing view: that AI could be more prepared, impartial, and less susceptible to personal interests, political influences, or intimidation. This potential is particularly relevant in less developed countries with lower judicial independence or higher corruption levels.[70] Regardless, at our current stage of civilization and technology, supervision by a human judge is essential. Additionally, judges may face an increased argumentative burden in cases where they intend to produce a result different from that proposed by AI.

VII. EDUCATION AND CULTURE

AI is going to transform the landscape of education worldwide, both in teaching methods and learning opportunities. Initially, the Internet, enhanced by AI, has exponentially expanded access to knowledge and information, broadening the horizons of everyone with access to the World Wide Web. Additionally, distance education has overcome the barriers of time and space, enabling people to learn at any time and from anywhere. Digital libraries eliminate the need for physical presence, allowing users to search repositories located anywhere in the world. From a teacher's perspective, AI can assist in preparing lessons, formulating questions, and even grading assignments. It can also handle administrative tasks, freeing up time for more academic activities. All of this, of course, must be done under human supervision.

From a student's perspective, AI, especially generative AI, greatly facilitates research by summarizing long texts, correcting grammatical errors, suggesting writing improvements, and overcoming language barriers. AI can also personalize teaching, tailoring it to the needs of students, and assisting people with disabilities by transforming text into speech or vice versa.[71] However, to ensure these benefits are evenly distributed across the population, high-quality connectivity for all (digital inclusion) is essential. We must also acknowledge the potential dysfunctions that can arise from AI use in education, such as plagiarism and limiting creativity and critical thinking. This is why international organizations such as the OECD[72] and UNESCO[73] have produced important documents with principles and guidelines for the use of AI in education.



Cover art of UNESCO's guidelines for the use of AI in education: *AI and education: guidance for policy-makers*

AI's impact on culture will also be immense. On a positive note, it will open doors for creativity, working synergistically with musicians, painters, writers, architects, graphic designers, and countless other creative professionals. Generative AI can assist in composing symphonies, literary works, poetry, and stories. However, this also raises numerous ethical questions about intellectual property and copyright.[74] Yuval Noah Harari's statement that AI has already hacked the operating system of human culture, which is language, warrants serious reflection. He further ponders what it would mean for humans to live in a world where a large portion of novels, music, images, and laws, among many other creations, are conceived by nonhuman intelligence.[75]

[70] For further information on this subject, see Gustavo Ariel, Are Artificial Intelligence Courts a Discrimination Risk?, EUROPEAN AI ALLIANCE (2021), https://futurium.ec.europa.eu/en/european-ai-alliance/open-discussion/are-artificial-intelligence-courts-discrimination-risk. See also Cass R. Sunstein, Governing By Algorithm? No Noise and (Potentially) Less Bias, 71 DUKE L. J. 1175 (2022).

[71] In the field of education, AI can also benefit high-ability (gifted) students, since the skills of perception, recognition, and recommendation enable supervision, understanding, and adaption to each student's learning process. It may also offer teachers more time for individual instruction. LEE, supra note 7 at 149.

[72] OECD, OPPORTUNITIES, GUIDELINES AND GUARDRAILS FOR EFFECTIVE AND EQUITABLE USE OF AI IN EDUCATION (2023), https://www.oecd-ilibrary.org/docserver/2b39e98b-en.pdf.

[73] UNESCO, AI and Education: Guidance for Policy-Makers, AI AND EDUCATION: GUIDANCE FOR POLICY-MAKERS 50 (2021), https://unesdoc.unesco.org/in/documentViewer.xhtml.

[74] GISELLE. BEIGUELMAN, POLÍTICAS DA IMAGEM: VIGILÂNCIA E RESISTÊNCIA NA DADOSFERA 59 (2021).

[75] Yuval Noah Harari, You Can Have the Blue Pill or the Red Pill, and We Are out of Blue Pills, N.Y. TIMES, (Mar. 24, 2023), https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html.

VIII. OTHER USEFUL APPLICATIONS OF AI

1. PRACTICAL EVERYDAY USES

AI technology is embedded in many applications on personal computers and smartphones, such as Google Maps, Waze, Uber, Spotify, Zoom, Facebook, and Instagram, as well as in personal assistants like Siri and Alexa. It also plays a crucial role in the streaming entertainment industry (Netflix, Amazon Prime, HBO Max) and gaming. Additionally, AI is integral to apps that facilitate bank transactions and credit card payments, among countless other uses.



Amazon's virtual assistant technology, Alexa.

2. ENVIRONMENTAL PROTECTION

AI will play an increasingly critical role in environmental protection, data analysis, weather prediction, and monitoring of emergency situations. Examples are manifold and include: analyzing climate change data, satellite and drone imaging, monitoring air, water, and soil pollution levels, rationalizing the distribution and consumption of energy and water, predicting natural disasters (such as hurricanes, earthquakes, and floods), aiding sustainable agriculture through soil sensors and other instruments, reducing the use of pesticides, guiding irrigation, and helping to plan reforestation.[76]

3. PERSONALIZATION OF COMMERCIAL RELATIONSHIPS, AMONG OTHERS

AI enables industry, commerce, services, media, and digital platforms to target consumers with information, news, and advertisements tailored to their interests. This optimizes people's time and facilitates decisions such as purchasing products or books, travel planning, and numerous other activities. Recommendations for movies, music, or other forms of entertainment may also stem from this use of AI. However, we must consider the negative aspects, such as the tribalization of life due to confirmation bias resulting from materials that reinforce existing preferences and convictions.

This phenomenon reduces the plurality of views, generates new forms of social control,[77] and can lead to polarization and radicalism.[78] In terms of personal relationships, studies have indicated that marriages resulting from online matches tend to be slightly more satisfactory than those in which partners meet through conventional offline methods.[79]

It is unfruitful to endlessly list all the uses and benefits of AI. The list continues to grow daily. These applications include the development of autonomous vehicles, monitoring equipment to detect possible infrastructure failures, fraud detection (especially financial), cybersecurity enhancement, aviation controls, and more. Instead, we must now turn our attention to the problems, risks, and threats that can arise from the large-scale use of AI.

Part II: ARTIFICIAL INTELLIGENCE AND ITS RISKS

Every new technology has a disruptive effect on production, consumption, the labor market, and social life. Moreover, like many things in life, innovations can have a negative side or be practiced by bad social actors. The spinning machine caused unemployment among seamstresses and craftsmen and off-

[76] Nevertheless, the energy costs of powering, operating, and maintaining AI, as well as its systemic impacts on different ecosystems, should not be ignored either, as some researchers have already observed. See Eva García-Martín, et al., Estimation of Energy Consumption in Machine Learning, 134 J. PARALLEL DISTRIB. COMPUT. 75, 75–88 (2019), https://doi.org/10.1016/j.jpdc.2019.07.007; Galaz, et al., supra note 22 at 1–10.

[77] On the use of AI as a form of social control: "Search engines presented another challenge: ten years ago, when search engines were powered by data mining (rather than by machine learning), if a person searched for 'gourmet restaurants', then for 'clothing', his or her search for the latter would be independent of his or her search for the former. Both times, a search engine would aggregate as much information as possible, then provide the inquirer options . . . . But contemporary search engines are guided by models informed by observed human behavior. . . . Designer clothing may be what the searcher is after. But there is a difference between choosing from a range of options and taking an action—in this case, making a purchase; in other cases, adopting a political or philosophical position or ideology—without ever knowing what the initial range of possibilities or implications was, entrusting a machine to preemptively shape the options." HENRY KISSINGER, ERIC SCHMIDT, & DANIEL HUTTENLOCHER, A ERA DA IA E NOSSO FUTURO COMO HUMANOS 20 (2023). [Originally published in English as The Age of AI And Our Human Future].

[78] See, e.g., Luís Roberto Barroso & Luna van Brussel Barroso, Democracia, Mídias Sociais, e Liberdade de Expressão: Ódio, Mentiras e a Busca Da Verdade Possível, 17 REV. BRAS. DIREITOS FUNDAM. JUSTIÇA 285 (2023), https://www.bbc.com/news/uk-politics-eu-referendum-36619444.

[79] Dolores Tropiano, More Internet Marriages Are Leading to Happy Marriages, ARIZONA STATE UNIVERSITY NEWS (2023), https://news.asu.edu/20230119-university-news-more-internet-matches-are-leading-happy-marriages (last visited Mar 31, 2024); William Harms, Meeting Online Leads to Happier, More Enduring Marriages, UCHICAGO NEWS (2013), ews.uchicago.edu/story/meeting-online-leads-happier-more-enduring-marriages.

set printing eliminated the jobs of linotypists. Computerization has reduced the need for bank employees in the financial system. Digital platforms have paved the way for extremist polarization,[80] disinformation,[81] and hate speech.[82] Even worse, the invention of caravels enabled transoceanic trade, but also facilitated the slave trade. [83]

> "Every new technology has a disruptive effect on production, consumption, the labor market, and social life."

For these reasons, it is crucial to observe the adverse effects of using AI to neutralize or mitigate them. These detrimental effects can have social, economic, and political implications, and could even undermine world peace. The following is a survey of some of the consequences, risks, and threats posed by AI.

I. IMPACT ON THE LABOR MARKET

This is the most obvious and predictable effect, a consequence of what generally occurs when new technology transforms the existing means of production. With the advance of automation, the job market landscape will change fundamentally, requiring workers from various different sectors to adapt to new roles. This transition is not always easy. In the case of AI, the impact will affect not only more mechanical jobs, but also skilled and creative functions,[84–85] It is certain that new technologies tend to generate new markets and, consequently, new jobs. However, there is a problem of timing and scale in this consideration. New jobs are unlikely to be created at the same rate and volume as those being displaced. [86] This poses a significant challenge, requiring governments to invest in social protection and worker training. Additionally, increased economic vulnerability can affect democratic stability, as it has historically served as a potential destabilizing factor.

II. USE FOR MILITARY PURPOSES

The literature on the use of AI for military purposes is relatively scarce, primarily due to the secrecy imposed on such matters for security reasons. Nevertheless, throughout history, new technologies often originate from or are quickly adapted for military use. It is not difficult to imagine countries like the United States and China competing to leverage AI for military purposes, using new technologies and robots. In fact, automated drones have already been employed in missions involving reconnaissance, surveillance, equipment delivery, and even air strikes. A particularly concerning issue is autonomous lethal weapons, which can engage in combat and attack targets autonomously, without human control. There are ongoing debates about the strict control of their use through international agreements.[87]

---

[80] MAX FISHER, A MÁQUINA DO CAOS: COMO AS REDES SOCIAIS REPROGRAMARAM NOSSA MENTE E NOSSO MUNDO 20 (2023). [Originally published in English as The Chaos Machine: The Inside Story of How Social Media Rewired Our Minds and Our World].

[81] MICHIKO KAKUTANI, A MORTE DA VERDADE: NOTAS SOBRE A MENTIRA NA ERA TRUMP 17 (2018). [Originally published in English as The Death of Truth: Notes on Falsehood in the Age of Trump].

[82] PATRÍCIA PERRONE CAMPOS MELLO, A MÁQUINA DO ÓDIO: NOTAS DE UMA REPÓRTER SOBRE FAKE NEWS E VIOLÊNCIA DIGITAL (2020); MATTHEW WILLIAMS, A CIÊNCIA DO ÓDIO 207 (2021). [Originally published in English as The Science of Hate].

[83] DARON ACEMOGLU & JOHNSON SIMON, POWER AND PROGRESS: OUR THOUSAND-YEAR STRUGGLE OVER TECHNOLOGY AND PROSPERITY 4–5 (2023). The authors list some inventions from the last thousand years that did not necessarily bring prosperity to all people.

84 Steve Lohr, Generative A.I.'s Biggest Impact Will Be in Banking and Tech, Report Says, N.Y. TIMES, (Feb. 1, 2024), https://www.nytimes.com/2024/02/01/business/ai-impact-jobs.html. ("The research estimates that banks and some tech companies spend 60 to 80 percent of their payrolls, or more, on workers in occupations most likely to be affected by the new technology."). Similarly, regarding the advertisement and marketing industry: Sapha Maheshwari, A.I. Fuels a New Era of Product Placement, N.Y. TIMES, (Feb. 1, 2024), https://www.nytimes.com/2024/02/01/business/media/artificial-intelligence-product-placement.html.

[85] Other studies indicate that jobs requiring social intelligence (public relations), creativity (biologists and designers), and perception and fine manipulation (surgeons) tend to be more spared. Carl B. Frey & Michael A. Osbourne, The Future of Employment: How Susceptible Are Jobs to Computerisation?, FUTURE OF HUMANITY INSTITUTE (2017), https://www-sciencedirect-com.ezp-prod1.hul.harvard.edu/science/article/pii/S0040162516302244. They also believe that positions in analysis, forecasting, and strategy will be the most affected. Michael Webb, The Impact of Artificial Intelligence on the Labor Market, 158713 SSRN ELECTRON. J. (2019).

[86] As Keynes observed almost a century ago, technological progress tends to generate at least a temporary disagreement in terms of work until new job opportunities are identified. John Maynard Keynes, Economic Possibilities for Our Grandchildren, ESSAYS PERSUAS. 358 (1930), http://www.mobihealthnews.com/content/ey-63-percent-patients-comfortable-tech-enabled-health-tracking.

[87] Michael T. Klare, UN to Address Autonomous Weapons Systems, ARMS CONTROL ASSOCIATION (2023), https://www.armscontrol.org/act/2023-12/news/un-address-autonomous-weapons-systems (last visited Mar 28, 2024).

The ethical implications of this type of weaponry are profound, making strict regulation—or even outright banning—crucial.



An American-made MQ-1C Gray Eagle autonomous military drone.

Additionally, communication and information technologies have been integral to military efforts for some time, leading to recurring tactics in the context of hybrid warfare. Alongside physical destruction, these new forms of aggression involve influence and disinformation campaigns (cognitive warfare) and cyberattacks aimed at compromising essential computer systems, such as power supply infrastructures. [88]

III. MASS DISINFORMATION

Since at least 2016, the spread of information through digital platforms and messaging apps has posed a serious threat to the democratic and electoral process. Studies have shown that lies and radicalism circulate online faster and with greater engagement than truthful and moderate discourses. Emotional, unlikely, and alarming content generates more engagement and mobilization. Deepfakes exacerbate the issue by simulating people saying things they never said, tampering with content and reality in ways that are undetectable to citizens.[89] This is not a hypothetical scenario; it is a worrying reality. The influence that mass disinformation has had on historical events, such as the United Kingdom's exit from the European Union (Brexit), the US elections in 2016, and the Brazilian elections in 2018, has become prominent. Democracy relies on the informed participation of citizens, which is seriously compromised by the widespread circulation of deliberate lies, character assassination, and conspiracy theories.

IV. VIOLATION OF PRIVACY

The business model of platforms using AI revolves around collecting as much personal data as possible from individuals, effectively turning privacy into a commodity.[90] Complex algorithms and multiple neural layers analyze this data to establish deep correlations, revealing insights into individuals' genetic data, psychological systems, vulnerabilities, and consumer, political, financial, sexual, and religious behaviors.[91] With this information, AI can make predictions, provide recommendations, manipulate interests, and produce the results desired by the algorithm. Consequently, access to private data, whether from individuals or companies, is crucial to the AI business model as it currently stands.[92] It is no accident that academic circles refer to data as the "oil" of this century. [93]

Regarding privacy, at least three aspects require attention.

1. The collection of Internet users' data without their consent by digital platforms and websites. This information is used for commercial purposes, such as targeting information and advertising, or even for manipulating the user's will, as neuroscience research has shown.

2. An aspect concerns surveillance and tracking by government and police authorities using facial recognition technologies and tracking tools. Although the objective is to fight crime, the risks of abuse are very high, especially in authoritarian governments.

[88] FREDERICO FRANCO ALVIM, RODRIGO LÓPEZ ZILIO & VOLGANE OLIVEIRA CARVALHO, GUERRAS COGNITIVAS NA ARENA ELEITORAL: O CONTROLE JUDICIAL DA DESINFORMAÇÃO 69 (2023).

[89] Luís Roberto Barroso & Luna van Brussel Barroso, Democracy, Social Media, and Freedom of Expression: Hate, Lies, and the Search for the Possible Truth, 24 CHIC. J. INT. LAW 51 (2023); Patrícia Perrone Campos Mello & Renata H. S. B. Rudolf, Redes Sociais e Democracia: Disrupção Tecnológica, Erosão Democrática e Novas Perspectivas, in DIREITO E POLÍTICA: UM DIÁLOGO POSSÍVEL? 53 (Eduarda Peixoto da Cunha França & Matheus Casimiro eds., 2023).

[90] EVGENY MOROZOV, BIG TECH: A ASCENSÃO DOS DADOS E A MORTE DA POLÍTICA 36 (2018).

[91] Aziz Z. Huq, Constitutional Rights in the Machine Learning State, 105 CORNELL L. REV. 1875, 37 (2020).

92 Shoshana Zuboff, Surveillance Capitalism or Democracy?: The Death Match of Institutional Orders and the Politics of Knowledge in Our Information Civilization, 3 ORGAN. THEORY 1 (2022). Many other rights restrictions derive from the restriction of privacy, such as damages related to physical, reputation, relational, psychological (emotional), economic, and discriminatory damages, as well as those related to the exercise of human autonomy (coercion, manipulation, misinformation, deformation of expectations, loss of control, among others). Danielle Keats Citron & Daniel Solove, Privacy Harms, 102 BOST. UNIV. L. REV. 793, 83 (2022), https://www.google.ca/search?q=privacy+2018; Huq, supra note 95 at 34–41.

[93] REBOLLO DELGADO, supra note 44 at 17.

3. AI systems require vast amounts of data to train their models, which introduce risks of data leakage and cyberattacks by malicious actors. For example, spear phishing[94] and doxing[95] can fuel harassment, political violence, malinformation, and disinformation.

V. ALGORITHMIC BIAS

Algorithms are trained with existing data, which inherently reflect past and present human behaviors imbued with biases and prejudices shaped by historical, cultural, and social circumstances.[96] Consequently, they tend to perpetuate current and past social structures of inclusion and exclusion. For instance, employability data often show lower hiring rates for women, Black, and Indigenous people, a trend that is unrelated to their abilities and productivity but can lead to the perpetuation of such discriminatory behaviors in the future. Public security data may indicate higher recidivism and violence rates among Black people, not because they are more violent, but because they often live in harsher environments.[97] Health cost data may overestimate expenses for some groups and underestimate for others, not necessarily due to their physical conditions.[98] Credit risk data may increase risks and financing costs for those with lower economic and social status, even if they have improved their conditions, depending on how the data was collected.[99] Consequently, some algorithms may disproportionately exclude women, criminalize Black men, and prevent the poorest from accessing credit. Under these conditions, AI can deeply reinforce existing inequalities, to the detriment of the most vulnerable groups in society.[100]



The New York Times headquarters in New York, NY.

VI. INTELLECTUAL PROPERTY AND COPYRIGHT ISSUES

The AI business model raises important questions about copyright and intellectual property. Who owns the copyright to the vast universe of songs, films, news reports, and content collected by Big Tech companies to feed their AIs? Is it the original authors and creators, or those who employ and exploit this content through algorithms? Generative AI is trained with an enormous amount of data, yet the answers it provides do not identify the source and author of the information. Debate on this subject has intensified and reached the courts. For example, AI companies collect content made by journalistic co panies and use it to train applications that then compete with these very outlets in producing informa-

[94] MUÑOZ VELA, supra note 72 at 64. These are malicious emails or messages, customized for a specific recipient, appearing credible, aimed at obtaining sensitive information (passwords, for example) or installing malware, which are malicious programs with damaging consequences on the affected systems.

[95] MICHELE PRADO, TEMPESTADE IDEOLÓGICA. BOLSONARISMO: A ALTRIGHT E O POPULISMO ILIBERAL NO BRASIL 162 (2023). Doxing means the malicious subtraction of information about someone, either from public files or by hacking computers, with the purpose of harassing, intimidating or extorting, among others.

[96] Regarding biases and heuristics in cognitive processes and human behavior, see the previous note. See also, Ricardo Lins Horta, Por Que Existem Vieses Cognitivos Na Tomada de Decisão Judicial?: A Contribuição Da Psicologia e Das Neurociências Para o Debate Jurídico, 9 REV. BRAS. POLÍTICAS PÚBLICAS 85 (2019).

[97] Jeff Larson, et al., How We Analyzed the COMPAS Recidivism Algorithm, PROPUBLICA (2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. A similar situation occurs with facial recognition technologies whose application has indicated racial bias, as well as the occurrence of repeated "false positives," leading to wrongful arrests and highly vexatious situations. A recent episode, related to the arbitrary detention of a fan in a football stadium, led the Brazilian state of Sergipe to suspend the use of the tool. See Uesley Durães, Reconhecimento Facial: Erros Expõem Falta de Transparência e Viés Racista, UOL (2024), https://noticias.uol.com.br/cotidiano/ultimas-noticias/2024/04/28/reconhecimento-facial-erros-falta-de-transparencia.htm.

[98] Ziad Obermeyer, et al., Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations, 366 SCIENCE 447 (2019), https://www.science.org/doi/10.1126/science.aax2342. In this case, the data used to train the algorithm was incomplete. Cost data on Black and White patients was used to understand the extent of their health needs. The resources spent on Black patients were lower than those spent on White patients, not because their needs were lower, but because they had greater difficulty accessing the service. As a result, AI wrongly underestimated the needs of Black patients.

[99] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 4 (2016).

[100] Huq, supra note 95 at 29–34; Silberg & Manyika, supra note 23 at 3.

tion.[101] This issue is at the heart of a lawsuit filed by The New York Times against OpenAI and Microsoft.[102] A similar claim involves Getty Images, a visual media company and image provider, and Stability AI, an AI company.[103]

There is no way to be exhaustive when exploring the risks involved in the development of AI, as countless possibilities must be considered, including those we have yet to imagine or anticipate. However, one last concern deserves adequate consideration: the concept of singularity. This term refers to the risk of computers gaining consciousness, acquiring a will of their own, and surpassing human capabilities. Since computers can process a much greater volume of data at much higher speeds, they could become superior to humans if they develop consciousness and willpower. The fear stems from the fact that AI systems can self-improve, potentially reaching a level of superintelligence that masters scientific knowledge, general culture, and social skills far exceeding those of the best human brains.

Skeptics of human potential might even suppose that a superintelligence beyond human capabilities could better address some of humanity's biggest unresolved issues, such as poverty, inequality, and environmental degradation. However, one could never be certain whether this uncontrolled intelligence would serve humanity's interests and values. This is why the governance of AI, both domestically and internationally, needs to establish safety protocols and ethical parameters designed to manage and mitigate this risk. If technology ever reaches this point—which many scientists doubt—the very future of civilization and humanity will be at stake.

Yuval Noah Harari makes a curious comment on the subject. In 2022, around 700 of the most important AI scientists and researchers were asked about the dangers of this technology, specifically whether it could affect human existence or cause severe disempowerment. Half of them said that the risk would be of 10% or more. Given this, Harari poses a fateful question: would you still board an airplane if the engineers who built it said there was a 10% chance of it crashing?[104] If that is the case, you would not be able to sleep peacefully.

Part III: SOME PRINCIPLES FOR REGULATING ARTIFICIAL INTELLIGENCE

I. COMPLEXITIES OF THE REGULATION

Based on everything discussed so far, it seems that the regulation of AI is indispensable. However, the task is not simple and faces numerous challenges and complexities. Below, we try to identify some of them.



Billionaire mogul Elon Musk was one of a range of experts who called for a pause in the development of powerful artificial intelligence.

Regulation needs to be done on a moving train. In March 2023, over a thousand scientists, researchers, and entrepreneurs signed an open letter calling for a pause in the development of the most advanced AI systems, given the "profound risks to society and humanity" they represented. The proposed pause, for at least six months, aimed to introduce "a set of shared security protocols.[105] While the concerns were fully justified, research was not suspended. The train con-

[101] For information on the crisis of the press business model and the effects of it on democracy, see MARTHA MINOW, SAVING THE NEWS: WHY THE CONSTITUTION CALLS FOR GOVERNMENT ACTION TO PRESERVE FREEDOM OF SPEECH 35 (2021); Vicki C Jackson, Knowledge Institutions in Constitutional Democracies: Preliminary Reflections, 14 J. MEDIA LAW 275, 280 (2022), https://heinonline.org/HOL/Page?handle=hein.journals/cajcl7&div=10&id=&page=. In Brazilian scholarship, see Roberto Barroso and van Brussel Barroso, supra note 82; Mello and Rudolf, supra note 93.

[102] Michael M. Grynbaum & Ryan Mac, The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work, N.Y. TIMES (Dec. 27, 2023), https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html; United States & Federal District Court of Manhattan, The New York Times v. Microsoft Corporation, OpenAI Inc., OpenAI LP, OpenAI GP LLC, OpenAI LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, LLC 1 (2023), https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

[103] Getty Images claims that Stability used the images produced by them to train an image-generating AI system called Stable Diffusion, without authorization, violating its intellectual property rights and the copyrights of its collaborators, for offering services similar to its own. James Vincent, Getty Images Is Suing the Creators of AI Art Tool Stable Diffusion for Scraping Its Content, THE VERGE (2023), https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit; Thomas M Barwick Ink v. Stability AI Ltd Getty Images (US) Inc, Getty Images International UC, Letty Images UK Limited, Getty Images Devco UK Limited, Istockphoto LP, Case No. IL-2023-000007 (2023), https://www.documentcloud.org/documents/24183636-getty-images-v-stability-ai-uk-ruling.

[104] Harari, supra note 79.

[105] Future of Life Institute, Pause AI Giant Experiments: An Open Letter, FUTURE OF LIFE INSTITUTE (2023), https://futureoflife.org/open-letter/

tinued at high speed, driven by the competitive race among nations, researchers, and entrepreneurs. Nevertheless, the letter reinforced the urgent need for governance, regulation, monitoring, and attention to the social, economic, and political impacts of new technologies.

The speed of change is astonishing. This makes it extremely difficult to predict future developments and adapt legal norms accordingly, which risk becoming obsolete quickly. For instance, the traditional landline telephone took 75 years to reach 100 million users, the mobile phone took 16 years, the Internet took 7 years, while ChatGPT reached 100 million users in just two months.[106] It is not easy for legislation and regulation to keep pace with such rapid innovation.

Risks of excessive regulation. While regulation has become essential, as highlighted above, it also involves risks. Two of them are worth mentioning.

1. Restrictions and civil liability must not be so severe as to stifle the momentum of innovation.

2. Disproportionate regulation can create a market reserve for established companies, widening the gap between them and the competition, and exacerbating economic concentration among the big players.

The prevailing view is that regulation should focus on results rather than the research itself.

Asymmetry of information and power between companies and regulators. AI technology is primarily controlled by the companies involved in its development, which possess far greater knowledge than potential regulators. This imbalance is further compounded by the economic power of big tech companies, which are among the most valuable companies in the world. This economic power can easily be converted into political influence. A notable example of this occurred when the National Congress in Brazil voted on a bill that regulated disinformation on social media. Some technology companies launched an intense campaign against the measure on their platforms and lobbied the National Congress, ultimately leading to the bill being removed from the agenda.[107]



The National Congress Palace of Brazil, Brasília.

Need for global harmonization of regulation. AI is a predominantly private technology, which does not observe national borders. Companies operate globally, often without having their headquarters in the main centers of their business. Data can be collected and used to train systems in different parts of the world. Under these conditions, the way AI works challenges some essential elements of law as we practice it. These elements include the opposability of fundamental and human rights to States (rather than private entities) and the scope of national jurisdictions, which are limited by the sovereignty of other countries. Additionally, the heterogeneous regulatory treatment of the issue in different countries can lead to an exodus of investment and hinder technological development in more restrictive States, while encouraging widespread violations of rights in more permissive regions.

II. SOME REGULATION EFFORTS

At the international level, some initiatives involving non-binding propositions (soft law) have been notable. Among them, the following stand out: a) the 2019 OECD Council Recom-

pause-giant-ai-experiments/.

[106] The Feed, ChatGPT Witnesses Massive Rise, Chatbot Gains 100 Million Users in Two Months, ECONOMIC TIMES, 2023, https://economictimes.indiatimes.com/news/new-updates/chatgpt-witnesses-massive-rise-chatbot-gains-100-million-users-in-two-months/articleshow/98428443.cms?from=mdr.

[107] Constança Rezende, PF Conclui Que Google e Telegram Agiram de Modo Abusivo Contra PL Das Fake News. Empresas, Que Lançaram Ofensiva Contra Projeto No Ano Passado, Negam Irregularidades, FOLHA DE SÃO PAULO, 2024, https://www1.folha.uol.com.br/poder/2024/01/pf-conclui-que-google-e-telegram-agiram-de-modo-abusivo-contra-pl-das-fake-news.shtml; Brazil, PL Das Fake News Dá "Poderes de Censura" Ao Governo, Diz Telegram, PODER 360 (2023), ttps://www.poder360.com.br/tecnologia/pl-das-fake-news-da-poderes-de-censura-ao-governo-diz-telegram/; Brazil, Google Inclui Texto Contra PL Da Fake News Na Página Inicial Do Buscador. Big Tech Exibe Link Para Artigo Nomeado "PL Das Fake News Pode Aumentar a Confusão Entre o Que é Verdade Ou Mentira No Brasil," PODER 360 (2023), https://www.poder360.com.br/tecnologia/google-inclui-texto-contra-pl-das-fake-news-na-home-do-buscador/. There seems to be little interest in containing fake news or moderating content. The more fake news, the greater the user engagement and the greater the interaction on the networks, therefore the greater the production of data, the raw material for Big Tech.

mendation on Artificial Intelligence;[108] and b) the 2021 UNESCO Recommendation on Ethics in Artificial Intelligence.[109–110] Both documents aim to address the previously mentioned risks. They are convergent and complementary, and present fairly general principles on AI to be detailed by the domestic regulations of the respective countries.



President Joe Biden signing an Executive Order on AI regulations.

At the national level, at the end of 2023, the US issued a comprehensive executive order (EO) on AI.[111] This broad regulation addresses multiple areas of technology risk. The President of the United States directed federal agencies to establish standards and measures to test, ensure the security and reliability of technology, and prevent fraud and algorithmic discrimination. The EO aims to protect the fundamental rights of citizens, consumers, competitors, and students. It also mandates the identification of AI-produced content with watermarks and defines good practices, including conducting studies on the impacts of AI on labor relations and measures to mitigate such impacts.

Additionally, the EO provides for research funding and support for small businesses to access technical assistance, resources, and the AI market. It also seeks to attract new talent through immigration measures. Developers of foundational models that may present risks to national security, national economy, and public health are required to notify the government when training their systems and share the results of their safety tests. The EO also calls on Congress to pass a law protecting privacy rights and safeguarding citizens' data.

In March 2024, the European Union (EU) approved the Artificial Intelligence Act (EU AI Act). Unlike the US's EO, the EU AI Act establishes rules and sanctions regarding the development, implementation, and operation of AI. It also mandates concentrated action by specific bodies for monitoring and enforcement. These standards are, however, proportional to the risk posed by the technology to people and goods (risk-based approach).[112] In this framework, AI systems are classified into three levels: a) AI systems that are subject to unacceptable risks, whose implementation is prohibited;[113] b) high-risk systems, whose implementation is permitted, as long as they meet mandatory standards; [114] and c) AI systems that do not offer high-risks, for which incentives are provided for the voluntary adoption of codes of conduct, a type of self-regulation.[115]

In Brazil, PL (Draft Bill) 21/2020 and PL 2,338/2023 are being processed in the National Congress, with a tendency to align with the standards set out in the European Union's proposed norms.[116] In general, the proposals aim to: a) guarantee rights to

[108] OECD, supra note 42. This recommendation was also adopted by the G20. G20, G20 Ministerial Statement on Trade and Digital Economy, 1 (2019), https://www.mofa.go.jp/mofaj/files/000486596.pdf.

[109] UNESCO, The Ethics of Artificial Intelligence, UNESCO 1 (2023), https://unesdoc.unesco.org/ark:/48223/pf0000381137. The document lists the following 10 principles: 1 – Proportionality and non-harm; 2 – Security; 3 – Justice and non-discrimination; 4 – Sustainability; 5 – Privacy and data protection; 6 – Human supervision and determination; 7 – Transparency and explainability; 8 – Awareness and literacy; 9 – Responsibility and accountability; 10 – Plurality of participants, adaptive governance, and collaboration.

[110] The United Nations has also adopted Principles for the Ethical Use of AI in the United Nations System, very similar to those recommended by UNESCO, see United Nations, Chief Executive Board for Coordination. Principles for the Ethical Use of Artificial Intelligence in the United Nations System [Advanced Unedited Version], INTER-AGENCY WORKING GROUP ON ARTIFICIAL INTELLIGENCE (2019), https://unsceb.org/sites/default/files/2023-03/CEB_2022_2_Add.1 %28AI ethics principles%29.pdf.

[111] An Executive Order is a piece of legislation, a directive, issued by the President of the United States of America, aimed at managing the federal government. It is similar to a decree in Brazil. It is limited in its capacity to regulate, as it is not a law produced by the legislature, and can be changed by decision of the next president.

[112] See European Union, Artificial Intelligence Act (2024), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

[113] This category includes social scoring technologies, biometric identification in public places for law enforcement purposes (with specific exceptions), as well as subliminal practices for manipulating people and/or exploiting the vulnerabilities of vulnerable groups. Id. at 5.

[114] For instance, when the AI system is designed to be used as a security component of a product or involves biometric data, emotion recognition systems, and critical infrastructures such as road traffic or the supply of water, gas, heating, or electricity, among others. Id. at 6, Annex III.

[115] Id. at 95. The rule also mentions a specific transparency risk category, applicable to certain AI systems which must clearly disclose to users that they are interacting with a machine, such as chatbots, or biometric categorization or emotion recognition systems, when allowed. Id. at 50.

[116] Chamber of Deputies, PL n. 21/2020, Initiative of Deputy Eduardo Bismark, Status: Approved with Amendments in Plenary and Sent to the Senate

people directly affected by AI systems; b) establish responsibilities based on the levels of risks imposed by AI-driven systems and algorithms; and c) determine governance measures applicable to companies and organizations in this field.

III. SOME GUIDELINES

Considering all that has been exposed so far, we can identify several values, principles, and objectives that should guide the regulation of AI to ensure these technologies serve the cause of humanity, maximizing their benefits while minimizing their risks. The regulation should focus on defending fundamental rights, protecting democracy, and promoting good governance. Some elements and aspects linked to each of these aims are presented below.

1. DEFENDING FUNDAMENTAL RIGHTS

a) Privacy. The use of AI must respect the individual data of natural and legal persons and cannot utilize such data without explicit consent. Invasive surveillance, such as facial recognition, biometrics, and location monitoring, must be strictly restricted and controlled. Moreover, given the vast amount of data used to power AI, there must be robust security mechanisms to prevent data leaks.

b) Equality (Non-Discrimination). One of the most valuable pillars of contemporary civilization is the recognition of all people as equal in both material and formal aspects. We have previously highlighted the dangers of algorithmic discrimination. AI regulation must prevent the unequal treatment of individuals based on suspect categories that exacerbate vulnerabilities, such as gender, race, sexual orientation, religion, age, and other characteristics. The track record in this area has shown significant issues that need to be addressed. [117]

c) Freedoms. With regard to individual autonomy, the use of neuroscience and microtargeting in AI has the potential to manipulate people's behavior and will by exploring feelings of fear, prejudice, euphoria, and other cognitive biases. This can induce them to buy goods, hire services, or adopt behaviors contrary to their interests, violating their cognitive freedom or mental self-determination. Moreover, the right to information, pluralism of ideas, and freedom of expression can be compromised by recommendation or moderation algorithms, which filter, direct, and exclude content, effectively acting as a private censorship. Regulation of AI must safeguard individual autonomy and one's access to marketplaces of ideas.

2. PROTECTING DEMOCRACY

a) Countering Disinformation. Democracy is a regime of collective self-government, which presupposes the enlightened and well-informed participation of citizens. The circulation of disinformation and conspiracy theories misleads or generates unfounded fears in people, compromising their judgment and choices. This issue is exacerbated by deep fakes, which create realistic but false videos and speeches. We are all educated to believe what we see and hear. Manipulations of this nature break the paradigms of experience[118] and undermine democracy. Accounting for manipulated data is critical to preserving democratic underpinnings.

b) Countering Hate Speech. Since the enshrinement of universal suffrage, democracy has involved the equal participation of all people. Hate speech consists of attacks on vulnerable groups with racist, discriminatory, or ableist statements against Black, gay, and Indigenous people, and those with disabilities, among others. By intending to disqualify, weaken, or silence certain social groups, hate speech undermines the protection of human dignity and weakens democracy.

c) Countering attacks on democratic institutions. Social media, aided by AI, has been instrumental in organizing attacks on democratic institutions with the aim of destabilizing them. Insurrectionary acts, such as those on January 6, 2021, in the United States, and January 8, 2021, in Brazil, involved coup attempts to undermine election results. These attacks put democracy at risk and cannot be tolerated.[119]

3. PROMOTING GOOD GOVERNANCE

Considering the international, regional, and domestic recommendations and enactments already mentioned, and the broad public debate underway in academia, civil society, and the press,

(2020), https://www.camara.leg.br/propostas-legislativas/2236340; Brazil; Federal Senate, PL n. 2338/2023, Initiative of Senator Rodrigo Pacheco, Current Rapporteur Senator Eduardo Gomes, Status: With the Rapporteur (2023), https://www25.senado.leg.br/web/atividade/materias/-/materia/157233.

[117] Jeffrey Dastin, Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women, REUTERS (2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/; Larson, et al., supra note 101; Obermeyer et al., supra note 97; Will Douglas Heaven, Predictive Policing Is Still Racist—Whatever Data It Uses, MIT TECH. REV. (2021), https://www.technologyreview.com/2021/02/05/1017560/predictive-policing-racist-algorithmic-bias-data-crime-predpol.

[118] As a result of our experience, we are used to think of videos and photographic records as elements that attest to the veracity of information. However, AI based on deep learning technologies allows for the production of fraudulent but ultra-realistic media, which confuse minds and allow for the synthetic production of false representations of reality. Michael Filimovicz, Introduction, in DEEP FAKES: ALGORITHMS AND SOCIETY 78, X–XI (2022).

[119] On the topic, see IGNACIO RAMONET, LA ERA DEL CONSPIRACIONISMO: TRUMP, EL CULTO A LA MENTIRA Y EL ASALTO AL CAPITOLIO (2022).

it is possible to arrive at some overlapping consensus regarding AI governance, outlined in the five guidelines set out below:

1. Centrality of the Common Good. AI must be developed and oriented towards the well-being of people, countries, and the planet. Its benefits must be distributed fairly among all, and its negative impacts must be mitigated through legislation and regulation.120

2. Plural Governance. The governance of AI must include the participation of a wide range of actors at different stages, with appropriate proportionality. This includes public authorities, scientists and researchers, civil society, academia, companies, and human rights entities. The diversity of perspectives and the balancing of values and interests are very important for the legitimacy of appropriate decisions and regulations.

3. Transparency and Explainability. Transparency means minimally informing users about how the system works and that they are interacting with an AI system. Explainability involves clearly communicating the reasons behind AI decisions and enabling the possibility of questioning the results. Both requirements help mitigate concerns about the accuracy and impartiality of algorithms and encourage the responsible use of automation technologies.121

4. Safety. AI systems must be internally safe to avoid errors that produce undesirable results and protected against external attacks. Security in AI includes impact analysis, ensuring data quality, cybersecurity measures, and mapping the processes and decisions that make up the AI lifecycle (traceability).

5. Control and Responsibility. Human supervision or control is fundamental to ensure AI operates within the boundaries of legality, ethics, and justice. Despite the relative autonomy of its decision-making processes, responsibility must always lie with a natural or legal person. This way, in the event of improper or malicious use, one or both of them will be subject to civil, administrative, and criminal liability.

CONCLUSION

The role of knowledge is to comfort the afflicted and afflict the comforted.[122] This article claims to have fulfilled that role by demonstrating that AI has both potential and risks in nearly every area of application. At the political level, it can help improve the representative system and better capture the feelings and will of citizens. However, it also holds the potential to disseminate disinformation, hate speech, and conspiracy theories, misleading voters, weakening vulnerable groups, or spreading unfounded fears, bringing out the worst in people.

On an economic level, AI can increase productivity across various sectors, from agribusiness to industry, and significantly improve the service sector. Nevertheless, it can also concentrate wealth in the most favored activities and in the richest nations, exacerbating global inequality. On a social level, it can be a vital tool in addressing issues related to poverty and unjust inequalities. However, it also poses the risk of causing mass unemployment as workers are displaced by automation. There are also ethical dualities. A deeper understanding of human nature can elevate the humanistic or spiritual level of society, but it also carries the risk of diminishing the centrality of the human person.

> **"AI must be developed and oriented towards the well-being of people, countries, and the planet."**

In short, we live in an era of ambiguity and decisive choices. In the authors' view, the history of the world has been a constant—albeit non-linear—flow towards good, justice, and civilizational progress. We have come from times of harshness, human sacrifice, and despotism to the era of human rights. For this reason, it is possible to adopt a constructive vision and attitude towards AI. This requires neither paralyzing fear nor naivety or fantasies. We need legislation, regulation, and, above all, education and awareness among scientists, companies, and citizens to ensure we do not lose our way. Moreover, as already mentioned, the compass and the course indicated by the stars are the values that lead to the good life: virtue, practical reason, and moral courage. If we lose our reference points of goodness, justice, and human dignity, then it might indeed be time to let the machines take over and hope they can do better.

[120] Regulation must be seen as a necessary but insufficient condition. The treatment of the risks associated with AI, in this sense, goes beyond the dimension of law to reach other fields, including ethics applied to economics and programming. For Lucrecio Rebollo, "[c]onceiving Law as the only way to order and equalize digital society is a serious mistake. Law should be, as it always has been, a way of resolving social conflicts with a perspective to the public well-being, but in all cases it requires the collaboration of other areas of knowledge, of all the elements that make up the social structure." REBOLLO DELGADO, supra note 44 at 52.

[121] Sarah M.L. Bender, Algorithmic Elections, 121 MICH. L. REV. 489, 12 (2022).

[122] This is a paraphrase of the quote "The job of the newspaper is to comfort the afflicted and afflict the comfortable," attributed to a fictional character - Mr. Dooley - created by Chicago Evening Post journalist Finley Peter Dunne. See David Shedden, Today in Media History: Mr. Dooley: 'The Job of the Newspaper Is to Comfort the Afflicted and Afflict the Comfortable,' POYNTER (2014), https://www.poynter.org/reporting-editing/2014/today-in-media-history-mr-dooley-the-job-of-the-newspaper-is-to-comfort-the-afflicted-and-afflict-the-comfortable/.

Carr Center for Human Rights Policy
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

Statements and views expressed in this report are solely those of the author and do not imply endorsement by Harvard University, the Harvard Kennedy School, or the Carr Center for Human Rights Policy.

Copyright 2024, President and Fellows of Harvard College
Printed in the United States of America

This publication was published by the Carr Center for Human Rights Policy at the John F. Kennedy School of Government at Harvard University

Copyright 2024, President and Fellows of Harvard College
Printed in the United States of America



carrcenter.hks.harvard.edu

79 JFK Street | Cambridge, MA 02138
carrcenter@hks.harvard.edu