Inter-American Development Bank

fAIr LAC

Responsible and Widespread Adoption of Artificial Intelligence in Latin America and the Caribbean

Marcelo Cabrol
Natalia González A.
Cristina Pombo
Roberto Sánchez A.

Social Sector

TECHNICAL NOTE Nº IDB-TN-01839

January 2020

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library
fAIr LAC: responsible and widespread adoption of artificial intelligence in Latin America and the Caribbean / Marcelo Cabrol, Natalia González Alarcón, Cristina Pombo, Roberto Sánchez Ávalos.
p. cm. — (IDB Technical Note ; 1839)
Includes bibliographic references.
1. Artificial intelligence-Social aspects-Latin America. 2. Artificial intelligence-Social aspects-Caribbean Area. 3. Artificial intelligence-Moral and ethical aspects-Latin America. 4. Artificial intelligence-Moral and ethical aspects-Caribbean Area. 5. Social service-Technological innovations-Latin America. 6. Social service-Technological innovations-Caribbean Area. I. Cabrol, Marcelo. II. González Alarcón, Natalia. III. Pombo, Cristina. IV. Sánchez Ávalos, Roberto. V. Inter-American Development Bank. Social Sector. VI. Series.
IDB-TN-1839

IDB
Inter-American Development Bank

fAIr LAC

RESPONSIBLE AND WIDESPREAD ADOPTION OF ARTIFICIAL INTELLIGENCE IN LATIN AMERICA AND THE CARIBBEAN

Marcelo Cabrol, Natalia González A.,
Cristina Pombo, Roberto Sánchez A.

Summary

Artificial Intelligence (AI) presents a unique occasion to **promote equality of opportunities and improve the quality of life of all people** of Latin America and the Caribbean. Beyond its technological possibilities, the responsible and human-centered use of AI is essential, although it also poses major challenges.

The Inter-American Development Bank (IDB) advocates building a shared understanding of AI, its possibilities, and its uses, as well as its risks and the potential actions to mitigate them.

Towards this end, the IDB, in cooperation with partners and strategic allies, is leading the fAIr LAC initiative to promote the responsible use of AI in order to improve the delivery of social services (mainly in the sectors of education, health, social protection, and labor markets, and for issues related to gender and diversity). The initiative also aims to create development opportunities to reduce gaps and attenuate growing social inequalities. Working jointly with the private and public sectors, civil society, and academia, the fAIr LAC initiative will lead the implementation of AI system experiments and pilot projects. Similarly, it will create models for ethical evaluation and other tools for governments, entrepreneurs, and civil society to deepen their knowledge of the subject, provide guidelines and frameworks for the responsible use of AI, and look to influence public policy and the entrepreneurial ecosystem.

This document addresses some of the challenges and opportunities that AI has to offer to society, along with the lines of action that the fAIr LAC initiative proposes for Latin America and the Caribbean.

Acknowledgements

For their valuable contributions and time, the authors would like to express their gratitude, to the advisory group comprised of Carolina Aguerre, Constanza Gómez-Mont, Daniel Korn, Elkin Echeverri, Fabrizio Scrollini, Gemma Galdon Clavell, Hugo Morales, Javier Barreiro, Joana Varón, Mario Arauz, Norberto Andrade, Renata Ávila, Ricardo Baeza-Yates, Ana Lucía Lenis, Ulises Cortés, and Virginia Pardo.

The authors are also grateful for the support and advice provided by Arturo Muente, Jaime Granados, César Buenadicha, Claudia May Del Pozo, Elena Arias, Manuel Urquidi, David Rosas, Marcelo Perez, Luis Tejerina, Jennifer Nelson, Tamar Colodenco, Rodrigo Galindo, Alejandro Noriega, Carlos Amunategui, Daniel Castaño (Atticus), Joan Manuel López, and Lucía Camacho (Karisma Foundation).

Contents

01. Introduction

As we have changed the way we live and work over the past decade, artificial intelligence (AI) has gained the spotlight in discussions across multiple domains. While there is some consensus that we are still years away from achieving widespread use of AI, weak and narrow AI applications – where machine learning specializes in the execution of a single task – have already exceeded levels of human precision for specific tasks in visual recognition and natural language analysis in domains such as medicine and law (Ardila et al. 2019; Wood 2018).

In Latin America and the Caribbean (LAC), just as in many other regions, AI promises to enhance the efficiency of the delivery of social services[1] and the transparency of public decision-making, as well as stimulate the economy through increased productivity. AI has the potential to help society overcome some of its most important challenges, from reducing poverty to achieving progress in education, improving health care delivery and eradicating diseases, and increasing food production to address the needs of the world´s population.

AI has the potential to help society overcome some of its most important challenges

Yet a number of questions arise: How far is LAC from achieving this potential? How can we ensure that we do not incur greater costs and create a more unequal society in pursuit of these benefits? What should AI look like for its use to be trustworthy? How do we ensure the responsible use of AI?

The region must prepare itself to reap the benefits of trustworthy use of AI that **places humans at the center of all decisions, identifies the ethical and privacy challenges posed by AI, and then puts in place the measures and standards to mitigate those risks.**

1 This refers to the set of services and actions directed towards improving social well-being through the provision of information, services, and support, particularly for the education, health, social protection, labor markets, and social security sectors, and for issues related to gender and diversity.

While it is important to promote innovation and efficiency in this endeavor, accessibility is particularly important. Promoting economic value is certainly praiseworthy, but supporting an equitable distribution of wealth is even more laudable.

The use of technology should be proactive in promoting social values such as integrity, tolerance, and diversity, and in decreasing consumption levels that may deepen the problem of waste, pollution, and climate change. Furthermore, in terms of social organization, we must avoid a situation in which the massive use of technology becomes a solution for the poor but a personalized privilege only for the rich.

The Inter-American Development Bank (IDB) aims to achieve a shared understanding of AI and how it can evolve in the near future, along with the opportunities it offers, its sectoral applications, and the risks it presents and the potential measures to mitigate that risk.

To that end, the IDB, in cooperation with several strategic partners, is leading an initiative called fAIr LAC with the aim of **promoting the responsible development and application of AI to improve the delivery of services** – thus reducing existing disparities– and eventually reduce growing inequalities. The fAIr LAC initiative leverages the public and private sectors, as well as civil society, in order to make an impact on both public policy and the entrepreneurial ecosystem. The initiative was designed by the IDB in collaboration with C Minds, which received input from a diverse group of experts in the region[2]. This document introduces the strategic definition and lines of action that fAIr LAC proposes in order to tackle the challenges and seize the opportunities that AI has to offer to societies of Latin America and the Caribbean.

fAIr LAC with the aim of promoting the responsible development and application of AI to improve the delivery of services and eventually reduce growing inequalities.

2 Those experts included Carolina Aguerre, Constanza Gómez-Mont, Daniel Korn, Elkin Echeverri, Fabrizio Scrollini, Gemma Galdon Clavell, Hugo Morales, Javier Barreiro, Joana Varón, Mario Arauz, Norberto Andrade, Renata Ávila, Ricardo Baeza-Yates, Ana Lucía Lenis, Ulises Cortés, and Virginia Pardo.

02. Artificial Intelligence

The idea of intelligent machines was posed by Alan Turing (1950) in his theoretical article titled "Computing Machinery and Intelligence" that explored whether it is possible for a computer to simulate human intelligence and learn.

The term "artificial intelligence," coined by cognitive scientist John McCarthy in the context of the "Dartmouth Summer Research Project on AI"[3], would appear a few years later and be recognized as a foundational milestone of AI as a field of study (McCarthy et al. 1955).

To this day there is still no universally accepted definition of AI, since it is a dynamic field of science that has progressively evolved and given way to multiple technologies. However, it can be said that AI is a field that centers on the development of computational system capabilities for performing tasks that were traditionally thought of as exclusively belonging to human "intelligence." Hence, the problem lies in defining what is meant by intelligence.

Turing (1950) proposed performing the following conversational evaluation to define intelligence: a machine would be considered intelligent if it were to engage in conversation with a human in such a natural way that the human could not distinguish that his/ her conversation partner was a machine. This evaluation has encountered numerous objections as a test for determining intelligence, some of which even pre-dated Turing.

Almost a century earlier, in her memoir on the Babbage analytical engine, mathematician Ada Lovelace had noted that a machine could not be considered intelligent if all it could do was what it was ordered to do (Epstein, Roberts, and Beber 2008, p. 65).[4] This argument implies that intelligence requires both autonomy and the capacity to innovate. In the same way, John Searle (1980) argued that the Turing test was inadequate and could be overcome through the use of syntactic rules without requiring a real understanding of meaning or semantics, and thus was not be a demonstration of human intelligence. [5]

3 The purpose of this event was to conduct a series of workshops based on the conjecture that "every aspect of learning or any other feature of intelligence can, in principle, be so precisely described that a machine can be made to simulate it."
4 Ada Lovelace (1815–1852) was an English mathematician and writer known to some as the world´s first programmer for having written the first computer program.
5 For example, a team led by Vladimir Veselov designed a chatbot called Eugene who portrayed himself to be a 13-year-old Ukrainian boy. In 2014 he convinced most of the judges in a contest organized by the University of Reading that the bot was a real human. Yet, this victory is controversial, since every time the chatbot did not master enough English it would excuse itself due to its inability to express itself correctly.

In the years since, activities and games have been developed as a metric to compare the intelligence capability of humans and computers, ranging from a game of chess to a game of Go. These tests are also questionable under the same parameter, however, since machines are incapable of explaining their strategies.

Author Pamela McCorduck (2004) refers to the aforementioned dynamic as the "AI paradox." According to McCorduck, some innovations at a certain point in time were thought of as quasi-human signals of intelligence but then progressively became superficial tests until they eventually lost the privilege of being classified as AI. Simultaneously, newer technologies emerged that began to take on this role (McCorduck 2004; Cristianini 2014).

Some innovations at a certain point in time were thought of as quasi-human signals of intelligence but then progressively became superficial tests until they eventually lost the privilege of being classified as AI. Simultaneously, newer technologies emerged that began to take on this role"

PAMELA MCCORDUCK, 2004 Y CRISTIANINI, 2014

This determination to measure intelligence while focusing on a single task has led to the division of the scope and conception of AI into three goals in terms of usage:

- Weak or **narrow intelligence**, which aims for the learning process to specialize in the execution of a single task, achieving a precision level that is equal or superior to that of a human (this type of AI is the only one that has been achieved to date)

- General or **strong intelligence**, which aspires for AI to succeed at generalizing the learning process by applying it to diverse tasks with creativity and self-awareness.

- **Artificial super-intelligence**, a type of AI which would outperform the cognitive capacity of humans at every level (Bostrom 2014). Although divergent views exist, most experts in the field consider that it will take many years to achieve even the second goal of general AI. Some experts even maintain that we will never get that far.

Although divergent views exist, most experts in the field consider that it will take many years to achieve even the second goal of general AI. Some experts even maintain that we will never get that far.

Towards the end of this past decade, Hao (2019), writing in the MIT Technology Review, analyzed 16,625 knowledge products on artificial intelligence with the purpose of examining the evolution of the terminology and techniques used in this scientific field.[6] Since the term artificial intelligence was born in the 1950s, each decade has witnessed a different technology at the forefront of AI development (Hao 2019). Two approaches that arose from the first workshop at Dartmouth in 1955, and which have modeled AI progress up until now, are symbolic AI and sub-symbolic or connectionist AI (Annex I).

The latter is comprised of a group of adaptation-based techniques with statistical components that allow a system to learn in an automated way through the extraction of patterns and inferences, without the need to receive explicit instructions from a human. In the 1950s and 1960s, the mathematical foundations of what are today known as neural networks were first developed (Ro-

6 This was achieved by using arXiv, the largest open-source databases of scientific papers articles. The analysis covered papers published from 1993 to November 2018.

senblatt 1958).[7] But while they were used extensively, they did not have the processing capacity and/or the required amount of information to successfully adapt the model.

For this reason, from the mid-1970s through to the 1990s AI symbolic systems – also known as **knowledge-based systems** – dominated AI development. Based on the generation of behavior by way of deduction from logical rules or axioms, these systems found optimal decisions through predefined rules within a specific domain (Cristianini 2014). One of their most famous applications occurred in 1996, when IBM´s Deep Blue computer beat the world´s champion chess player Garry Kasparov. However, the limitations of expert systems became increasingly evident, in particular due to their lack of autonomously acquired knowledge and also the complexity of their construction: there were too many rules that had to be coded in order to create an "intelligent" system, which increased system costs and construction time.

Consequently, automated learning, better known as machine learning (a type of sub-symbolic AI), recovered its popularity by the end of the 1990s and at the beginning of the 2000s with the emergence of Bayesian network techniques[8] and support vector machines.[9] However, since 2010 neural networks have gained importance once again and have prevailed ever since (Hao 2019).

Within this historical context, three breakthroughs have made sub-symbolic AI become dominant and revolutionize the field of AI:

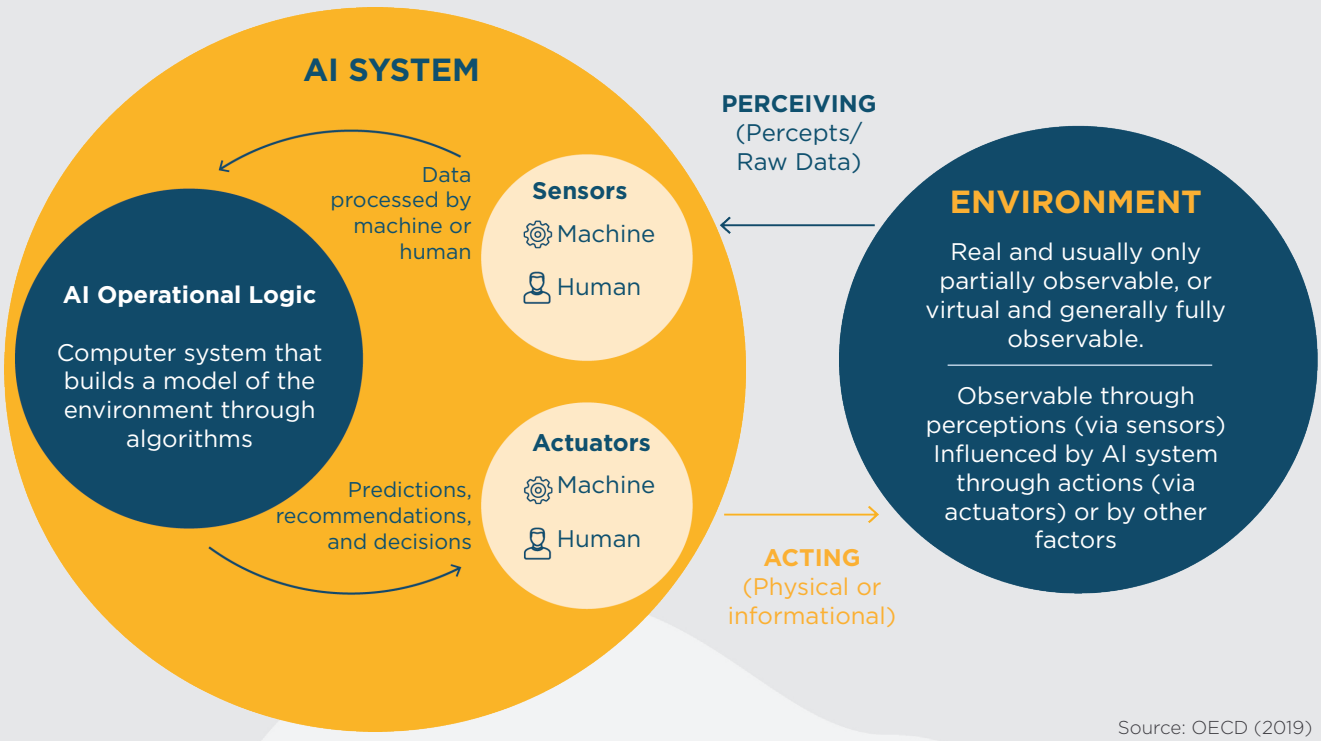
- **Improved neural network algorithms,** thanks to the efforts undertaken in 1986 by scientists Rumelhart, Hinton, and Williams (1986) in their back-propagation adjustment proposal [10].
- **Access and massive data recollection.**
- **An increase in processing capacity** through the development of graphics processing units (GPU).

The foregoing does not imply that symbolic AI has become obsolete, since its use remains extensive in planning and optimization applications. Furthermore, in recent years applications have been developed that combine both paradigms, such as "natural language processing algorithms [that] often combine statistical approaches (that build on large amounts of data) and symbolic approaches (that consider issues such as grammatical rules)" (OECD 2019, p. 27). Similarly, there is a global conviction that in order to continue with general AI development, sub-symbolic AI will be insufficient and therefore better ways will have to be found to combine both approaches.

7 The mathematical model upon which the neural networks (or "perceptrons," as they were known in those days) were based did not allow for the design of multilayered models, whereby, in practice, they could not be applied to scenarios different than those that contained negative or positive linear decision problems. This resulted in the failure to recognize the exclusive disjunction function, giving way to the criticism of Marvin Minsky and Seymour Papert (1969) of the first neural networks. This led to the virtual abandonment of the study of such networks for over 20 years.
8 A multivariate probabilistic model that relates a set of random variables in a directed graph that explicitly identifies causal influence.
9 A machine learning model whereby an algorithm centers on learning to discriminate between positive and negative results of a class of previously established n-dimensional vectors.
10 Backpropagation is an iterative supervised learning method used to train artificial neural networks through the error correction of a network node.

OECD (2019) includes both paradigms when describing AI as **"a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy"** (OECD 2019, p. 24). This definition covers systems that can use information or entries provided by humans or machines, whether it be through analysis algorithms or by manually using statistical inference to formulate options for information or action. Likewise, the definition focuses on the impact and influence that the technology may have on the social environment (Figure 1).

HIGH-LEVEL CONCEPTUAL VIEW OF AI (figure 1)



Source: OECD (2019)

The fAIr LAC initiative adopts the OECD´s definition of AI, and its lines of action will center on any system that can create information for autonomous decision-making (ADM/S)[11] in matters of social services delivery. This approach is independent from the type of learning and the type of algorithm, since AI model results constitute a data creation process that can interact in different ways with individuals responsible for making decisions to affect the environment, regardless of whether its result is introduced as a recommendation or if the system is capable of initiating an interim and final action.

Nonetheless, the fact that fAIr LAC uses a broad definition of what constitutes AI, the fAIr LAC initiative will mainly center on risk assessment and the creation of sub-symbolic AI mitigation mechanisms, specifically for machine learning algorithms (Annex I), since they have become so popular in recent years due to their potential to improve the delivery of social services.

11 Backpropagation is an iterative supervised learning method used to train artificial neural networks through the

02.1 The Potential of AI to Improve Social Well-being

Ever since discussions about technology for social well-being started, there have been numerous approaches to its use. The first applications centered on digital government and improving governance processes. Recently, the approach has shifted towards solving large-scale social problems such as education, poverty, and inequality, among others. To the extent that AI gradually becomes known as a technology that is accessible and has a general purpose in daily life, its impact will be greater in terms of its broader application to all aspects of human existence.

AI applications are diverse and their growth is noticeable in different spheres of life where patterns may be detected from big volumes of data and complex models, as well as in the availability of interdependent systems that can improve decision-making and generate more egalitarian and efficient policies. AI research areas such as natural language processing, computer vision, and classification and prediction models have significant potential to influence the welfare of society. In particular, these applications are being used in four major areas where their potential use and reach is immense: health and poverty, education, equity and social inclusion, and security and justice (McKinsey Global Institute 2018a).

A few examples serve to illustrate this last point. According to the World Health Organization (WHO), close to 138 million patients fall victim to medical errors every year, and 2.6 million of them die. During 2015, medical errors due to the incorrect use of pharmaceuticals and diagnostic errors represented the third cause of death in the United States and accounted for 10 percent of all deaths (GE Healthcare and UCSF 2016). In this sense, the strengthening of IT systems in the health sector will facilitate the future implementation of preventive algorithms that can make care safer and more efficient.

In the educational sphere, AI can be used to adapt course contents based on the progress and learning achieved by each student (McKinsey Global Institute 2018a). In terms of employability, AI can render labor intermediation more efficient, fair, and inclusive. Nowadays, intermediating and matching candidates to jobs are not the only areas where AI can make an impact on public employment services. With the use of AI, it is now possible to provide integrated services to those individuals who are seeking jobs as well as to company human resource departments, training centers, and the public at large, by providing them with relevant employment information.

02.2 Ethical and Responsible AI

The search for and definition of ethical AI is still an issue under discussion. In recent years, national and international organizations have created expert committees to debate these challenges and have published a set of principles (Jobin, Ienca, and Vayena 2019; Mittelstadt 2019). Ethical principles offer high-level guidance on what must and must not be done in the development and deployment of AI systems by all of those parties that have active functions in the lifecycle of an AI system (including its development, deployment, operation, maintenance, etc.).[12]

The fAIr LAC initiative will use the OECD´s ethical principles, adopted in May 2019, as a development guide (Box 1). These principles have been adhered to by all OECD member countries and by six non-member countries, as well as by the G20 in June 2019 (OECD 2019).

While six Latin American countries (Argentina, Brazil, Colombia, Costa Rica, Mexico and Peru) have adopted the OECD principles, the application of specific recommendations in this regard is still incipient. With the purpose of adopting responsible AI in the region, the fAIr LAC initiative will concentrate on defining challenges that arise from trying to operationalize ethical principles so as to create adequate implementation strategies on the basis of LAC issues and perspectives.

List of the OCDE's Ethical [12] (Box 1)

Inclusive Growth, Sustainable Development, and Well-Being

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

Human-Centered Values and Equity

AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor rights. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

Transparency and Explicability

AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: (i) to foster a general understanding of AI systems, (ii) to make stakeholders aware of their interactions with AI systems, including in the workplace, (iii) to enable those affected by an AI system to understand the outcome, and, (iv) to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

12 The OECD (2019) describes the AI system lifecycle as follows: (i) design, data, and models, (ii) verification and validation, (iii) deployment or publication, and (iv) operation and monitoring.

Robust, Secure, and Safe

Robustness, security, and safety are essential elements of every AI system for the following reasons:

- AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.
- To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.
- AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.

Accountability

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

Source: OECD (2019)

CHALLENGES IN LAC

DATA SOURCES

Past discrimination

Non-representative samples for populations of interest and missing data

MODEL DEVELOPMENT

Leakage of Information

Temporary changes

Over-adjustment and under-adjustment

GOVERNANCE AND SECURITY

Governance and cyber-attacks

Illegal AI penetration and adversarial attacks

USE AND DECISION-MAKING

Algorithmic bias (output)

Disparate impact (outcome)

CONCEPTUALIZATION AND DESIGN

Necessity and proportionality

Correct definition of the public policy problem and response

ACCOUNTABILITY

Interpretability and explicability

Informed consent and limits on privacy

Accountability and responsibility

DIGITAL INTEROPERABILITY AND TRANSFORMATION

Availability, interoperability, and data quality

Lack of technical

SYSTEMIC IMPACT

IHuman – AI interaction

Economic, educational, and employment impacts

Competition and monopolistic consolidation of information

Source: Prepared by the authors.

02.3 AI Challenges

As AI begins to expand, along with the recognition of its benefits and opportunities, some voices have also warned about the challenges posed by its responsible adoption and the undesirable outcomes it may bring about through its widespread incorporation into the processes and decisions of society (Figure 2). These concerns are highly varied and do not only center on AI; they also consider the impact on the AI ecosystem and on the way in which human beings should interact with and relate decisions to AI.

In order to create successful mitigation tools that allow for designing systems that have a positive impact on individuals and society, it is critical to understand the types of challenges and their differences. The list in Figure 2, prepared on the basis of discussions that took place in multidisciplinary workshops with regional experts,[13] contains some of the most important challenges on which fAIr LAC will focus its lines of action.

Digital Interoperability and Transformation

The following are some of the challenges resulting from digital transformation and the technical capabilities that it requires:

- **Availability, interoperability, and quality of data:** The availability of data is a major requirement to be able to adjust AI systems to machine learning models. While it is true that in recent years LAC countries have made progress in the digitalization of social services, there is still significant fragmentation between these services. In addition, a bias is detected in the region in terms of low or no data available for rural areas with high marginalization levels. There is a great need for the continued development of interoperable digital systems so that relevant information "be made available to authorized staff and contain precise, complete, necessary and sufficient data. All the foregoing within an adequate legal and regulatory framework that respects privacy standards, ethics, legislation and applicable regulations" (Pombo et al. 2019, p. 12).

- Lack of technical capacity: In 2017, PricewaterhouseCoopers estimated that AI could generate US$15.7 billion for the global economy, which would represent an increase of 14 percent in global GDP (PwC 2017). The greatest economic benefits of AI were in China (26 percent increase in GDP for 2030) and North America (14.5 percent increase in the same period). According to the PwC report, if the current trend continues, LAC would only capture 5.4 percent of the global increase, leaving it at a competitive disadvantage and increasingly lagging behind other parts of the world. In order to increase the AI adoption

13 This refers to fAIr LAC advisory group experts and participants. For additional information, visit https://www.iadb.org/en/fairlac

rate in the region it is necessary to create technical capacity and promote its responsible use. To do this, LAC needs to make efforts to update the available expertise related to AI, not only to train enough specialists in this field of knowledge, but also to enable a large number of persons to live and work with AI systems. The capacity-building of human resources is fundamental to absorb and harness these new technologies (McKinsey Global Institute 2018b).

Conceptualization and Design

Efforts to assimilate AI into the formulation of relevant public policy face the following challenges:

- **Necessity and proportionality:** Although AI has sufficient potential to improve processes and reduce inequalities, AI is not a panacea. Its limitations must be understood so as to not fall into what Morozov (2014) calls "technological solutionism,"[14] that is, believing that technology in itself has the capacity to solve all social problems even in the absence of adequate public policies. Accepting the limits implies acceptance that there are social issues for which the creation of an AI system may have little impact or not be adequate. This means that for every instance for which it is considered a necessity to apply AI and perform proportionality analysis, consideration will need to be given to such issues as the risk of affecting individual rights, the number of persons who may be affected, and their level of vulnerability.

- **Correct definition of the public policy problem and response:** In line with the preceding point, there is a risk of proposing AI projects from a technology perspective and not on the basis of a particular social problem. A necessary and functional AI project may present risks if the correct public policy is not developed. In sum, it is necessary to separate AI´s real capabilities as an information-generating system from the responsibility of public policy decision-makers to design interventions to solve social issues.

14 Morozov (2014) identifies "technological solutionism" as an endemic ideology that reformulates complex social phenomena such as politics, public health, education, and law enforcement as "neatly defined problems with a definite, computable solution or as transparent and self-evident processes that can be easily optimized, if only the right algorithms are in place."

There is a risk of proposing AI projects from a technology perspective and not on the basis of a particular social problem. A necessary and functional AI project may present risks if the correct public policy is not developed.

Governance and Security

A major theme that arises in the AI field relates to the security of its infrastructure. The challenges here include:

- **Governance and cyber-attacks:** Information risks are related to governance and security protocols that are used all along the project lifecycle. In recent years, there has been an increase in personal data breach cases in LAC. Occasionally, risks are triggered by human error due to lack of knowledge and understanding about security standards and good practices. For example, in 2016 the voter registry of Mexico´s National Electoral Institute, which contained information on nearly 94 million people, was exposed without a password in a storage system due to human error (Baraniuk 2016). Although the lack of protocols and theft of information through cyber-attacks are not exclusive to AI, the growing expansion of AI applications in digital services does increase the exposure of citizens' personal data. In the years to come, it will become increasingly important that the region implement information security standards and protocols.

- **Illegal AI penetration and adversarial attacks:** Several papers on computer vision and neural networks for image classification explore the way an algorithm can be confused with opposite examples. Szegedy et al. (2014) added images created synthetically with quasi-imperceptible perturbations, whose purpose was to make the model classify them incorrectly. The applications of these types of attacks may be very different. For example, in the case of facial recognition as a widely used security measure on mobile phones, illegal penetration would aim to confuse the algorithm by making it classify the face of the attacker as if it were the face of the owner of the device, based on the owner's personal information.

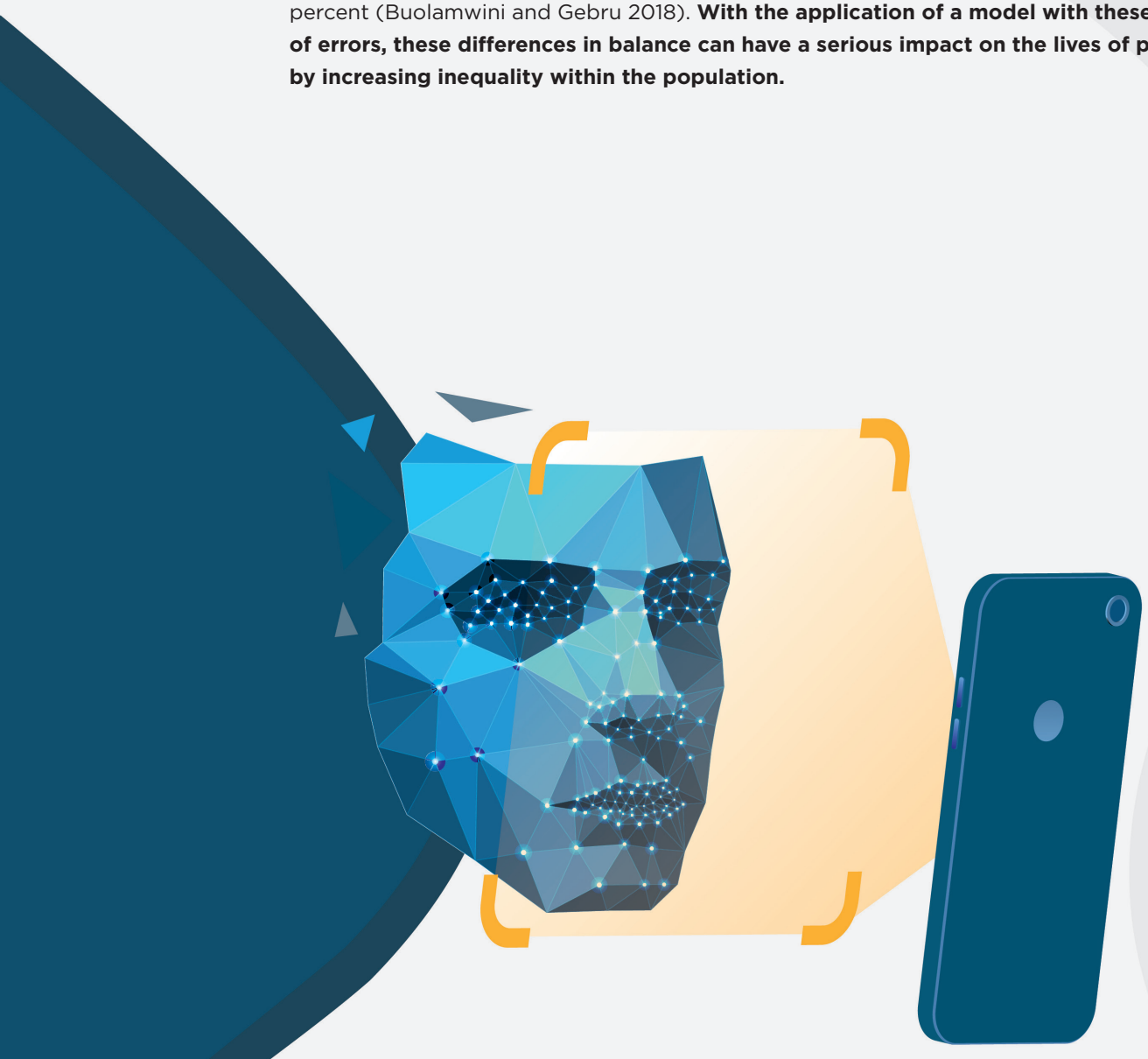
Data Sources

Challenges here stem from preexisting features in the data that will be used to adjust the model. The challenges are a result of phenomena intrinsic to the training data, mainly involving failing to understand the data-generating process and trying to reach results on populations not represented in the samples or through historical discrimination phenomena already contained in the data. These phenomena can create biases by replicating behaviors observed in the data, privileging one group or damaging another.

- **Past discrimination:** When learning happens with historical information or one that is previously tagged, data used for training may contain implicit biases observed in society. For example, in 2015 Amazon piloted a human resources recommendation system using supervised learning techniques. The model provided training with historical results from candidate selection processes of the previous 10 years. Every curriculum was assigned a binary label: 1 if the candidate had been accepted for the position and 0 if the candidate had been turned down. What the model did not account for was that the technology industry has been characterized as predominantly masculine,[15] so an algorithm trained with this information was capturing these patterns of exclusion and ended up consistently recommending a higher proportion of men, thus reinforcing gender discrimination (Dastin 2018).

15 In recent years, many companies have launched initiatives to achieve gender parity.

- **Non-representative samples for populations of interest and missing data:** This error occurs when a model is trained on the basis of a database that is not representative of the population to which it seeks to generalize,[16] or when observations for the different subpopulations are insufficient. This can cause the model to generalize with a sufficient degree of precision for the population as a whole, but not for subpopulations specifically (Guo et al. 2008). One application where this phenomenon has been identified repeatedly is computer vision and facial recognition. A study by Buolamwini and Gebru (2018) indicated that datasets used to train commercial face recognition services were overwhelmingly comprised of lighter-skinned subjects. This training showed an error rate for white males of 0. 8 percent, while for darker-skinned women the error rate was 34.7 percent (Buolamwini and Gebru 2018). **With the application of a model with these types of errors, these differences in balance can have a serious impact on the lives of people by increasing inequality within the population.**



16 There is a significant difference between an external validity error, as it is understood by studies in econometrics, and class imbalance errors for machine learning models. Seeking to create causality arguments, the external validity error use representativeness to describe a phenomenon that affects the population. On the other hand, machine learning models aim to minimize errors in the prediction of observations that are external to training.

Model Development

These challenges are generated in the development of algorithms, mainly due to methodological errors in the execution of a model´s training process and in the management of information. These errors may also create biases or can worsen the performance of systems over time.

- **Leakage of Information:** This error is essentially methodological and is produced when, during model design, appropriate division is not performed between training subsets, testing, and validation. This leads the model to learn and evaluate with the same information, which results in high precision levels that are unrealistic. A comparison with a real-life event would be to take an exam when you already have the correct answers. The student will undoubtedly receive a passing grade, but that does not mean the student is prepared to solve those problems independently, since he or she may have only memorized the answers.

- **Temporary changes:** This error is generated when a model that had a good precision level at one time continues to be used indefinitely, while the real world is complex and changes constantly. This is why the application of these types of models requires a responsible team to continuously ask whether the assumptions and the way in which the model is used continue to be useful for society.

- **Over-adjustment and under-adjustment:** In machine learning models trained with supervised analysis, the goal of the training is to generalize that learning. This means that the relationships or patterns learned by the model through the observation of a portion of data allow them to obtain useful information to classify examples that have not yet been observed. When an algorithm is over-adjusted, it can perfectly learn the training data at the level that negatively impacts performance of the model in information outside of that subset of learning. That is, it may possibly not obtain good results when classifying new data. Furthermore, sub-adjustment occurs when the model is too simple and does not learn enough to allow it to create a useful classification. Both phenomena can occur for subpopulations of a population when there is an over-adjustment for groups with specific characteristics or an under-adjusted for other groups.

Use and Decision-making

These challenges occur when the algorithm is being used for decision-making in a real situation. They relate to the effect that the use of an algorithm can have by segmenting or categorizing the population, and they describe the risk that arises from the disconnection between the algorithm´s decision-making and the process administrator.

- **Algorithmic bias (output):** This occurs when the AI system makes systematic errors that create unfair outcomes for subpopulations or specific individuals. As witnessed in the previous challenges (data sources: model development), while these biases can be created by past discrimination, class imbalance, information leakages, temporary changes, and over- and under-adjustments, among others, their impact comes once they are used to take a decision or public policy action. The way the system evaluates whether the algorithm is making an unfair decision is not general for every problem and can even be different for the same problem in different contexts or countries. This is because what is understood as "justice" can vary according to the culture and/or tradition of a given population group, or in those instances when a society considers that it is necessary to implement positive discrimination policies during a determined timeframe in order to offset a prior discrimination situation (Table 1). Open discussions in this regard must be fostered, since having clear definitions is the only way to operationalize the mitigation of risk considering the specific conditions of every application. pacto desigual no requiere que exista un sesgo algorítmico en la recomendación (output), pues describe una situación en la cual el uso del producto o sistema impone una carga desproporcionada para los miembros de grupos específicos, creando la posibilidad de que su uso amplíe el estado de marginación o exclusión. Por ejemplo, las evaluaciones automatizadas de riesgos de reincidencia criminal, aunque estén bien calibradas mediante una definición de justicia como "paridad demográfica", pueden resultar en efectos acumulativos muy marcados para ciertos grupos, perpetuando así una condición de marginación.

- **Disparate impact (outcome):** A disparate impact does not require the existence of a biased algorithm in the recommendation (output) because it describes a situation where the use of a product or system imposes an unreasonable burden on the members of spe cific groups, raising the possibility that its use will increase their marginality or exclusion. For example, while automated criminal recidivism risk assessments may be well calibrated through a definition of justice as "demographic parity," they can result in very marked cumulative effects for certain groups, thereby perpetuating their marginalization.

SELECTED DEFINITIONS OF JUSTICE [17] (table 1)

Definition of Justice	Description
Counterfactual Justice	Considers that a predictor is "fair" if its output remains the same when the protected attribute is flipped to its counterfactual value (for example, when introducing a change in race, gender, or other condition).
Demographic Parity	Establishes that the portion of each segment of a protected class (gender, for example) should obtain an equally positive result (such as the assignment of scholarships).
Equal Opportunities	Assuming that all persons have the same qualifications, this implies that each group should obtain a positive result under equal conditions.

Source: Prepared by the authors.

Accountability

One of the major challenges of AI systems is the type of information collected about persons and the pressing need to safeguard it and ensure that it will only be used for the purposes specifically authorized by the concerned individual. Several challenges are faced in this regard:

- **Interpretability and explicability:** These two related terms are used to describe the level of understanding that one may have of models. On the one hand, interpretability is the capacity to observe a system bi-directionally, which involves both understanding the reasons for making a concrete prediction, such as predicting what will happen next if a change is made to any of the input variables. On the other hand, explicability is a broader concept that describes the capacity of understanding the functioning of a model in human terms by considering its inputs, outputs, and parameters. For the most part, symbolic algorithms are highly explicable and interpretable because they follow logical rules. That is not always the case with sub-symbolic algorithms (especially those based in deep neural networks), and in general the application of other methodologies (sometimes called "black box" models) is required to achieve interpretability. In many countries of Latin America, institutions are legally

17 An example is the Local Interpretable Model-Agnostic Explanations (LIME) Methodology.

required to substantiate the reasons for granting or denying social rights services and, in turn, individuals have the right to challenge such decisions, which compels institutions to develop the capacity to explain the outcomes of models.

- **Informed consent and limits on privacy:** Informed consent implies that an individual who is the owner of personal data understands and accepts how they will be used or treated. This requirement becomes very complicated when referring to AI applications and technology in general, because the fast pace of innovation makes it difficult for regulatory authorities, experts, and opinion makers to develop standard laws and good practices.

- **Accountability and responsibility:** This challenge arises from the lack of legal clarity about a system´s responsibility over its decisions. For example, if an algorithm decides the order in which patients arriving to an emergency room are to be attended and one of them dies for lack of timely care, who assumes the responsibility for that decision? Furthermore, financial resources for indemnity need to be appropriated. Nowadays, different enterprises rely on business model programs such as Software as a Service (SaaS).[18] If a third party makes a contract with one of these platforms and ends up improperly using a product or service or causing quantifiable damage, to what extent is the SaaS supplier company responsible? Furthermore, how can we ensure that people who are affected are able to dispute outcomes? The need to establish accountability frameworks cannot be underestimated.

Systemic Impact

These are impacts presented in an indirect manner in the system, understood as the social environment. The main challenges are threefold:

- **Human – AI interaction:** In this instance, the challenges involve the indirect implications of AI use in social sectors, including the definition of the role of the user as a critical recipient of AI system recommendations for decision-making. Relying on new technologies is an increasingly widespread practice in the public sector. The optimal way for responsible decision-makers (users) to use AI tools is combining results with their own professional intuition. Nonetheless, the National Endowment for Science, Technology and the Arts (Nesta)[19] after an extensive literature review, interviews with public officials, and discussions with experts, found that some users simply ignore the results of these tools (algorithmic aversion),[20] while others resort to their technical knowledge and generally biased common sense to report on the decision process. This implies that bias persists as a feature of human decision-making, despite the introduction of tools such as AI (Snow 2019). Given the existing limitations of AI systems in high-risk situations, as well as aversion towards the application of results on the part of some users, one could suggest that AI

18 SaaS is a model where the system or application is administered and maintained by the provider; the client only consumes the outcomes without making any changes or adjustments.
19 Nesta is an innovation foundation based in the United Kingdom. See Snow (2019).
20 People are averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithms than they do in people after seeing them commit a mistake (Dietvorst, Simmons, and Massey 2015).

only be employed as a support and as a complementary input for information under consideration. Nesta highlights three key principles in the interaction between AI and human beings that, if taken into account, would improve the adoption and utilization of an AI system´s recommendations: context, understanding, and agency (Snow 2019).

- **Economic, educational, and employment impacts:** This challenge includes the changes that are taking place in job markets by way of massive automation of functions and tasks, while from another perspective this also includes the effects of the emergence of new jobs and the need for people to acquire the skills required for those emerging jobs.

- **Competition and monopolistic consolidation of information**: In general, it can be said that a sub-symbolic AI model will be more accurate if trained with the largest quantity of information, as long as it is of good quality.[21] For example, the privileged position of consolidated governments and enterprises regarding data collection access and capacity may favor the emergence of information monopolies. A specific case can be observed when an AI product becomes a market leader by increasing the number of users of its services. The larger the number of users, the larger the volume of information that feeds the system, which in turn will improve the model's quality. This generates a winner-take-all market. This means that a given product will gain a larger proportion of users and income for a certain class of products and services, due to its slightly higher quality than competitors' products. Furthermore, it will consolidate its leadership by capturing a larger market share. It is important that mechanisms to reduce such effects be found during the development of solutions geared towards providing social services. This can be achieved by promoting open data projects and collaborative initiatives as an alternative to AI construction, and by promoting the creation of structures such as data trusts and data commons.[22]

21 There are cases when a greater amount of data may not lead to a better model or superior methodological approximations (such as Bayesian statistics), especially if the lack of information can be offset by expert knowledge. However, the relationship between data quality and data precision is unquestionable.
22 "Data trusts" refers to a repeatable framework of terms and mechanisms for managing information. "Data commons" can refer to both the technological platform for storing and manipulating shareable data sets, or to the set of principles and governance strategies for the use of those data sets.

03. The fAIr LAC initiative

The Latin America and Caribbean region has made major advances in reducing poverty and inequality over the past decade. However, social inequality in the region continues to be a significant problem, and LAC is considered to be the most unequal region in the world.

Since 2015, setbacks have been recorded, particularly regarding the 167 million people that still live in extreme poverty (ECLAC 2019). A large majority of this impoverished population live in rural areas with insufficient and low-quality access to basic health, water, and sanitation services. This population is also burdened with a legacy of discrimination based on gender, race, and social class. In this context, LAC should strive to find new and better ways to reduce poverty, promote economic growth, and foster the fair distribution of wealth. AI emerges as an innovative tool to achieve a greater socioeconomic impact in these areas.

LAC should strive to find new and better ways to reduce poverty, promote economic growth, and foster the fair distribution of wealth. AI emerges as an innovative tool to achieve a greater socioeconomic impact in these areas.

The 2019 Government AI Readiness Index,[23] produced with support from Oxford Insights and the International Development Research Centre (IDRC), shows that countries of the region face challenges in three areas when trying to benefit from using AI for the common good: adequate policies, capacity, and resources. First, LAC currently has neither a coherent **policy** nor defined ethical standards, although Mexico, Colombia, Uruguay, and Argentina are in the process of setting out AI policies and strategies. For example, Colombia defined its Digital Transformation

23 The Government AI Readiness Index is a classification system created by Oxford Insights and the IDRC. The index is the sum of an average normalization of indexed metrics on a scale of 0 to 10 taken from sources including the United Nations, World Economic Forum, Global Open Data Index, and World Bank, in addition to Gartner, Nesta, and Crunchbase. These metrics are grouped under four high-level topics: governance, infrastructure and data, skills and education, and government and public services.

and Artificial Intelligence National Policy in the CONPES 3975 document, identifying specific guidelines that, when implemented, will generate a coherent policy framework for the ethical and responsible development of AI.

Second, **capacity** is a challenge for the region and for its governments, in particular. Although there are some enterprises and scholars working in the AI field, there is an absence of thorough knowledge of AI in economic sectors and no clear understanding about its applicability in the public sector. Lastly, when compared with countries such as Canada, the United States, and the United Kingdom, the Latin American nations still have not been able to connect their public and private capital with the technical and academic resources available in the region to establish AI development hubs.

The untapped potential is significant. It is estimated that by 2035 AI development could add 1 percentage point to the annual economic growth of LAC (Ovanessoff and Plastino 2017). To take advantage of this opportunity, policymakers, entrepreneurs, and civil society should view AI as a tool that has the potential to generate economic growth and social well-being in the long term, and not just another productivity engine. In turn, the analysis of current and emerging AI risks and their impact on the labor force will help create a mitigation strategy to address those issues. Even though nowadays AI tends to be considered a black box, it is possible to demand transparency, explicability, and traceability throughout its development and implementation.

For this purpose, the IDB is leading fAIr LAC with the support of different strategic partners. The initiative seeks to promote responsible application of AI to improve the provision of social services and thereby lessen growing social inequality in the region. fAIr LAC leverages the public and private sectors and civil society in order to influence public policy and the entrepreneurial ecosystem.

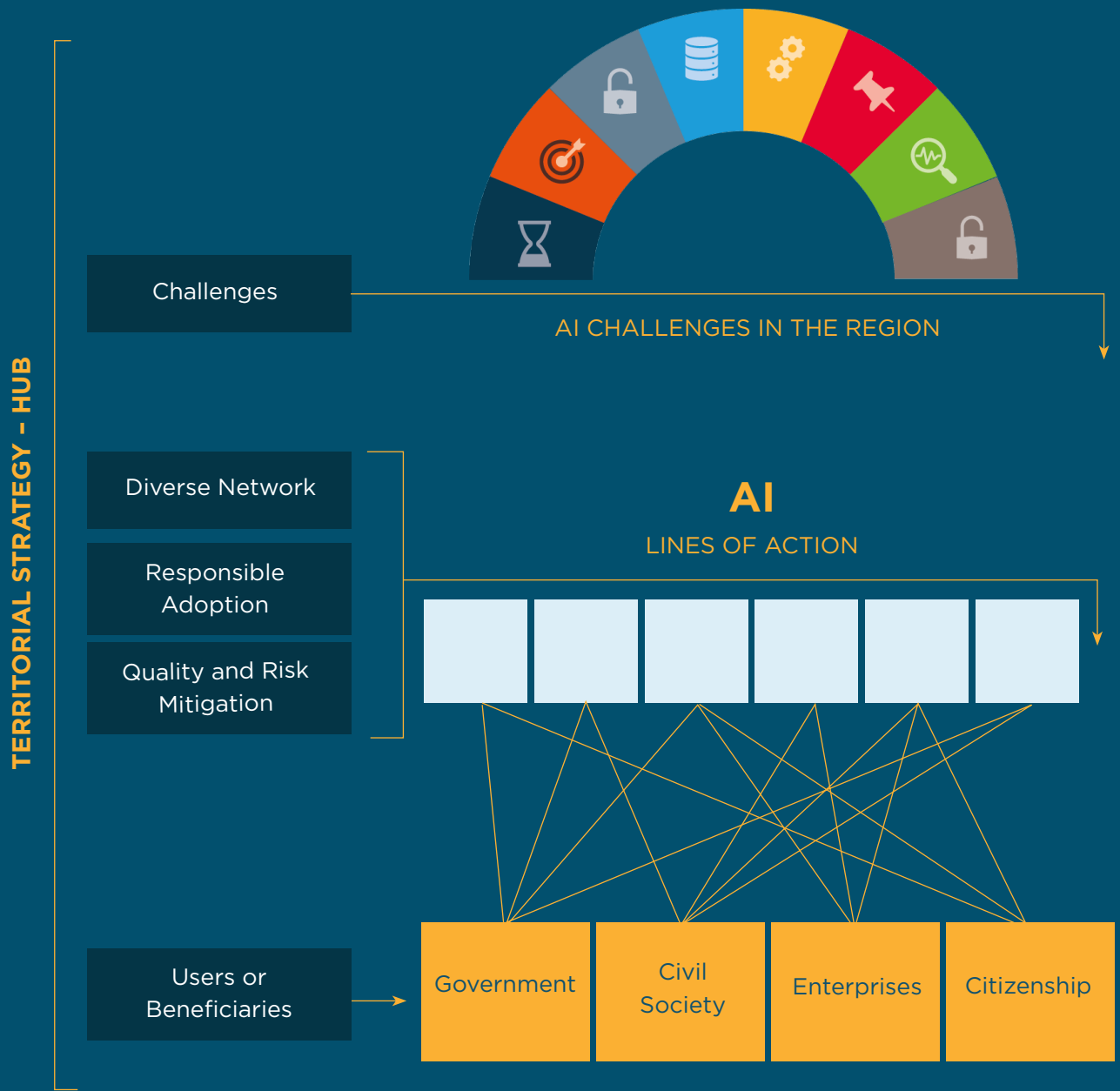
Strategic Dimensions and Lines of Action

Given the challenges that have been identified and the scope determined for fAIr LAC, three strategic dimensions have been suggested that articulate strategies to achieve the objective to promote the responsible use of AI in the provision of social services. These three dimensions are (I) development of a diverse network, (II) training for responsible AI adoption, and (III) the promotion of quality and risk mitigation (Annex II).

Within each strategic dimension, lines of action are set forth with activities and specific interventions targeted at one or several of the identified users or beneficiaries, namely government, civil society organizations, enterprises, and citizenship (Figure 3).[24]

24 A detailed account of every line of action exceeds the scope of this study. This will be further addressed in specific documents in the future.

fAIr LAC Strategic Dimensions (figure 3)



Source: Prepared by the authors.

03.1 Dimension 1: Develop a Diverse Network (Highlight, Diffuse, Build, and Connect)

To ensure that a variety of voices are heard and the challenges that affect the LAC population are identified, it is necessary to create a diverse network that contributes to foster discussions about AI, its potential, and its consequences. The goal is to **create forums for meetings, dialogue, and exchange** such as those described below where citizens, enterprises, and the government can generate and share knowledge.

- ***AI Experts Consulting Group.*** This is a network of professionals and experts from academia, government, civil society, industry, and the entrepreneurial sector that advises on the development and implementation of fAIr LAC lines of action. The group will promote an ethical application of AI in LAC and emphasize an understanding of the consequences and particularities of the regional context regarding global discussions. In addition, it will provide guidance to government initiatives on the responsible use of AI.

- ***Observatory of Use Cases.*** A platform will be created with fact sheets for those AI use cases in social services delivery that are being developed in the region by governments, enterprises, academia, and civil society. The objective is to provide a regional public asset to make the public and private sectors more aware of successful cases, identify good practices, and systematize learning through applications that countries and/or enterprises can reproduce, thus giving visibility to the region´s best practices. The observatory will be constantly updated with the aim of making it a documentation benchmark on the state of AI in the region, including initiatives unrelated to the IDB and/or financed by the Bank.

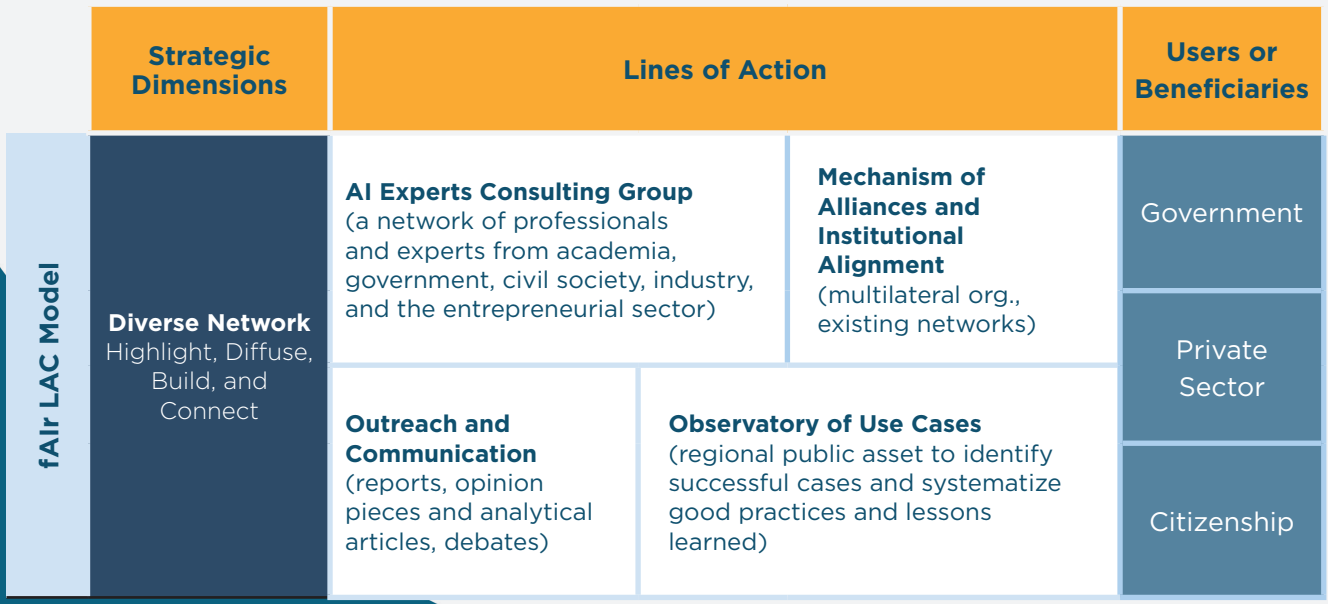
- ***Alliances and Institutional Alignment Mechanism.*** The purpose of this mechanism is to identify those AI initiatives that are being implemented simultaneously by partner institutions in order to explore cooperation synergies and opportunities between them, as well as to establish institutional alignment at the regional and international levels.

- ***Outreach and Communication.*** There is no reason to believe that the region will become a leader in the field of AI in the near future, but there is reason to believe that AI will at least be generally applied, used, and developed in LAC countries.

The objective is to provide a regional public asset to make the public and private sectors more aware of successful cases, identify good practices, and systematize learning.

In this regard, it is vital that decision-makers, public officials, entrepreneurs, and citizens understand the advantages and risks of AI. For this purpose, fAIr LAC has an outreach strategy that includes knowledge dissemination and communication media campaigns (Figure 4).

DIVERSE NETWORK AXIS LINES OF ACTION (figure 4)



Source: Prepared by the authors.

03.2 Dimension 2: Capacity Training for Responsible Adoption of AI

To ensure that the region adopts AI in a responsible manner, it is first important to help policymakers, citizens, and the entrepreneurial ecosystem fully understand the challenges and opportunities posed by AI. Furthermore, the region must stimulate the establishment of training opportunities that involve public officials, civil society representatives, and entrepreneurs (Figure 5). Different actions and products are being considered towards this end are discussed below.

- ***Experiments and Pilot Projects.*** Support programs (consulting, financing, mentoring networks) will be created for experiments and use cases and have two objectives: to accumulate institutional experience in the application of analytical and practical projects; and to systematize application experiences where AI helps create a greater social impact while respecting human rights. Projects that can be scaled and emulated in the region will be prioritized, considering at all times the need for domain adaptation given the multiplicity of contexts and different needs of every country, city, and municipality.

- ***Incentives System.*** This system will establish incentive models for public officials, enterprises, and citizens (acknowledgments, access to finance and resources, institutional endorsements, challenges).

- ***Diverse Tools.*** The main idea here is that public officials and entrepreneurs must be able to access AI educational training, as well as understand the benefits and risks of AI, and that they can deepen their knowledge on the subject through the development and publication of guides, frameworks, and other methodological tools for the responsible use of AI.

RESPONSIBLE ADOPTION AXIS LINES OF ACTION (figure 5)

	Strategic Dimensions	Lines of Action		Users or Beneficiaries
fAIr LAC Model	Responsible Adoption (AI for social services)	**Support Program for Experiments and Use Cases** (consulting, financing, mentoring network)	**Pilot Projects** (with scaling potential in the region)	Government
		Incentives Model acknowledgments, access to finance and resources, institutional endorsements)	**Diverse Tools:** -Methodologies, guides, impact evaluation frameworks - Open Model Repositories and Tools Index	Private Sector

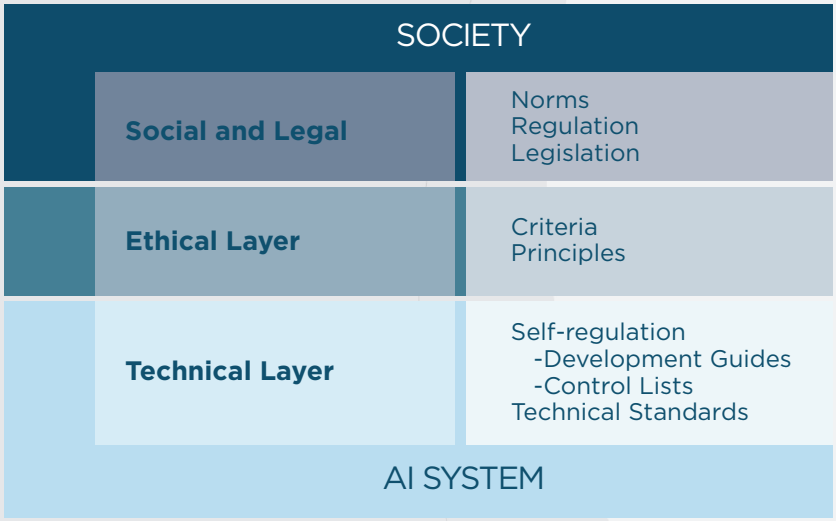
Source: Prepared by the authors.

03.3 Dimension 3: Quality Promotion and Risk Mitigation

In light of AI´s transformational benefits, as well as its risks, the design of a public policy that balances these two factors must be prioritized. Societies must establish monitoring mechanisms throughout the lifecycle of AI systems. This includes guaranteeing the development of algorithms with verification of minimum standards, auditing of data quality and design processes, establishment of security protocols and governance of personal information, and creation of legal mechanisms for responsibility and accountability.

Varied risk mitigation mechanisms may be established, including self-regulation, issuance of standards, audit processes, and official regulatory frameworks (Figure 6). All of them are desirable – they differ only in terms of whether or not they are legally binding. Gasser and Almeida (2017) propose a layered governance representation model to regulate the use of AI systems when interacting with society. The model presents different instruments that can be developed in parallel. The fAIr LAC initiative will encourage their development at the regional and territorial levels.

LAYERED GOVERNANCE MODEL (figure 6)



Source: Adapted from Gasser and Almeida

- ***Guides and Guidelines.***
  The fAIr LAC initiative will work to create tools that can guide developers and those persons responsible for the formulation of public policies, and on how to approach the necessity and proportionality challenges,[25] as well as on policies concerning development and infrastructure.[26] These documents will cover fields of study that include the following subjects:

  - ***Evaluation of necessity and proportionally:*** With the support of the IDB´s Social Sector specialists and other regional network experts, proposed AI systems to be applied to social services in pilot projects will be evaluated with a comprehensive development analysis that considers the effect that resulting public policy proposals may have.

  - ***Data governance:*** Information security standards, protection of personal data, and entrepreneurial governance and interoperability processes will be described here.

  - ***Evaluation of algorithmic equity:*** Non-discriminatory criteria such as demographic parity, equal opportunities, and threshold groups, among others, will be addressed (Dignum 2019; Gajane and Pechenizkiy 2018).[27] Since these criteria depend on the culturally determined concept of equity, their analysis will be conducted on case studies of pilot projects implemented by fAIr LAC.

- ***Self-regulation.***
  Since industry generally moves forward faster than government and regulatory frameworks, encouraging self-regulation among enterprises is a very important step to generate good practices because it serves as a mirror image of the main problems faced by enterprises developing cutting-edge technologies. Guides and checklists are some of the tools for internal use that establish minimum quality levels for processes. It is important to mention that there are limitations to self-regulation as a governance tool, since enterprises have incentives that could run in the opposite direction of general societal well-being. But even then, the fAIr LAC initiative considers self-regulation to be a key initial step that, due to the characteristics and accelerated pace of implementation of AI, will allow for experimenting with norms and guidelines and evaluating them. Nevertheless, this should be accompanied by an effort to develop and put in place official standards and regulatory frameworks. For this reason, fAIr LAC seeks to create an ethical self-evaluation process that can be implemented by any enterprise and that allows for the identification of main risks and the appropriate actions to mitigate them.

25 See the section in this document entitled "Conceptualization and Design."
26 See the section in this document entitled "Model Development."
27 These concepts are not defined in this document, since they exceed its scope. This will be further addressed in fAIr LAC documents in the future.

- ***Standards.***
  Standards are a set of good practices related to specific systems or tools that aim to establish a quality system. Unlike self-regulation, for standards there are independent government agencies with technical committees that validate a compendium of standards. The formulation of standards allows for issuing certificates and conducting audit processes that generate market incentives, to the extent that they operate as a differentiating factor in a competitive world. The fundamental concept here is that of a "certification authority," namely, a third party that verifies compliance with stipulated processes defined by the standard and that issues the certificate. fAIr LAC will work with AI ecosystem actors in the development of quality standards for public and private initiative projects.

- ***Official Regulatory Framework.***
  Possible differences may exist between the regulatory frameworks of various countries in the region. This document describes long-term mitigation strategies as those official modifications introduced by a government to established norms, regulations, and laws that are legally binding, meaning they are compulsory. When a good practice or standard becomes a norm or law, it becomes enforceable, and the consequences of its non-compliance should be clear. Work will be carried out to develop regulatory experimentation initiatives (regulatory sandboxes) and other mechanisms that report on the development of local regulatory frameworks (Figure 7).

QUALITY AND RISK MITIGATION AXIS MAIN LINES OF ACTION (figure 7)

	Strategi Dimensions for fAIr LAC	Lines of Action				Users or Beneficiaries
fAIr LAC Model	Quality and Risk Mitigation	Self-regulation, Guidelines, or Directives	Standards and Certifications	Initiatives on Regulatory	Regulatory Frameworks	Government
		Reinforce the participation of Latin American and Caribbean experts and practitioners in international efforts (IEEE, OECD, etc.) and bring international initiatives to the region				Private sector

Source: Prepared by the authors.

03.4 Regional and Territorial Strategy

This strategy must be anchored in key Latin American and Caribbean territories and actors through local fAIr LAC hubs that promote the responsible use of AI to ensure that it will yield a social impact, apart from having sectoral effects. This territorial anchorage is based on the adaptation to local realities of elements such as the relationship between public and private actors, capacity-building, data curation and enrichment, incentives to the demand of AI-based solutions, and the promotion of use cases (from identification to implementation). At the same time, efforts should be made to harmonize initiatives in the region as a whole in order to prevent legal fragmentation from undermining the establishment of a robust AI ecosystem.

An AI hub is an enabling ecosystem with the desirable conditions for the development and implementation of the fAIr LAC initiative. The strategic dimensions will be emulated at the local level with the aim of encouraging implementation of responsible AI and establishment of alliances between public and private institutions to promote LAC as an AI innovation pole for social impact (Figure 8).

Feedback and learning between regional and territorial initiatives will be a two-way process because, **given the execution of experiments and use cases, the development of AI will generate lessons that may be used for the production of knowledge for the whole region.**

As of December 2019, fAIr LAC had three regional hubs: Jalisco (Mexico), Uruguay, and Costa Rica. These hubs will serve as territorial benchmarks for the responsible use of AI for North America, Central America, and the Southern Cone, respectively.

REGIONAL AND TERRITORIAL MODEL (figure 8)

Source: Prepared by the authors.

Conclusion

The importance of artificial intelligence (AI) and other related technological developments for the future of mankind is unquestionable. These advances are radically transforming the way we work and live, and their effects – many of them positive and others less so – are the subject of discussion across many elements of society in all of the countries in the world.

As discussed throughout this document, one of the areas where AI is expected to have a substantial impact is on the **social welfare of citizens, by helping to improve the efficient and effective delivery of social services mainly related** to education, health, and poverty alleviation in order to reduce inequality.

Since these subjects have held a privileged position on the public agenda of Latin America and the Caribbean, the activities that the IDB and its public and private partners, academia, and civil society organizations will develop through the fAIr LAC initiative will be fundamental to helping address the challenges that AI represents for the region.

Even though the challenges identified are numerous and comprise matters of a technical, regulatory, data management, and public policy nature, among others, most efforts will have to concentrate on guaranteeing responsible and ethical use of AI and on preventing the deepening of social inequalities by neutralizing the harmful effects of biased automated learning against disadvantaged groups.

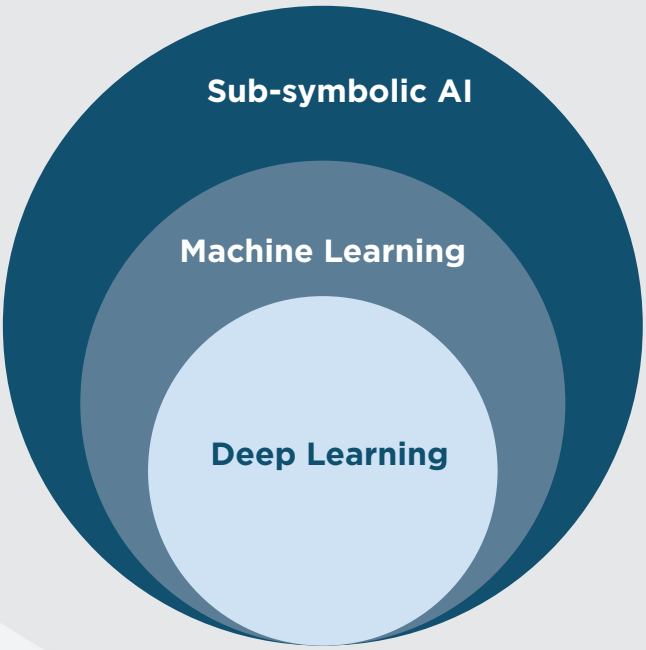
Most efforts will have to concentrate on guaranteeing responsible and ethical use of AI and on preventing the deepening of social inequalities.

04. Anex I

AI Sub-symbolic Learning Paradigms

Sub-symbolic AI encompasses different methodologies, among which is machine learning, which is the most used methodology at present (Figure A2.1).

SUB-SYMBOLIC AI (figure A2.1)



Source: OECD (2019)

Machine learning utilizes three learning paradigms: supervised, non-supervised, and reinforcement.

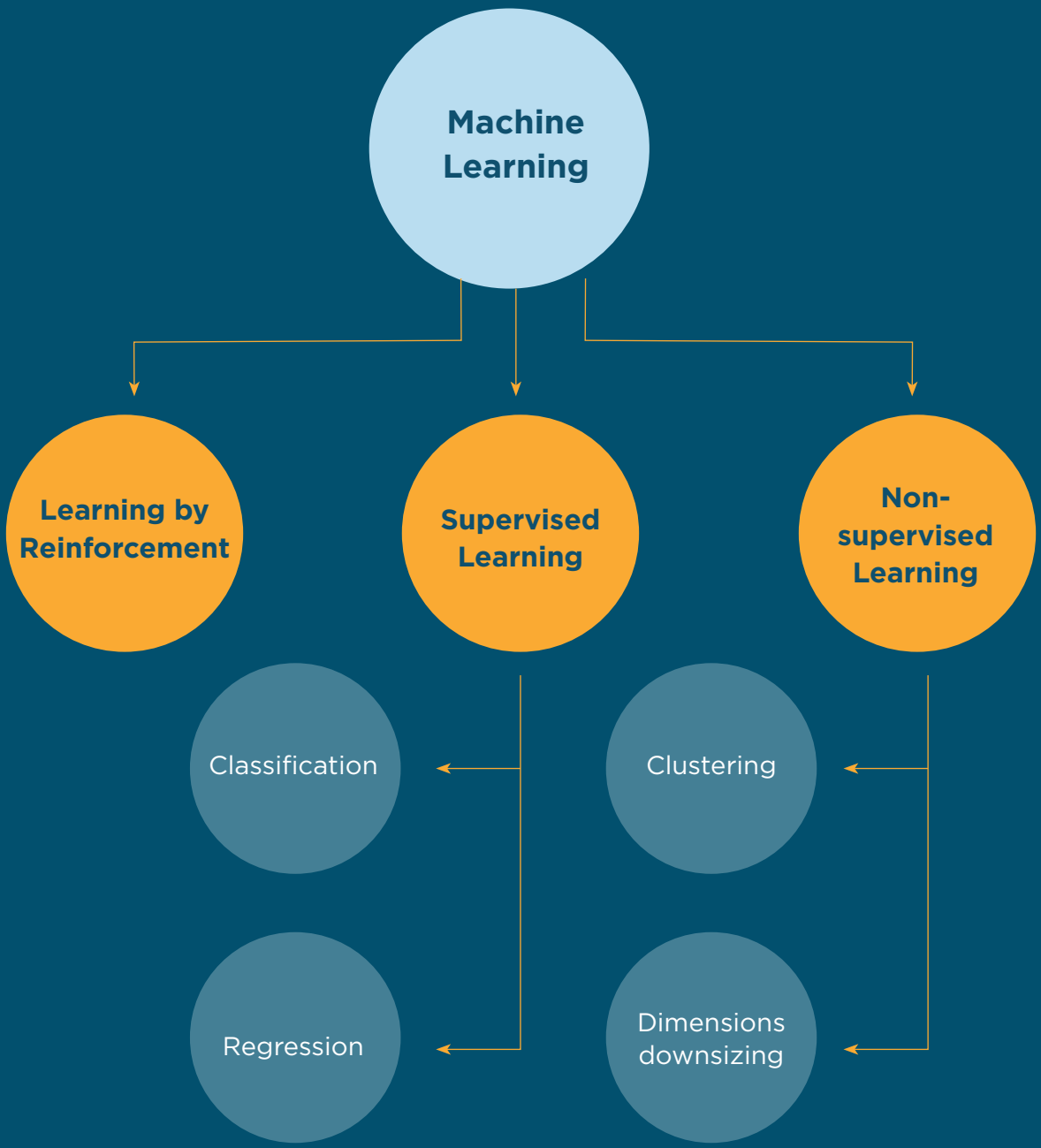
- **Supervised learning** occurs when information is used when the desired output or "label" is known in advance. The algorithm takes the variables related to the problem and learns the relationship patterns between them and its result. The objective of this type of learning is for the model to generalize and make a forecast or classification with a certain level of accuracy about the information that was not observed during the training. Some of the algorithms used in this paradigm are neuronal nets, vectoral support machines, and logistic regressions, among others (Bishop 2006).

- **Non-supervised learning** occurs when the algorithm does not observe a desired result or label during the training. Then the algorithm has to find structural patterns in the information that allow it to create associations, either by discovering similar groups within the training information (clustering) or by reducing the number of dimensions. Some of the algorithms utilized in this paradigm are hierarchical clustering methods, k-means, and tree-based models, among others (Bishop 2006).

- In the case of **learning by reinforcement**, no optimal solution examples are given to the algorithm and it must discover them through a trial-and-error process. This is done by creating a learning mechanism guided by a reinforcement system. The algorithm becomes an agent that must interact with an environment in which each action it takes carries a reward or a penalty. Seeking to maximize its gains, the agent learns guided by this system without interference and/or prior knowledge. Inside the learning by reinforcement paradigm, the agent´s decisions are known as policies, namely mathematical rules with which the agent decides "to exploit" [28] or to explore. Two of those rules are QLearning and the deterministic policy gradient (Bishop 2006).

28 Exploitation occurs when an agent repeats an action that is known, whereas exploration occurs when an agent experiments with new actions.

MACHINE LEARNING PARADIGMS (figure A2.2)



Source: Prepared by the authors.

05. Anex II

THE fAIr LAC MODEL (figure A1.1)

fAIr LAC Model

Strategic Dimensions	Lines of Action				Users or Beneficiaries
Diverse Network Highlight, Diffuse, Build, and Connect	**AI Experts Consulting Group** (a network of professionals and experts from academia, government, civil society, industry, and the entrepreneurial sector)		**Mechanism of Alliances and Institutional Alignment** (multilateral org., existing networks)		Government
	Outreach and Communication (reports, opinion pieces and analytical articles, debates)		**Observatory of Use Cases** (regional public asset to identify successful cases and systematize good practices and lessons learned)		Private Sector / Citizenship
Responsible Adoption (AI for social services)	**Support Program for Experiments and Use Cases** (consulting, financing, mentoring network)		**Pilot Projects** (with scaling potential in the region)		Government
	Incentives Model acknowledgments, access to finance and resources, institutional endorsements)		**Diverse Tools:** -Methodologies, guides, impact evaluation frameworks - Open Model Repositories and Tools Index		Private Sector
Quality and Risk Mitigation	Self-regulation, Guidelines, or Directives	Standards and Certifications	Initiatives on Regulatory	Regulatory Framework	Government
	Reinforce the participation of Latin American and Caribbean experts and practitioners in international efforts (IEEE, OECD, etc.) and bring international initiatives to the region				Private Sector

Well-being and Quality of Life (at the personal level)
Trust + Social Impact + Equality

Source: Prepared by the authors.

06. References

- Ardila, D., A. Kiraly, S. Bharadwaj, B. Choi, J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. Naidich and S. Shetty. 2019. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography". Nature Medicine 25 (1). doi: 10.1038/s41591-019-0447-x.
- Baraniuk, C. (2016). Millions of Mexican Voter Records '*Were Accessible Online*'. Abril. Obtenida de BBC: https://www.bbc.com/news/technology-36128745.
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Nueva York: Springer-Verlag.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.
- Buolamwini, J. y T. Gebru. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Conference on Fairness, Accountability, and Transparency.
- CEPAL (Comisión Económica para América Latina y el Caribe). (2019). *Panorama Social de América Latina 2018.* Santiago: Naciones Unidas. Obtenido de LC/PUB.2019/3-P.
- Cristianini, N. (2014). On the Current Paradigm in Artificial Intelligence. *AI Communications 27*, No. 1. Obtenido de https://doi. org/10.3233/AIC-130582.
- Dastin, J. (2018). *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women.* Obtenido de Reuters: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
- Dietvorst, B., J. Simmons y C. Massey. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General,* 1, 114–126. doi: 10.1037/xge0000033.
- Dignum, V. (2019). *Responsible Artificial Intelligence*. Springer. ISBN 978-3 -030-30371-6.
- Epstein, R., G. Roberts y G. Beber (2008). *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer.* Springer. p. 65. ISBN 978-1-4020-6710-5.
- Gajane, P. y M. Pechenizkiy. (2018). On Formalizing Fairness in Prediction with Machine Learning. Cornell University. Obtenido de https://arxiv.org/pdf/1710.03184.pdf
- Gasser, U. y V.A.F. Almeida. (2017). A Layered Model for *AI Governance. IEEE Internet Computing*, 21 (6) (Noviembre): 58–62. doi:10.1109/mic.2017.4180835.
- General Electric Healthcare y University of California-SF. (2016). Big Data, Analytics & Artificial Intelligence. Obtenido de http://newsroom.gehealthcare.com/wp-content/uploads/2016/12/GE-Healthcare-White-Paper_FINAL.pdf
- Guo, X., Y. Yilong, C. Dong, G. Yang y G. Zhou. (2008). On the Class Imbalance Problem. Fourth International Conference on Natural Computation.
- Hao, K.. 2019. *We Analyzed 16,625 Papers to Figure Out Where AI is Headed Next*. 25 de enero. Obtenido de https://www.technologyreview.com/s/612768/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/
- Jobin, A., M. Ienca y E.Vayena. (2019). The Global Landscape of AI Ethics Guidelines. *Nat Mach Intell* 1, 389–399. doi:10.1038/s42256-019-0088-2.

- cCarthy, J., M. L. Minsky, N. Rochester y C.E. Shannon. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Obtenido de http://jmc.stanford.edu/articles/dartmouth/dartmouth.pd
- McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. A K Peters/CRC Press; 2a edición
- McKinsey Global Institute. (2018a). *Notes from the AI Frontier: Applying AI for Social Good*. McKinsey & Company.
- ----------. (2018b). *Notes from the AI Frontier: Modeling the Impact of AI on the World Economy.* McKinsey & Company.
- Minsky, M. y S. Papert. (1969). *Perceptrons: An Introduction to Computational Geometry. Cambridge*: MIT Press
- Mittelstadt, B. (2019) Principles Alone Cannot Guarantee Ethical AI. Obtenido de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3391293.
- Morozov, E. (2014). T*o Save Everything, Click Here: The Folly of Technological Solutionism.* PUBLICAFFAIRS. ISBN 9781610391382
- OECD (Organizatión for Economic Co-operation and Development). (2019). *Artificial Intelligence in Society*. París: OECD Publishing.
- Ovanessoff, A. y E. Plastino. (2017). *How Artificial Intelligence Can Drive South America's Growth. Accenture.*
- Pombo, C., G. Ortega, F. Olmedo, M. Solalinde y A. Cubo. (2019) El ABC de la interoperabilidad de los servicios sociales: Marco conceptual y metodológico. Obtenido de https://publications.iadb.org/es/el-abc-de-la-interoperabilidad-de-los-servicios-sociales-marco-conceptual-y-metodologico.
- PwC (PricewaterhouseCoopers). (2017). *Sizing the Prize: What's the Real Value of AI for your Business and How Can you Capitalise?* Obtenido de https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf
- Rosenblatt, F. (1958) The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6), 386–408. https://doi.org/10.1037/h0042519
- Rumelhart, D., G. Hinton y R. Williams. (1986). Learning Representations by Back-propagating Errors. *Nature* 323: 533–536. doi:10.1038/323533a0
- Searle, J.R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*. 3 (3): 417-457.
- Snow, T. (2019). *Decision-making in the Age if the Algorithm: Three Key Principles to Help Public Sector Organisations Make the Most of AI Tools. Nesta*, Londres: Nesta.
- Szegedy, C., Z. Wojciech, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow y R. Fergus. (2014). Intriguing Properties of Neuronal Networks. Cornell University. arXiv:1312.6199v4 [cs.CV]
- Turing, A. (1950). *Computing Machinery and Intelligence*. Obtenido de https://www.csee.umbc.edu/courses/471/papers/turing.pdf
- Wood, J. (2018). This AI Outperformed 20 Corporate Lawyers at Legal Work. Obtenido de https://www.weforum.org/agenda/2018/11/this-ai-outperformed-20-corporate-lawyers-at-legal-work/