Recommendations for reliable artificial intelligence

1. Preliminary considerations

The irruption of Artificial Intelligence (AI), which is expressed in the growing importance of data and algorithms in people's lives, pushes States to define strategies to channel the transformative potential of this technology in solving specific problems and for the common good.

AI-based technological solutions allow higher levels of automation and the leap towards decentralized and predictive systems for decision making. On the productive level, AI is promising due to its ability to promote innovation, add value, increase labor productivity, give rise to new goods and services, boost exports, among other possibilities.

In the public sphere, AI offers solutions that make State management more efficient and improve the design and implementation of policies and the provision of essential services in health, education, security, transportation, environmental care, etc. Governments can also use AI to improve communication and engagement with citizens.

In this sense, the State plays a fundamental role not only by promoting research and development of AI solutions that are designed to meet the real needs of people, but also by guaranteeing that AI is transparent, equitable and responsible. This implies establishing clear rules to guarantee that the benefits of any technological development can be used by all sectors of society; to promote responsible collection and use of personal data, avoid algorithmic discrimination, and manage the risks of using AI to prevent harm.

Argentina has a dynamic scientific and technological ecosystem, with proven capacities for innovation, development and production of technological solutions based on AI. It is key to generate the political and institutional conditions so that these capacities are put in value at the service of a broader strategy that prioritizes technological sovereignty and allows a response to the social, productive and environmental problems of the country.

1.1 Objective

Through this document we seek to compile and provide tools for those who carry out public innovation projects through technology,

but specifically in those that import the use of artificial intelligence. In this sense, it is recommended to adopt a multidisciplinary approach, comprehensively conceiving the implications of the use, adoption, development and public innovation through artificial intelligence.

The manual is intended to provide a framework for the technological adoption of artificial intelligence centered on the citizen and their rights, conceiving its social and strategic aspect, ensuring optimal functioning of the provision of services and an ethical approach.

1.2 Scope

This manual seeks to offer theoretical and practical tools to those who are part of the public sector, whether leading innovation projects, developing technologies, adopting technologies developed by other technical teams/suppliers, formulating the technical specifications for these acquisitions.

2. Marco conceptual

Artificial intelligence currently groups together a set of technologies and is named after an ability that for a long time was considered exclusive to people: intelligence.

At the time this set of technologies was baptized with that name, the concept of intelligence was quite different from the ideas and theories that are currently being discussed about what we understand today as human intelligence.

Thus, in the middle of the 20th century, the study of intelligence was focused exclusively on cognitive abilities and, within them, on logical, mathematical, and linguistic abilities. At that time, the mind-computer analogy was also beginning to become popular, with which professionals in the area of cognitive psychology made computational metaphors to explain the advances, theories, and discoveries of the human mind, as well as professionals in computer sciences used that analogy with the human mind as inspiration to define the architecture of the first computers.

Current studies on human intelligence broadened this specific concept of intelligence and reformulated its understanding, expanding to different areas that were not previously considered as belonging to human intelligence. Howard Gardner, when developing multiple intelligences, exposes musical, bodily-kinesthetic, visual-spatial, intrapersonal, interpersonal, and natural intelligence as additional to logical-mathematical and verbal-linguistic. These theories represent an open framework that, as studies progress, are defined and

adding new types of intelligence, such as emotional intelligence. Multiple intelligences contribute to understanding the scope of current artificial intelligence technologies, since the scope of some of these multiple intelligences can be associated with the destinations of use or types of artificial intelligence technology.

Unlike humans, who possess all these intelligences to a greater or lesser degree, some more developed than others, artificial intelligence, until very recently, could only cover one type of these intelligences at a time. For example, there are currently artificial intelligences used only for natural language processing, they only perform that task, they are, let's say, *"good writers."* But, for example, not all of them can *"hear" us,* only *"read what we write" (after* our writing is encoded in binary, the language in which artificial intelligence processes information). Although none of them can yet speak to us, look at us, infer what we think, and simultaneously empathize with us, as well as other similar actions that are (so far) typically human.

This narrowness is a characteristic for which the vast majority of artificial intelligences that we use today are called *“narrow artificial intelligence” (also* called weak). In other words, these technologies are considered *"smart"* in a very specific aspect, taking into account the broad spectrum of human cognition.

There is also conceptually general artificial intelligence (also called strong), which would be equivalent to human intelligence, but so far it is a theoretical approach. However, technological evolution is advancing very quickly, and today there are already types of multimodal artificial intelligences. Multimodality allows the addition of two or more intelligences that work with a single type of data. For example, one that works with text and another with images, and make them work together expanding the reach of narrow artificial intelligences.

Said artificial intelligences, which in addition to receiving text-type inputs receive images, they in turn can contain text, and can also recognize it and take it as input. While multimodality seems to be a step in the right direction on the path towards artificial general intelligence, any predictions that can be made in this regard today are purely speculative. To contextualize the latter, we can conceptually list different similarities and differences between artificial intelligence systems and humans.

The framework that describes artificial intelligence systems developed by the Organization for Economic Cooperation and Development (OECD) shows that artificial intelligence models interact with the context by receiving different types of data (generated by people, sensors, cured by experts, public, private, dynamic, static, etc.), which are used to build the model of

artificial intelligence, which, once trained, processes this type of data to provide different output responses (recognition, event detection, prognosis, and other actions) with different destinations of use that can be human language, artificial vision, automation, robotic optimization, etc. All these actions affect the context, that is, they can alter the environment in which we live.

Conceptually, an analogy can be drawn with this way of describing artificial intelligence systems with the way in which human beings interact with their contexts. We perceive information from the environment with our senses, we mentally represent that information and there we can process it by performing different mental operations, then we can act in different ways and as a consequence of these mental operations we can speak, write, recognize people, create, etc.

At this level of description, the similarities are general and are being contrasted with the *"black boxes"* of artificial intelligence1 that , that is, with those models little are transparent and incapable of explaining their results. However, if compared by looking inside these boxes, there are substantial differences between machines and humans that mean that the road to artificial general intelligence does not occur, at least in the short term. One of the main differences is consciousness. Antonio Damasio, when addressing the issue of consciousness, describes three stages. One is consciousness, which allows us to have the ability to perceive what happens inside our body, which is different from the ability to perceive what happens outside. and the environment, the second stage, on these two the third stage of consciousness is built, called autobiographical, which allows us to remember the past and project or imagine the future.

Self-determination is the ability of a person to decide something for themselves, and this is also a human capacity, which allows them to act freely and choose actions with intention and purpose, while understanding the consequences of such actions and the responsibility that we have on them.
It serves to build our *self-concept.* That is, understanding the image we have of ourselves, for example, with the skills and competencies we possess to perform certain tasks effectively; At the same time, it also pushes to cover the need to integrate belonging groups in which we participate by affinity with other people. These human characteristics refer to the basic psychological needs for autonomy, competence, and affinity, defined in Deci and Ryan's Self-Determination Theory.

[1] These are machine learning algorithms or deep neural networks, among others, that do not reveal how they process information or make decisions. That is, models whose internal functioning is unknown or not transparent to external observers. Faced with black boxes, outside observers can only enter input data and receive output results, without having a clear understanding of the intermediate steps or the factors that influence the decisions made. Although black boxes can be highly effective in solving complex problems and achieving accurate results, they pose challenges in terms of explainability and ethics.

All these aspects, together with the accumulated experience provided by knowledge of the world, our body sensitive to the environment and the emotions that modulate thoughts, make up the subjective human experience which (at least for now) artificial intelligences do not enjoy.

However, through statistics, mathematics, large volumes of data, computer infrastructure and different interfaces that can provide a certain degree of action in the environment in which people live, these technologies are a reflection of our own humanity, a partial reflection but reflection at the end, built with its own virtues and defects. These concepts were addressed by the philosopher Shannon Vallor through the theory of the mirror, establishing that this reflection must be observed and optimized, not only through the development and evolution of technologies, but fundamentally aiming to improve ourselves as people.

3. Recommendations and principles for the implementation of AI projects.

The development and implementation of AI can, however, generate challenges, which demand that its adoption be projected following a series of ethical principles in order to maintain the protection of fundamental rights, respect democratic values, prevent or reduce risks, promote innovation and people-centered design.

To explain in an orderly way how these principles play, they will be framed in a time line that contemplates the life cycle of artificial intelligence.

3.1 How is it advisable to conceive artificial intelligence?

The starting point, prior to the cycle, refers to the conception of artificial intelligence. That is, how it should be conceived, how it is understood before working with it. This point is relevant given the human tendency to anthropomorphize technology. In this sense, a recommendable aspect is given by **clearly differentiating the concepts of responsibility and execution.**

When technological services are contracted, what is transferred to the provider is the execution of different tasks but not the responsibility for their effective completion. The same thing happens with artificial intelligence. When you use artificial intelligence algorithms, as before, you are shifting the execution, but not the responsibility. In other words, artificial intelligence only carries out an execution without its own intention and in a reactive manner to a human request, who has decided to program it, train it and implement it with a specific destination of use in order to execute different actions.

Consequently, it appears that **an algorithm does not have self-determination and/or agency to freely make decisions** (although the concept of *"decision"* is often used in colloquial language to describe a classification executed by an algorithm after training), and **therefore Therefore, responsibility for the actions that are executed through said algorithm in question cannot be attributed to it.**

In other words, for a human person to be legally responsible for the decisions they make to carry out one or more actions, there must be discernment (full human mental faculties), intention (human drive or desire) and freedom (to act in a calculated and premeditated). Therefore, to avoid falling into anthropomorphisms that could hinder eventual regulations and/or mistaken attributions, it is important to establish the conception of artificial intelligence as artifice, that is, as technology, a thing, an artificial means to achieve human objectives but that they should not be confused with a human person. **That is, the algorithm can execute, but the decision must necessarily fall on the person and therefore also the responsibility.**

International background.

From the very moment of conception, it is also relevant to address certain principles that all the actors involved must comply with, which should be taken as principles of design, development, implementation and use of artificial intelligence. In this sense, the United Nations Organization (UN) through the United Nations Educational, Scientific and Cultural Organization (UNESCO) issued the Recommendation on the Ethics of Artificial Intelligence, to which all countries adhered. members at the General Assembly in November 2021, including Argentina. This recommendation contains a set of principles that are summarized below.

Proportionality and safety. It should be recognized that AI technologies do not necessarily guarantee, by themselves, the prosperity of humans or of the environment and ecosystems. In the event that any harm to humans may occur, the application of risk assessment procedures and the adoption of measures to prevent such harm from occurring should be ensured.

Protection and security. Unwanted damage (security risks) and vulnerabilities to attack (protection risks) should be avoided and should be taken into account, prevented and eliminated throughout the life cycle of AI systems to ensure safety and security. protection of human beings, the environment and ecosystems.

Equity and non-discrimination. AI actors should promote the

diversity and inclusion, guarantee social justice, safeguard equity and fight against all types of discrimination, in accordance with international law. AI players should do everything reasonably possible to minimize and avoid reinforcing or perpetuating discriminatory or biased applications and results throughout the life cycle of AI systems, in order to ensure the fairness of such systems.

Sustainability. Continuous assessment of the human, social, cultural, economic and environmental effects of AI technologies should be carried out with full knowledge of the impact of such technologies on sustainability.

Right to privacy and data protection. It is important that data for AI systems is collected, used, shared, archived and deleted in a manner consistent with international law and in accordance with these stated values and principles, while respecting relevant national, regional and international legal frameworks. .

Human supervision and decision. Humans may sometimes decide to rely on AI systems for efficiency reasons, but the decision to relinquish control in limited contexts will still rest with humans as they can call on AI systems in decision-making and task execution, but an AI system can never replace the ultimate responsibility of humans and their accountability.

Transparency and explainability. The transparency and explainability of AI systems are often fundamental preconditions for ensuring respect, protection and promotion of human rights, fundamental freedoms and ethical principles. People should have the opportunity to request explanations and information from the AI manager or from the relevant public sector institutions. Such controllers should inform users when a product or service is provided directly or with the help of AI systems in a proper and timely manner.

Responsibility and accountability. Adequate monitoring, impact assessment, audit and due diligence mechanisms, including with regard to the protection of whistleblowers, should be developed to ensure accountability for AI systems and their impact over the long term. of its life cycle.

Sensitization and education. Public awareness and understanding of AI technologies and the value of data should be promoted through open and accessible education, civic engagement,

digital skills and training in the ethics of using AI, media and information literacy and training led jointly by governments, intergovernmental organizations, civil society, universities, the media, community leaders and the private sector, and taking into account the existing linguistic, social and cultural diversity, in order to guarantee effective public participation.

Adaptive multi-stakeholder governance and collaboration.
The involvement of different stakeholders throughout the life cycle of AI systems is necessary to ensure inclusive approaches to AI governance. These include governments, intergovernmental organizations, the technical community, civil society, researchers and academia, the media, education officials, policy makers, private sector companies, human rights institutions and equality promotion bodies, anti-discrimination watchdog bodies and youth and children's groups, among others.

Although the aforementioned Recommendation is currently the one with the greatest adherence by governments, other meetings have been developed and gestated with different actors in the AI ecosystem for the purpose of addressing and agreeing on common principles. Thus, in January 2017, the **Asilomar Conference** organized by the *"Future of Life"* Institute was held with the aim of making visible the vision of academia and industry on the opportunities and threats created by AI. In this framework, the participants made various contributions based on which a list of 23 principles on how AI should be managed was compiled, based on three axes: (i) research problems, (ii) ethics and values and (iii)) long-term problems.

Among the principles listed in the first category, Principle 4 stands out, related to fostering a culture of cooperation, trust and transparency among AI researchers and developers, and Principle 5, aimed at promoting the cooperation of the teams that develop AI systems to avoid taking shortcuts in security standards.

Regarding committed ethics and values, the following principles are highlighted:

6) Security: AI systems must be safe and secure throughout their operational life, and in a verifiable manner where applicable and feasible.

7) Failure transparency: If an AI system causes damage, it should be possible to determine why.

9) Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with the responsibility and opportunity to shape those

implications.

10) Values Alignment: Highly autonomous AI systems should be designed in such a way that their goals and behaviors can be ensured to align with human values throughout their operation.

Regarding the difficulties that AI can represent in the long term, in the Conference referenced those linked to:

11) Human values: AI systems must be designed and operated in a way that is compatible with the ideals of human dignity, rights, freedoms and cultural diversity.

12) Personal Privacy: People should have the right to access, manage and control the data they generate, given the power of artificial intelligence systems to analyze and use that data.

13) Freedom and privacy: the application of AI to personal data must not unreasonably restrict the real or perceived freedom of individuals.

14) Shared benefit: AI technologies should benefit and empower as many people as possible.

15) Shared Prosperity: The economic prosperity created by AI must be widely shared to benefit all of humanity.

16) Human Control: Humans must choose how and whether to delegate decisions to AI systems to achieve human-chosen goals.

In May 2019, the 36 OECD member countries, together with Argentina, Brazil, Colombia, Costa Rica, Peru and Romania, signed the OECD Principles on AI, within the framework of the Organization's Council of Ministers Meeting. , with the aim of guiding governments, organizations and individuals so that in the design and management of AI systems, the interests of people are prioritized, and responsibility for their proper functioning is guaranteed.

Based on the guidelines provided by governments, academic institutions, representatives of the private sector, civil society, international organizations, the technological community and trade unions, the following five principles were agreed upon, based on values for the responsible development of AI systems. :

Inclusive growth, sustainable development and well-being.
Stakeholders must proactively engage in responsible management of trustworthy AI that drives results

beneficial to people and the planet, such as increasing human capabilities and enhancing creativity, promoting the inclusion of underrepresented populations, reducing economic, social, gender and all kinds of inequalities, and protecting of natural environments, thus reinforcing inclusive growth, sustainable development and well-being.

Values and fairness AI **focused on the** **human being.**
actors must respect the rule of law, human rights and democratic values throughout the life cycle of the AI system. These values include freedom, dignity and autonomy, privacy and data protection, non-discrimination and internationally recognized equality, diversity, equity, social justice and labor rights.

To do this, AI actors must apply mechanisms and safeguards, such as human determination, that are appropriate to the context and consistent with the state of the art.

Transparency explainability. and AI actors must commit to transparency and responsible disclosure in relation to AI systems. To this end, they must provide meaningful information, appropriate to the context and consistent with the state of the art:

i) to foster a general understanding of AI systems, ii) to make stakeholders aware of their interactions with AI systems, workplace, iii) to enable those affected by an artificial intelligence system to understand the outcome , and; iv) to allow those adversely affected by an AI system to challenge its result based on clear and easy to understand information about the factors, and the logic that served as the basis for the prediction, recommendation or decision.

even in the

and

Robustness, protection. security
AI systems must be robust, secure, and protected throughout their entire life cycle so that, under conditions of normal use, foreseeable use or misuse, or other adverse conditions, they will function properly and do not pose unreasonable risks. of security
To this end, AI actors need to ensure traceability, including in relation to data sets, processes, and decisions made during the AI system lifecycle, to enable analysis of AI system results. and the answers to the questions, appropriate to the context and consistent with the state of the art.

AI actors must, depending on their roles, context and their ability

policy, apply a systematic risk management approach to each phase of the AI system lifecycle on an ongoing basis to address related risks, including privacy, digital security, security, and bias.

Responsibility.
AI actors must be responsible for the proper functioning of AI systems and respect for the above principles, depending on their roles, the context and in line with the state of the art.

Likewise, within the framework of the G20 Ministerial Meeting on Trade and the Digital Economy that took place in June 2019, and under the premise that innovative digital technologies continue to provide immense opportunities for the economic and productive development of nations While creating challenges, the G20 adopted the **Human-Centered AI Principles,** which were fundamentally replicated in the above-mentioned OECD AI Principles.

Having exposed the principles that it is recommended to incorporate into all phases of the design and implementation of the AI project and some of the international instances in which the subject is being addressed, the main points to be addressed in each stage and the information with which should be counted

3.2 What is recommended to do before starting the AI cycle?

Before thinking about the design, development, implementation and/or use of artificial intelligence, it is advisable to work on some relevant issues that must be defined before addressing the resolution of problems with this type of technology.

3.2.1 Form a diverse and multidisciplinary human team

The diversity of knowledge and perspectives in these teams is essential to address ethical challenges, understand the social implications, prioritize user-centered solutions, avoid bias and discrimination, and foster innovation.

Having a human team with diverse perspectives, knowledge and varied experiences in different areas, can help to achieve a deeper understanding of users and their contexts, and therefore, to address the challenges of AI from different points of view. view. It can lead to more complete and creative solutions, more intuitive and adapted to the real needs of people. Diversity in teams can also help identify and address biases

inherent in data, algorithms and automated decisions, helping to mitigate discrimination and ensure that AI systems are designed and/or implemented in a responsible, fair and equitable manner.

In addition, it is always advisable to create communication channels with actors outside the government, who can be consulted and listened to even if they are not directly involved in the development, implementation or execution of the project.
For example, actors from civil society, from universities, academia, business, specialists in ethics and the disciplines involved, among others.

3.2.2 What is the level of awareness in the organization?

It is important to start by understanding the level of knowledge of the people who make up the organization on the subject of artificial intelligence. But not only regarding technical knowledge of the subject, but also about different ethical aspects related to the adoption model to be used, destination of use and human control, risk management, as well as the best practices for public innovation.

Consequently, it is advisable to start by raising awareness through different means, such as communication campaigns, talks and training, describing adherence to the principles, indicating training actions and relocation of jobs depending on the type of adoption model. to be used in each case, as well as the existence of humans as agents of control and/or interpretation of results and decision-making. These aspects, and other similar ones that are disseminated in the same sense, will contribute to lowering the resistance to adoption of this type of technology, increasing the chances of success, sustainability and innovation.

3.2.3 Is the use of artificial intelligence exclusive for the problem to be solved?

Since the use of artificial intelligence carries various risks, and it is advisable not to start troubleshooting with the sole purpose of *“using artificial intelligence”,* **it is important that exploration of different types of technology be done before concluding that artificial intelligence offers the best solution to the problem to be solved. On certain occasions, other simpler, less risky and equally efficient software solutions can be used to respond to the same challenge.**

3.2.4 What is the scope of the basic models for AI adoption?

You can basically define two types of models on which you can choose to adopt artificial intelligence. One of them is automation. Basically

It consists of substituting human work for hardware, software and/or algorithms to carry out certain tasks, operations or repetitive, sequential processes, of different degrees of complexity, but which respond to duly typified problems. In this type of model, it is possible to reduce the processing times of different orders and respond to them automatically, as long as they are typified and maintain the same type of data (which is what artificial intelligence will work with).

It is important **to define that this adoption model,** like any other, is not simply based on the incorporation, configuration and start-up of a new technology, it must also consider other organizational aspects, among which is a fundamental and decisive one: training. and relocation of people whose tasks will be automated, these aspects must be addressed in a planning prior to the start of process automation.

Likewise, regardless of the degree of automation achieved, **it is always essential to have human intervention to verify and control the correct execution of automated processes;** to offer a personal route to the demand of those people who do not have the technological means to make one or more requests automatically; to evaluate if there is degradation of the algorithm; and to observe new types of requests that are not covered by the automation.

The other adoption model implies human participation far beyond control, it is a model where machines and humans collaborate to solve problems, it is known as the human-machine model, "augmented intelligence", or also " *human* in *the loop".* All of them are expressions of conceptually similar technological solutions that take advantage of the unique capabilities and characteristics of humans at various points in the AI lifecycle. In this model, artificial intelligence technologies contribute an important part of the work that is very expensive for humans, such as statistical thinking based on large volumes of data, but the results of these analyzes are presented to humans who contribute the work. difficult for machines, and they complement the analysis performed by the machine by making decisions, or by re-requesting new analysis and reports in order to make those decisions in a better informed way.

Understanding the differences of each adoption model and the actions that must be taken in each case is key to understanding the possible risks that exist in the destination of use of this technology.

3.2.5 What is the intended use of the AI, and how will human control be performed?

The versatility of artificial intelligence technologies allow their implementation in a very wide variety of destinations of use. But each of the intended uses carries different levels of risk, which in turn implies various levels of risk treatment and human control (for example, auditability and traceability control). In some areas this is very important given that certain destinations of use, for example in cybersecurity, may represent potential risks if their use is not controllable, auditable and traceable. These risks have a negative impact on transparency and consequent accountability.

Again, technology executes different actions that are ordered by a human person with intention, discernment and freedom of action. This is why, in the event that these intentions provided by one or more humans are not aligned with the common good and human rights, there must be control instruments designed to identify responsibility and accountability.

3.2.6 What is a premortem analysis?

An interesting way to identify eventual risks in an artificial intelligence project is to use the premortem technique. As in the prospective, the idea is to imagine a future, but instead of imagining the desired future, one imagines one where after implementing that project, the results were different from those expected. That is, imagine that the project was a failure. Once people are in that unwanted future, they analyze why or where the project failed.

The entire diverse and multidisciplinary team that will design the artificial intelligence project participates in this analysis, and each person will try to identify the reasons why the project failed. Then, each participant communicates their findings and together they classify the causes according to their probability of occurrence and their impact.
They select those with the highest probability and/or the most negative impact, identify them with a name, and then manage the risk that each one represents. Risk treatment can choose to accept, mitigate, eliminate or transfer these identified risks.

This technique allows, in a simple and fast way, to identify potential causes of failure, project risks and how to deal with them during the design phase.

3.2.7 What now...?

After answering these questions and working on the recommendations before

mentioned, but before starting the AI life cycle, it is interesting to identify the actors that will participate in the adoption of this technology and understand the contribution that each of them makes within the public innovation process2 .

3.3 Ethical aspects to consider within the AI cycle

Given that ethical aspects are specific to people, throughout each of the stages of the AI life cycle it must be ensured that the people who make up the diverse and multidisciplinary team in charge of design (which is the first activity of the life cycle), know and understand the necessary basic ethical aspects involved.

Stage N°1: Design and data modeling

This is the first stage of the AI life cycle. It begins with the design of the data and the models involved. It is important that from this first stage ethical aspects are included as design criteria that will facilitate compliance with the defined principles and consequently increase the chances of success of the project.

3.3.1. Common starting point for the diverse and multidisciplinary team

Since each of the members of a diverse and multidisciplinary team has varied knowledge with different experiences, it is advisable to clearly agree on the purpose of the project. Consequently, each and every one of the members must know, understand, agree, and commit to carrying out the following minimum aspects:

to. The principles of design, development, implementation and ethical use of artificial intelligence, defined by UNESCO. b.
The impact on society in general and the needs to be covered in the particular recipients.
c. The potential risks evaluated by level of impact and probability of occurrence, and the treatments defined for each of them. d. The
transparency and accountability mechanisms to be used for traceability and auditing (either of what is executed by machines and/or what is decided by people). and. The role,
scope of activities and distribution of responsibilities of each member of the team. F. The definition and formal assignment of the person
responsible for ensuring

[2] To facilitate identification and understanding, the use of the **Guide for the design and technological adoption of public innovation projects is recommended,** which facilitates the identification and understanding of the participation and interdependence of the actors involved.

the sustainability of the project over time.

g. The survey and understanding of the various profiles of recipients, whether all citizens or part of it (taxpayers, public employees, social security beneficiaries, students, patients, etc.). This includes aspects that could eventually give rise to different biases. It is also recommended that each of these profiles be represented by at least one person. h. The survey and understanding of the scope, implications and impact

of the regulations involved.

Yo. The documentation, registration and socialization of the experience to promote good practices and lessons learned necessary for organizational learning and public innovation.

3.3.2 Ethical considerations regarding data design

The treatment that should be given to the data involved in the project should not be underestimated. These should be dealt with by professionals based on good data science practices. The quality of the data used will determine not only the quality of the trained model but also contribute to the success of the project. The data is the raw material to build the trained artificial intelligence model that will be used so that, by entering different inputs, a correct answer is obtained.

In this sense, the different aspects detailed below must be considered in order to carry out an ethical data design.

to. The classification of the data according to its confidentiality. It is recommended that there be an agreement regarding said classification and that it be elaborated on the basis of international standards related to information security. As an example, a general classification agreed at an international level is outlined.

Yo. Confidential data. It refers to sensitive data or information that may refer to intelligence, defense, security, and other similar issues.

ii. Personal information. It refers to those data or information of people that have been specifically defined as such by current regulations.

iii. Internal data. It refers to those internal management data or information, which are neither confidential nor personal, but which are not classified as public information. iv.

Public data. Refers to those data or information in the public domain that are generally available, either as open data sets, and/or as content in

websites.

b. The data sources that will be used to design and build the data set corresponding to the training of the model.

Yo. Data available on the internet. It is the least expensive case, however, it must be taken into account that there is a high degree of probability that they are inaccurate, have different types of biases, may be subject to intellectual property, among various aspects that not only degrade the quality of the data, but also prevent the creation of training data in an ethical manner. ii. Existing data in

the organization. In this case, it is necessary to measure the associated costs: prior to its use, the classification must be taken into account according to its confidentiality, rights of use, consent of the owners, possibility of anonymizing said data and other aspects established by current regulations.

iii. Data requested from third parties. In this case, the associated costs must also be dimensioned, since they are not available on the internet, and to obtain and use them, the classification must be taken into account according to their confidentiality, rights of use, consent of the owner, the traceability of the entire process. obtaining and creating data for training and other aspects established by current regulations. c. The quality of the data. In all cases,

the quality of the data must be ensured. For example, avoiding the existence of biases, verifying that they are accurate or that they reflect the reality they claim to represent, among others. This treatment must be carried out by data science professionals, who must continuously evaluate the different data sets in order to ensure that the training of the AI models is carried out according to the aforementioned UNESCO principles.

3.3.3 Ethical considerations regarding model design

The models must be designed in such a way that they do not introduce biases inherent to their conception. For example, through a definition that omits aspects of the context that privilege or disadvantage some people over others, or using algorithms that work better with certain variables or characteristics than with others, which could generate eventual inaccuracies or biases in the results.

In line with UNESCO principles, models must be transparent and explainable. That is, the execution that led to its result must be able to be understood by the people who operate these systems, so that they, in turn,

can make decisions with those results, and also, to be able to clearly explain to the people affected by the decision made or to third parties how said result was reached.

Stage N°2: Verification/Validation

In a second stage, within the AI life cycle, it is important to carry out the corresponding verifications and validations of the designs carried out in the first stage. For this, the design of the equipment, the data and the models involved must be taken into account. These verifications and validations are carried out taking into account both the principles defined by UNESCO, as well as the interaction of the recipients with the designed prototypes (first conceptual solutions of the trained model or models), in conditions similar to those that will have their final implementation. .

3.3.4 How is the specific ethical knowledge needed for the AI project validated?

Likewise, after the training (or awareness) carried out, so that the members can know, understand, agree, and commit to carrying out the minimum necessary ethical aspects, they can be dumped in writing and signed in an ethical commitment act of the AI project.

3.3.5 Data ethics

Data sets built specifically to train artificial intelligence models must be validated prior to implementation in the field. People on the team who are data science professionals should be in charge of evaluating the quality of data that will be used to train the AI models.

A risk classification should be established (for example, of three traffic light-type levels, or with values from one to five) regarding how much they conform to the aforementioned UNESCO principles.

3.3.6 How are aspects related to the ethical design of AI models validated?

Prototype tests must be carried out by professionals with knowledge of agile methodologies and it is recommended that the diverse and multidisciplinary team in charge of the project be present during the tests. In this instance, the models trained in conditions similar to those that they will have in their implementation will also be validated. To carry out these tests, one or more prototypes of the trained models will be used, with a minimal user interface but similar in appearance to the final one.

To test the models, at least one person from each defined profile will be invited so that the work team can observe how the model is used and take advantage of this interaction to verify different ethical aspects of the design. For example, that there are no biases, that the person in charge of making the decision can understand the result of the execution of the model (in the case of the human-machine adoption model), that it can be easily explained to the people affected, validating that they clearly and fully understand the result of the model and the consequent human decision.

That is, in this test with prototypes, different ethical aspects will be validated such as; the congruence between the results and the expectations of the design; the absence of bias; the explainability of the model, as well as other ethical aspects of the design that are susceptible to improvements. A risk classification should be established (for example, of three traffic light-type levels, or with values from one to five) regarding how much they conform to the aforementioned UNESCO principles.

3.3.7 How are verifications/validations recorded?

All actions and decisions taken within an AI project, including those related to verifications and validations of ethical aspects carried out in the design stage, must be recorded. This point is critical to be able to comply with the principles related to transparency and accountability corresponding to the actions and decisions involved in each AI project. A formal registration medium must be used that allows traceability and audits of each and every one of the verification and validation actions.

Stage N°3: Implementation

At this stage, the infrastructure professionals who are part of the diverse and multidisciplinary team of the IA project come into play. In this case, there are implementation options that can be based on contracting cloud services, on the deployment of your own infrastructure, or on a solution that includes both options.

Whatever the case, it must be ensured that the implementation allows: establishing an adequate degree of information security; carry out traceability on the actions and decisions that occurred in the project, identifying the people who carried them out; carry out audits (this point is especially important when contracting cloud services) and offer the user facilities for accessibility to information and communication technologies (ICT).

3.3.8 How to establish an adequate degree of information security?

It is important that information security best practices are followed. For this, those responsible for the security of the information that

form the diverse and multidisciplinary work team, they must take into account the following aspects:

to. The survey, knowledge and understanding of the scope of international standards and regulations, and best practices in information security. b. The survey, knowledge and understanding of current regulations in information security.

c. The use of accessory applications in charge of managing the records (logins, events, etc.) of the systems involved in such a way as to facilitate the treatment of eventual security incidents; automate the creation of audit reports; and improve transparency through control of the people who access the systems, applications and equipment. d. Carrying out different tests to find security vulnerabilities that could cause eventual unwanted incidents. and. The participation of the area or authority with primary responsibility for information security, which understands all aspects related to cybersecurity and the protection of critical information infrastructures, as well as the generation of prevention capacities, detection, defense, response and recovery from computer security incidents. This is particularly important in the event that the institution adopting the development based on AI does not have a specific area of information security.

3.3.9 What aspects must be taken into account when establishing traceability?

The systems involved in the deployment of infrastructure for the implementation of the AI project, as well as the procedures defined for their management, must have the appropriate means of recording all the actions carried out in the system (registrations, cancellations, modifications in configuration) for all hierarchies and all user profiles (Administrator, operator, users, etc.), in such a way as to reliably identify all the people who carried out the different actions and decisions in the project.

In the case of deployment through cloud services, whether total or partial, it is necessary to understand, prior to contracting, the traceability offered by cloud service providers in order to understand if the scope of traceability offered allows implementing the ethical principles corresponding to said matter.

3.3.10 What aspects must be taken into account for the systems to be auditable?

To guarantee compliance with ethical principles, it is necessary to audit the model and traceability is the best tool to achieve this objective. It is key to be able to identify and understand the record of actions, decisions and/or any other event that affects the systems involved in the AI project.

In the case of deploying on-premise solutions (within the organization's infrastructure), it is important to ensure, in addition to controlling access to systems, physical access control where the infrastructure involved is housed.

In the case of deployment through cloud services, it is important to understand, prior to contracting, the audit facilities offered by cloud service providers, in order to understand if the scope offered allows the implementation of the ethical principles corresponding to said matter.

3.3.11 What aspects must be taken into account so that the systems have ICT accessibility?

It is necessary to carry out the best practices of ICT accessibility, whether through web pages or mobile applications. To this end, the professionals in charge of ICT accessibility, who make up the diverse and multidisciplinary work team, must take into account the following aspects:

to. The relay, knowledge and understanding of the scope of international regulations and best practices in terms of ICT accessibility. b. The relay, knowledge and understanding of the scope of the regulations national in terms of ICT accessibility.

c. The evaluation of the website. In the particular case of web accessibility, it is recommended to use specific applications available in charge of evaluating the accessibility of websites that users will use to access the systems involved in such a way as to ensure a minimum level of accessibility (level A). Likewise, it is recommended to request the assistance of the application authority of Law 26,653 of Web Accessibility.

Stage No. 4: Operation and maintenance

Technological innovation projects do not end with implementation; Operation and maintenance is the final stage of the AI life cycle. A common problem is that these two actions, despite their importance, are often not considered in project design. These tasks are the operations and maintenance of both the infrastructure where the AI-based technological solution is deployed, as well as the model itself, given that, for

For example, many times the models degrade and stop responding correctly. These actions allow for availability, continuity, and sustainability of the service provided through the AI solution.

3.3.12 How could you monitor and what should be monitored considering the ethical use of AI?

Monitoring is an action taken at this stage to make sure everything is working as expected. Different variables can be monitored that will be chosen according to the purpose pursued. For example, if what is sought is to understand if the model responds as it was validated in the tests with prototypes, its performance can be monitored through the measurement of different parameters automatically, and manually, that is, carried out carried out by people who inspect and perform assessments of the behavior of the model.

Monitoring manual assessments allows the people involved to understand the outputs generated by the model based on the inputs provided by the users. Therefore, it not only enables verification of the model's performance in terms of the quality of the response provided, but also with respect to possible biases that may have been omitted or overlooked in the design and testing process. Likewise, other types of undesirable results can be detected that, if not monitored, could have different degrees of negative or detrimental impact on people.

With this type of assessment it is also possible to understand the degree of applicability that the operator can offer and the level of transparency that it can provide to the end user.

3.3.13 What general aspects should be considered regarding the existence of ethical incidents?

Ethical incidents can be caused by different reasons. For example, they can be caused by an involuntary human error in one of the stages of the life cycle that causes a malfunction in one or more technologies involved, an intentional and improper use of one or more people within the organization, an improper use end users, an attack on the security of the organization (internal and/or external), among others.

If the principles and recommendations included in this document were received, the minimum bases are in place to be able to provide correct treatment to a possible ethical incident, whatever its cause.

Since the occurrence of incidents cannot be eliminated, the correct and complete documentation of them will be a fundamental input to be able to take account of the details and conditions in which they occurred. Subsequently, these records

They will be useful to prepare the necessary accountability reports and comply with the principles defined by UNESCO.

The treatment of incidents allows learning from them to avoid their repetition, highlighting those aspects that failed in order to correct them.

3.3.14 What precautions from the ethical point of view should be considered for the control of internal users?

Like any other computer system, the minimum necessary controls for authentication and authorization of internal users must be carried out regardless of the role they have (administrator, operator, user, etc.), the existence of generic users such as "maintenance", " monitoring", etc., since they do not allow the person who uses them to be identified.

Internal users who have not been part of the diverse and multidisciplinary team involved in the design of the project must have the same treatment as said members. That is, each and every one of the internal users must clearly understand the purpose of the project and formally register their ethical commitment, whether in the administration, operation or simple use of the technologies involved within the AI project.

All changes in the configurations, replacements, updates, improvements, or any action carried out in the technologies involved within the AI project must be planned, registered and formally authorized by the person responsible for the project (and/or the impact of the services it provides). are provided through AI technologies) who in turn will be accountable to the authorities, the ethics committee (if it exists) and various control and audit bodies.

None of the changes in the configurations, replacements, updates, improvements, or any action carried out in the technologies involved within the AI project must be carried out individually, privately, unilaterally, discretionally, and/or without being formally registered.

3.4 What ethical aspects should be considered outside the AI cycle?

The chronological order established in this document established the different ethical aspects to be considered at different times. The time of conception of the AI, the time before the start of the AI cycle, and the time when the AI cycle takes place. Now it is time to work on the ethical aspects in the moment after the cycle.

Of course, the operation and maintenance of the systems are still in force to ensure their availability and sustainability, although at this time some

issues may change: perhaps the diverse and multidisciplinary team has dissolved; the priorities have changed, those responsible, the approaches have changed, etc. However, as long as the service (or services) provided through AI technologies remain in force, it is necessary to carry out the operation and maintenance tasks, although the responsibilities are not limited to them.

Although in the design stage different risks were surveyed and different treatments were planned to avoid their occurrence or negative impact, this does not prevent their occurrence and potential damages that may be derived from them.

The persons formally designated as responsible must act immediately and personally to understand the scope of the damage and arbitrate by possible means the necessary actions to correct and/or reverse the damage caused. The necessary means to be able to carry out said correction must be previously defined as formal and institutionalized procedures. The actions of responsibility and accountability involved must be duly recorded, and be defined as witness cases to be studied and discussed as part of the lessons learned, necessary for organizational learning and continuous process improvement: both aspects that favor the public innovation.

Without exception, for all cases in which there are services provided through AI technologies, a human route must be established (with face-to-face attention) to attend to those people who, due to their profile and/or contextual situation, do not have access to the devices. and minimum universal basic technological services necessary to be able to be users of said services, or prefer the attention of a human person.

4. Glossary

Technological adoption: It is a necessary requirement for innovation that occurs both when organizations are end users of one or more technologies and/or when they are the ones who implement one or a set of technologies that are contracted through a technology adoption project and these They turn out to be an effective means to provide services and/or implement public policies since they are adopted by the recipients.

Algorithms: OECD documents define them as exact sequential sets of commands that are executed on input designed to produce output in a clearly defined format. Algorithms can be represented in plain language, diagrams, computer code, and other languages.

Machine learning: The United Nations defines it as a branch of artificial intelligence (AI) focused on creating applications that

they learn from data and improve their accuracy over time without being programmed to do so. OECD documents define it as a subset of artificial intelligence in which machines take advantage of statistical approaches to learn from historical data and make predictions in new situations.

Deep learning: OECD documents refer to learning models inspired by biological neurons, however, neural networks do not necessarily learn in the same way as humans. Such networks organize computing through large collections of simple computational units. The term *"deep"* refers to the number of layers in the network. Until recently, a lack of computing power and training data meant that only small networks could be explored. Several decades of research into algorithm improvements, combined with graphical processing units originally developed for video games, finally made it possible to train large networks using massive amounts of data. This has led to systems that perform much better than previous approaches on tasks like image captioning, facial recognition, speech recognition, and natural language machine translation.

Automation (through AI): AI systems designed to automate typical, monotonous, massive and repetitive tasks. Automation represents a way of adopting AI that must be accompanied by a retraining process for people displaced by said automation for their relocation within the organization.

Iterative trial and error cycles: Organic development methodology that allows designers and developers to get real-time feedback on their work and make quick and effective adjustments. It also makes it easy to detect problems and errors early, so you can fix them before they become bigger problems. This methodology is widely used in the development of innovative products and services in various fields, including technology, design, engineering, among others. One of these cycles is represented by the iterative sequence of create-measure-learn.

Data science: Discipline that, through the combination of mathematical and statistical models, computer programming and data visualization techniques, supports decision-making processes, for example, to design public innovation projects, based on data processing. large volumes of data.

Value construction: Ability of the solution to provide a significant and objectively measurable benefit for the recipients. The service or public policy must generate value for the recipients, through, for example, its ability to meet the needs, difficulties and frustrations of these people by creating or improving their experiences as users of the technologies.

that were used as a means to implement said services and/or policies.

Creativity: From an individual perspective, it refers to the capacity or ability of the person to make contributions that are both new and valuable. It can also be understood as a process (composed of different stages), as products (which must have value and novelty), as contexts (which are cultivated to favor it). It also refers to practices or actions that a person performs taking advantage of their accumulated experience and knowledge to interact with their social and material context, thus allowing them to carry out said contributions that must be new and valuable in the contexts for which they were created.

Skewed data: Presence of imbalances or distortions in the training data, for example, used to develop an artificial intelligence model. They may be due to a variety of factors, such as a lack of diversity in the data, the inclusion of incorrect or incomplete data, the exclusion of certain categories of data, or the selection of data that reflects partial or limited reality.
For example, when a machine learning model is trained on data that does not fully represent the population to which it is applied, it can lead to incorrect or biased predictions in the real world.

Organic development: Refers to a process of creation and evolution of products and/or services that is based on iterative cycles of trial and error. The process implies continuous feedback and adaptation based on the results obtained in each cycle, which allows a natural and fluid evolution of the product or service. This approach is essential when working with agile methodologies, since it allows greater flexibility and adaptability in the process of designing and creating potentially innovative technological solutions.

Human-Centered Design: Exercise of design activities that focuses on the needs, desires, difficulties and frustrations of the people who will use the designed product or service (not on the technical or technological aspects).
It involves the exploration of the behavior of the target people and promotes an iteration from the beginning of the design and throughout it for feedback from those people. Prioritize people by identifying opportunities to improve their experiences, proposing solutions that are intuitive, useful, effective and easy to adopt.

Explicability: OECD documents define it as that aspect that allows people affected by the result of an AI system to understand how it was arrived at. This involves providing easy-to-understand information to people affected by the result of an AI system that allows them to question the particular result, to the extent possible, the factors and logic that led to a result.

Human-Centered AI (Augmented Intelligence): AI Systems

designed to amplify and augment human capabilities and human control over machines, not replace them. They are systems that prioritize the interests and rights of people over automation. It also represents a way of adopting AI where one or more technologies do not replace the people involved but rather they work collaboratively commonly known as human-machine modality.

Innovation: Action and effect produced by creating something new or altering/ modifying something existing, giving rise to something substantially different that adds value in a certain context, given that said novelty is adopted by the people who integrate it, improving or transforming some aspect of their to do.

Public Innovation: Processes, products or services, which deliver value, and turn out to be new or improved to respond to collective challenges and improve citizen satisfaction, increase the productivity of the state administration, the democratic opening of its institutions, the production of services and public policies, among others.

Artificial Intelligence: There is no universally accepted definition of AI. In November 2018, the OECD AI Expert Group (AIGO) established a subgroup to develop a description of an AI system. This group defines it as a system based on machines that is capable of influencing the environment producing a result (predictions, recommendations or decisions) for a determined set of objectives. It uses data and inputs based on machines and/or humans to (i) perceive real and/or virtual environments; (ii) abstract these insights into models through analysis in an automated way (for example, with machine learning), or manually; and (iii) use model inference to formulate options for the results. AI systems are designed to operate with different levels of autonomy. Likewise, the United Nations defines artificial intelligence as the ability of a computer or a computer-enabled robotic system to process information and produce results similar to the thought process of human beings in learning, decision making and resolution. from problems.

Narrow Artificial Intelligence: According to the OECD, *“narrow”, “weak”,* or *“applied”* artificial intelligence is designed to perform a specific reasoning or problem-solving task within a limited domain. While these tasks may be driven by highly complex algorithms and neural networks, they remain singular and goal-oriented. Narrow AI does not have the ability to adapt to new situations without prior reprogramming.

General Artificial Intelligence: Also known as *“strong”* artificial intelligence or “generalized” artificial intelligence , it is a constantly evolving area of research and development (OECD). It refers to artificial intelligence systems that would have the ability to learn, generalize, induce and abstract knowledge to

through different cognitive functions. They would have a strong associative memory and would be able to judge and make decisions. They could solve multifaceted problems, learn through reading or experience, create concepts, perceive the world and themselves, invent and be creative, react to the unexpected in complex environments, and anticipate. It only exists as a theoretical concept, its advent is uncertain.

Technological adoption project: Way in which the adoption of one or more technologies is organized and achieved through different actions with the purpose of carrying out some type of innovation, such as public innovation.

Artificial neural network: OECD documents define it as a sophisticated statistical modeling technique. This technique is accompanied by increasing computational power and the availability of massive data sets *(“big data”).* Neural networks involve the repeated interconnection of thousands or millions of simple transformations into a larger statistical machine that can learn sophisticated relationships between inputs and outputs. In other words, neural networks modify their own code to find and optimize links between inputs and outputs. Finally, deep learning is a phrase that refers to particularly large neural networks; there is no defined threshold as to when a neural network becomes *"deep".*

Bias: In OECD documents describing common-base key terms for G20 discussions, they define four types of biases that can occur in AI systems.

Perception bias: Occurs when the data collected over- or under-represents a certain population and makes the system work better (or worse) for that population compared to others.

Technical bias: Occurs when the technology itself introduces biases or inaccuracies due to, for example, algorithms that work better with certain variables or features of the AI system that are introduced with different variables or features.

Modeling bias: It occurs when the manual design of a model by experts does not take into account some aspects of the environment, either consciously or unconsciously.

Activation bias: Occurs when the outputs of the AI system are used in the environment in a biased way.

Argentine Republic - National Executive Branch
1983/2023 - 40 YEARS OF DEMOCRACY

Additional Signature Sheet
Attachment Disposition

Number:

Reference: Recommendations for trustworthy artificial intelligence

---

The document was imported by the GEDO system with a total of 29 pages/s.