



MEXICO | AUGUST 2023

# Open Loop Mexico:

Public Policy Prototype on the Transparency and Explainability of Artificial Intelligence Systems



CLAUDIA MAY DEL POZO  
NORBERTO DE ANDRADE  
DANIELA ROJAS ARROYO



## About Open Loop

Open Loop is a global program that connects policymakers and technology companies to help develop effective, evidence-based policies around AI and other emerging technologies. The program, supported by Meta (previously Facebook), builds on the collaboration and contributions of a consortium of regulators, governments, tech businesses, academics, and civil society representatives. Through experimental governance methods, Open Loop members co-create policy prototypes and test new and different approaches to laws and regulations before they are enacted, improving the quality of rule-making processes in the field of tech policy. This report presents the findings and recommendations of the Open Loop policy prototype creation program on the transparency and explainability of Artificial Intelligence systems that was conducted in Mexico between February and August 2021.

This report is licensed under a Creative Commons Attribution 4.0 International License.













## Quote this report

Del Pozo, C., Nuno Gomes de Andrade, N., & Rojas Arroyo, D. "Prototipo de Políticas Públicas sobre Transparencia y Explicabilidad de Sistemas de Inteligencia Artificial [Public Policy Prototype on the Transparency and Explainability of Artificial Intelligence Systems] (2023), at: <https://openloop.org/reports/2023/10/Public-Policy-Prototype-on-the-Transparencyaand-Explainability-of-Artificial-Intelligence-Systems.pdf>

## Acknowledgements

This policy prototype program was co-designed and led by Open Loop, a global experimental governance program supported by Meta, C Minds' Eon Resilience Lab, and the Inter-American Development Bank (IDB), via its fAIr LAC initiative, with support from Mexico's National Institute for Transparency, Access to Information and Personal Data Protection (INAI).

We would like to thank the following companies for their collaboration and commitment, in addition to their active participation, without which this report would not have been possible:

<b>Ai360</b>	 The logo for ai360 consists of the text 'ai360' in a bold, blue, sans-serif font. Below it, the words 'Analítica Inmobiliaria' are written in a smaller, black, sans-serif font.
<b>Fincomún</b>	 The logo for Fincomún features a stylized 'F' icon composed of three curved lines in pink, orange, and blue, followed by the word 'Fincomún' in a blue, sans-serif font.
<b>helKi</b>	 The logo for helKi is the word 'helKi' in a teal, lowercase, sans-serif font.
<b>Hitch</b>	 The logo for hitch features the word 'hitch' in a bold, blue, lowercase, sans-serif font. A small red square is positioned at the bottom right of the 'h'.
<b>Inndot</b>	 The logo for inndot consists of the word 'inndot' in a bold, black, lowercase, sans-serif font. To the right of the 't' is a yellow triangle pointing to the right. Below the text, the words 'PIENSA SOLUCIONES' are written in a smaller, black, uppercase, sans-serif font.
<b>LUZi</b>	 The logo for LUZi features a circular icon with a blue and pink yin-yang-like design. Below the icon, the word 'LUZi' is written in a blue, sans-serif font.
<b>Nauphilus</b>	 The logo for nauphilus features a shield-shaped icon with a blue and white design. To the right of the icon, the word 'nauphilus' is written in a blue, lowercase, sans-serif font.
<b>Nowports</b>	 The logo for nowports features a stylized 'N' icon composed of three curved lines in blue and pink, followed by the word 'nowports' in a blue, lowercase, sans-serif font.
<b>OS City</b>	 The logo for OSCITY features a hexagonal icon with a blue and pink design. To the right of the icon, the word 'OSCITY' is written in a pink, uppercase, sans-serif font.
<b>Rhisco</b>	 The logo for rhisco features the word 'rhisco' in a bold, black, lowercase, sans-serif font. Below it, the words 'TO COMPLY & COMPETE' are written in a smaller, black, uppercase, sans-serif font.



Special thanks to the following experts for their invaluable time, ideas, and contributions to the development and implementation of this public policy prototype and report (who are listed in alphabetical order):

- **Carla Vázquez Wallach**  
Founder and General Director of Legal + Innovation in Mexico and Member of C Minds' Brain Hive
- **César Said Rosales**  
Project Director for the Inter-American Development Bank (IDB) Lab's fAIr LAC initiative
- **Constanza Gómez-Mont**  
President and Founder of C Minds
- **Cristian Guerrero**  
Former Technical Consultant for C Minds
- **Cristina Pombo**  
Consultant for the IDB's Social Sector
- **Daniel Castaño**  
Founder and Partner of Mokzy, Professor at the Universidad Externado de Colombia, and Researcher and Consultant focused on AI, digital ethics, and regulation
- **Edson Prestes**  
Researcher at the Informatics Institute of the Universidad Federal de Rio Grande do Sul, Brazil, Senior Member of the Institute for Electrical and Electronics Engineers Robotics and Automation Society and the Association for Rules, and Member of C Minds' Brain Hive
- **Guillermo Larrea**  
Corporate Attorney focusing on Latin America at Jones Day
- **Jesús Sánchez**  
Deputy Director of Research for the National Institute for Transparency, Access to Public Information and Data Protection (INAI)
- **Jonathan Mendoza**  
Secretary of Personal Data Protection for the INAI
- **Laura Galindo Romero**  
AI Policy Manager at Meta
- **Natalia González**  
Former Coordinator of the IDB's fAIr LAC initiative
- **Paula Vargas**  
Director of Privacy Policy and Engagement at Meta
- **Rafael Ramírez de Alba**  
Professor at the Economic Environment Department of the IPADE Business School and Member of C Minds' Brain Hive
- **Ricardo Baeza-Yates**  
Research Director for the Experiential Artificial Intelligence Institute at Northeastern University in Silicon Valley and Member of C Minds' Brain Hive
- **Tetsuro Narita**  
Senior Specialist in the IDB Lab Investment Unit
- **Verena Kotschieder**  
Former AI Policy Program Manager at Meta
- **Victoria Martín del Campo**  
Former Tech Philosopher at C Minds

---

In addition, we would like to thank our communication partners Wizeline and MC Luhan, who were key to the call for proposals stage of the program.

<b>Forewords</b>	<b>7</b>
<b>Executive summary</b>	<b>11</b>
<b>1 Introduction</b>	<b>19</b>
<b>2 Open Loop Mexico and Public Policy Prototypes</b>	<b>22</b>
What is Open Loop? .....	23
What is a public policy prototype? .....	23
Why design public policy prototypes? .....	23
<b>3 Open Loop Mexico</b>	<b>24</b>
Responsible AI Context in Mexico .....	25
What are transparency and explainability in the context of AI systems? .....	26
Goals for the Open Loop Mexico program .....	28
Key players .....	29
Methodology .....	34
Prototype implementation .....	35
Evaluation criteria .....	40
Limitations of the exercise .....	40
<b>4 Assessment of the public policy prototype</b>	<b>41</b>
Framework clarity .....	42
Public policy effectiveness .....	42
Feasibility .....	43

<b>5 Specific amendments to the regulatory framework and playbook</b>	<b>44</b>
<b>6 Recommendations for the formulation of public policies focused on AI/ADM systems' transparency and explainability</b>	<b>48</b>
<b>7 Conclusion</b>	<b>53</b>
<b>Bibliography</b>	<b>55</b>
<b>Annexes</b>	<b>58</b>
<b>Endnotes</b>	<b>90</b>



## Forewords

### **National Institute for Transparency, Access to Information and Protection of Personal Data (INAI)**

In the digital era, we face imminent technological progress that never ceases. We are part of a generation that uses technology in their daily activities. In this context, it is of utmost importance to keep in mind the advantages and disadvantages of the virtual world, so it is essential to stop for a minute to reflect on the guarantees that we must demand from both developers and authorities and the role of the user in the use of the tools that make our lives easier.

From a privacy perspective, trust is critical and essential to keep up with the fast pace of innovation, so it is of utmost importance to consider ethical aspects of the design that demonstrate responsible use in processing personal data.

As mentioned by Yuval Noah Harari, Israeli Historian and Writer, "The first regulation I would suggest is to make it mandatory for AI to disclose that it is an AI. If I am having a conversation with someone, and I cannot tell whether it is a human or an AI—that's the end of democracy. This text has been generated by a human" .

Ethics must go hand in hand with innovation to generate regulatory mechanisms and public policy. It is crucial to implement a comprehensive framework for the ethical utilization of data throughout its entire life cycle, spanning from generation and utilization to elimination. This framework should guarantee that the treatment of data is based on ethical principles, instilling confidence in data owners. Therefore, prioritizing an ethical approach is essential to ensure that users remain at the core of decision-making processes.

As former European Data Protection Supervisor Giovanni Butarelli noted: "Human innovation has always been the result of the activities of specific social groups and specific contexts, generally reflecting the social norms of the time. However, technological design decisions should not dictate our social interactions and the structure of our communities, but should support our fundamental values and rights."

INAI's role in this project consisted in guiding the participating companies on the best practices to guarantee the protection of personal data, for which we suggested actions of privacy by design and by default, as well as the strict adherence to the principles for personal data processing.

INAI is pleased to participate in this type of initiatives that allow us to get closer and go hand in hand with developers and companies that implement new technologies and who are concerned, at the same time, with guaranteeing the respect of their users' human rights. Open Loop represented a unique opportunity for INAI, which allowed us to highlight the importance of personal data protection and the respect of data subjects' privacy. As regulatory authorities, we must encourage public-private collaborations that shall be a key instrument and an effective communication channel to face the future challenges presented by emerging technologies, allowing us to strengthen and implement various human-centred prevention mechanisms.

#### **Jonathan Mendoza Iserte**

Personal Data Protection Secretary

National Institute for Transparency, Access to Public Information and Data Protection (INAI)



## **C Minds**

Artificial intelligence (AI) systems play an increasingly prominent role in our daily lives, so it is imperative that we prioritize responsible and rights-centered use. In this sense, the C Minds team is committed to creating strategies that minimize potential risks and maximize the positive social impacts of AI systems and other emerging technologies. We believe that a collaborative approach is essential to achieving this goal, as bridging different sectors and perspectives can lead to a more comprehensive understanding of the nuances associated with bringing principles to practice. This, in turn, leads to more holistic, inclusive, and pragmatic strategies for the development and use of AI systems.

In 2019, C Minds coauthored Mexico's National AI Strategy, which focused on the responsible and ethical use of technology, positioning Mexico among the 10 countries to have created a strategy pertaining to AI. Since then, we have continued our mission to improve the quality of life in Mexico and other Latin American and Caribbean countries via the responsible use of new technologies. We have achieved this by developing pioneering AI ethics projects in the region, generating public policy recommendations, and creating guidelines and frameworks focused on responsible development and use.

The project presented in this report is exciting, especially considering its unique character in the region and its multisectoral governance model. The creation of public policy prototypes is a dynamic and pioneering methodology that has emerged as a promising approach to mitigating some of the challenges currently present in the development of public policy, especially those related to technology. It also underscores the importance of learning from feedback and involving stakeholders at an early stage. Among other things, this allows for the development of contextualized public policy and pragmatic solutions to challenges—in our case, those in the field of AI ethics. This project is based on a philosophy of inclusivity, bringing together all sectors to ensure complementary perspectives. It is the result of a proud collaboration between Meta and C Minds' Eon Resilience Lab, together with the Inter-American Development Bank (IDB) - fAIr LAC's initiative, with the support of Mexico's National Institute for Transparency, Access to Public Information and Data Protection (INAI), participating companies and subject matter experts.

With this document, we hope to contribute to the creation of regulatory frameworks for the development and use of human-centered AI systems in Mexico. We hope that the recommendations presented inspire regulatory institutions not only in Mexico but throughout the Latin American and Caribbean regions to continue promoting the responsible use and development of emerging technologies and to create more inclusive processes and more equitable technological benefits for all.

### **Constanza Gómez Mont**

Founder and President

C Minds





## Meta

Transparency and explainability are fundamental for the responsible development of Artificial Intelligence (AI). The OECD identifies transparency and explainability as one of its core principles, stating in its Recommendation on AI that "AI actors should commit to transparency and responsible disclosure of AI systems.... [providing] meaningful information [that is] appropriate to the context, and consistent with the state of the art." Indeed, at Meta these are central aspects of our five pillars of responsible AI.

Our interdisciplinary Responsible AI (RAI) team has collaborated closely with academia, civil society, governments, and other industry partners on various transparency and explainability projects. For instance, we have introduced AI system cards as a way to explain how the AI powering our products works in a manner that users can understand. In our first pilot, we explain the categorization process of the Instagram Feed, providing a simplified step-by-step explanation of the system's backend operations. Additionally, we offer an interactive exercise where users can experiment with hypothetical profiles to predict the appearance of their feeds. The intention is to provide users with the necessary technical information to understand our products and make informed decisions about their experiences.

Another example is our long-standing Why am I seeing this? (WAIST) tool, which underwent an update in early 2023. By personalizing users' experiences with our products, we use AI to present them with content and ads that are most relevant to their interests. The updated WAIST tool allows users to gain a better understanding of this personalization process for both News Feed and ads. The updates include summarized information about how user activity, both on and off our technologies, informs the machine learning models we use to shape and deliver ads. We have also included new examples and illustrations that explain how our machine learning models connect various topics to present relevant ads to users.

In addition to our internal efforts, external collaboration plays a vital role in our responsible AI endeavors.

Approaching policy development in an experimental, evidence-based manner enables policymakers and regulators to systematically assess the impacts of their proposals on people and businesses. This approach helps them gain a deeper understanding of how these proposals resonate in the real world before they become concrete laws and regulations. Our aim with Open Loop is to share its learnings with policymakers and stakeholders worldwide, encouraging them to adopt similar prototyping initiatives and embrace an innovative and collaborative approach to public policy development.

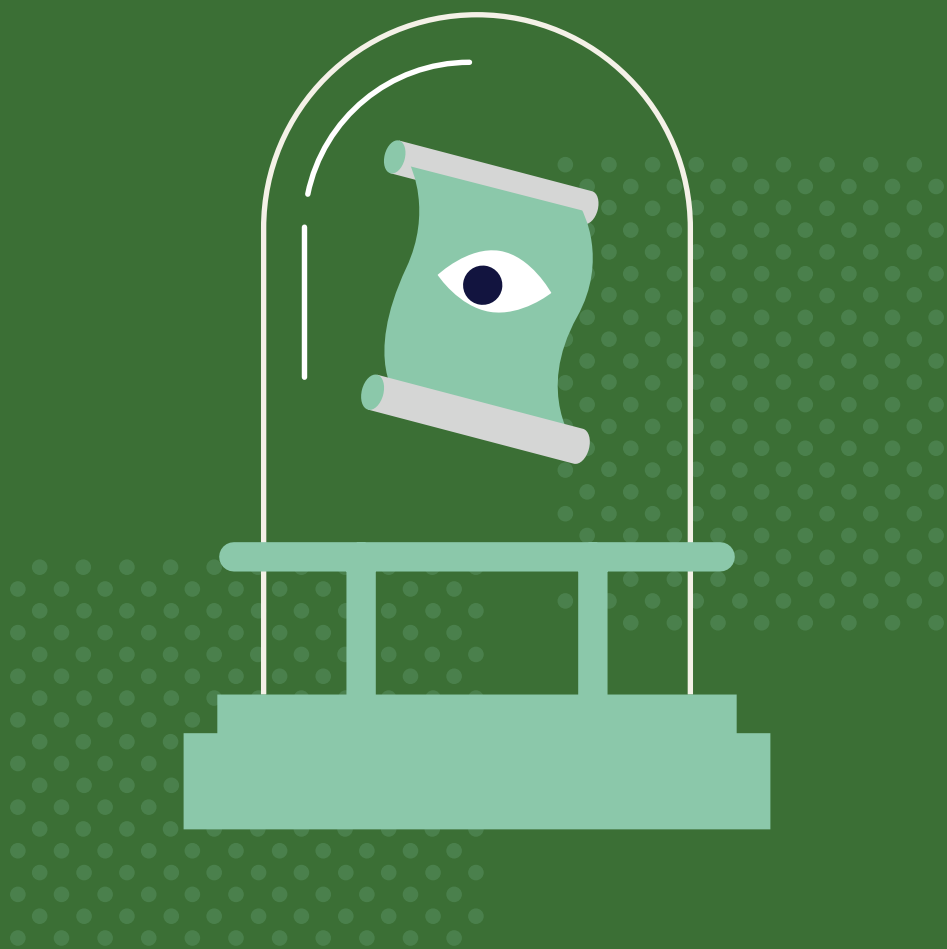


The Open Loop project described in this report revolves around the creation of a governance framework and a practical manual that embody the principles of transparency and explainability. Serving as a public policy prototype, these principles were subsequently tested by Mexican companies that use automated decision support systems to offer goods and/or services. The primary objective of this initiative was to explore how these companies could effectively incorporate transparency and explainability, empowering users to access essential information about their interactions with AI systems. By fostering a deeper understanding of the capabilities, limitations, and processes leading to specific outcomes, users would be better equipped to navigate their experiences. The design of the public policy prototype on transparency and explainability of automated decision support systems was guided by international best practices, taking into account the contextual and scope considerations. Open Loop Mexico primarily focused its efforts on nascent and early maturity companies, seeking to understand the circumstances faced by organizations with limited technical, economic, and human resources. This emphasis is crucial as such entities constitute a significant portion of the business landscape in developing economies like Mexico (representing 98% of the composition).

I would like to mention that this would not have been possible without the collaborative spirit of Mexico's National Institute for Transparency, Access to Information and Protection of Personal Data (INAI), the fAIr LAC initiative of the Inter-American Development Bank (IDB), and the tireless and professional colleagues of C Minds' Eon Resilience Lab. We are also grateful, of course, for the participation of the Mexican companies that joined the project, and for the support of an excellent group of experts.

**Paula Vargas**

Director, Privacy Policy & Engagement LATAM  
Meta



## **Executive summary**

## Executive summary

### What is Open Loop?

Open Loop is a global program that connects policymakers and technology companies to help develop effective and evidence-based policies around AI and other emerging technologies. The program, supported by Meta, builds on the collaboration and contributions of a consortium composed of regulators, governments, tech businesses, academics, and civil society representatives. Through experimental governance methods, Open Loop members co-create policy prototypes and test new and different approaches to laws and regulations before they are enacted, improving the quality of rulemaking processes in the field of tech policy.

### What is Open Loop Mexico?

In the case of Mexico, the **“Public Policy Prototype on the Transparency and Explainability of Artificial Intelligence Systems”** (hereinafter AI systems will be referred to as AI/ADM systems, so as to also refer to Automated Decision-Making (ADM) systems, maintaining tech neutrality in light of possible future technology developments)<sup>1</sup> was carried out by Meta and C Minds’ Eon Resilience Lab, in collaboration with the Inter-American Development Bank (IDB), through its fAIr LAC initiative, and with support from Mexico’s National Institute of Transparency, Access to Information and Personal Data Protection (INAI), together with the industry and thematic experts. The purpose of this program was to design a governance framework and a practical manual (playbook) that outlines the principles of transparency and explainability (T&E). These documents (policy prototype) were tested by Mexican companies that utilize AI/ADM systems to provide goods or services. The overall policy aim was to strengthen responsible AI in Mexico, focusing on T&E.

This exercise aimed to ensure that people know when they are interacting with an AI/ADM system and understand its limitations and capabilities, as well as how it achieves specific results.

#### Why focus on transparency and explainability?

Any human who interacts with an AI/ADM system should be able to know how and why certain results, conclusions, or predictions are being produced and, consequently, understand the logical reasoning behind a decision or recommendation given by this type of system.

The principle of transparency refers to people’s capability to understand and describe the internal functioning of a system. In turn, explainability safeguards the right to know the internal mechanics of an AI/ADM system and to understand it in human terms.

The public policy prototype on the T&E of AI/ADM systems (hereinafter, “prototype”) was designed based on the context and scope of international best practices around T&E principles. The idea was to test it for the following questions:

- **Clarity:** To what extent do the participating companies understand the requirements established in the prototype?
- **Effectiveness:** To what extent does the prototype help achieve the general policy objective?
- **Viability:** To what extent do the benefits outweigh the costs of achieving the objectives of the public policy prototype?

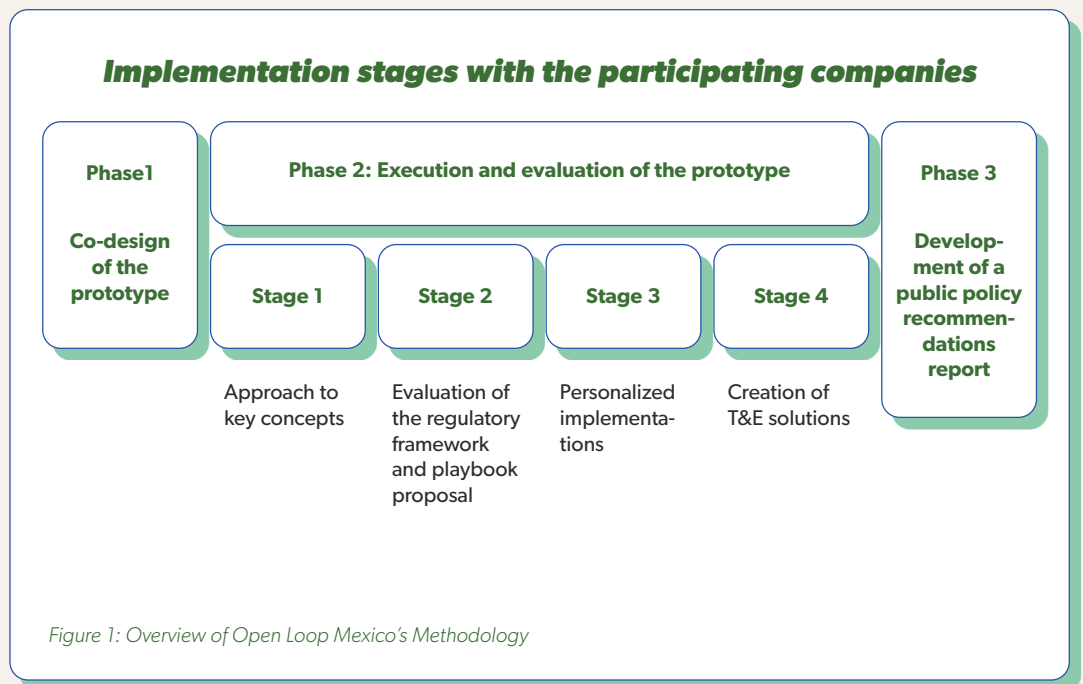
### How was Open Loop Mexico carried out?

The program was carried out with 10 Mexican companies that carried out a series of activities to implement the set of norms and practices contained in the normative framework proposal and accompanying implementation guide (playbook) created by the project partners with the support of topical experts. They then provided feedback on the clarity, viability, and effectiveness of these documents.

By executing a series of activities, the companies adjusted to and tested the scope of the prototype from a compliance perspective. The comments and suggestions provided by the companies allowed for the adjustment to the content of the prototype to improve it, in terms of clarity, viability, and effectiveness, and promote an understanding of what it means to observe T&E principles.

### Methodology

The Open Loop Mexico program was carried out between February and August 2021 and was structured into three phases:



### Results

Open Loop Mexico demonstrated that the proposed methodology to evaluate the suitability of the prototype was appropriate for testing the normative governance framework and the playbook with participating companies. The program allowed for the reception of feedback on some of the successes and challenges experienced in implementing a possible regulation on the T&E of AI/ADM systems under the evaluation of the three established criteria. We obtained the following results:

- **Clarity:** The prototype was clear to participating companies. In terms of T&E, companies' understanding of the topic went from 5.3 out of 10 prior to reading the documents and 8 out of 10 afterwards, with 0 being "no knowledge" and 10 "expertise". The objectives and specific actions required for compliance with the established requirements in the framework were also clear, especially with the support of the playbook. The latter significantly impacted the implementation effectiveness, allowing companies to ground their case in the scope of prescriptive norms in the proposal. Overall, they gave the documents a 7.4 out of 10 for clarity, with 10 being extremely clear.
- **Effectiveness:** Most companies designed and/or published explainability messages as part of the user experience with the products or services. Some chose to include notifications, messages or videos for the user to understand how the AI/ADM system worked. While the prototype demonstrated that it encompasses the essential elements for companies to develop T&E solutions for their products or services, there remains an opportunity for enhancing the understanding of the potential risks associated with AI systems. This highlights the need for comprehensive awareness campaigns regarding the impacts of AI/ADM systems before considering the implementation of a mandatory governance framework. Without adequate awareness, there is a risk of significant noncompliance with regulations.
- **Viability:** In general, companies expressed that they experienced some level of difficulty in complying with the governance framework established in the prototype due to a lack of time and sufficiently trained workforce in technical matters. They also noted that the feasibility of implementing T&E mechanisms would differ depending on the complexity and impact (risk level) of the model used by companies.

The Open Loop Mexico program gave rise to a governance framework that can serve as input for regulatory institutions to develop public policies on the T&E of AI/ADM systems, with the advantage of having been tested and enriched with recommendations from the companies that implemented it. The content of the framework and playbook was strengthened by the following actions:

- Clearer wording was integrated; individual definitions were established for transparency, explainability, and global and local interpretability; and hypothetical cases were added to improve the understandability of the prototype.
- A new stage was included that considered the importance and implementation of impact assessment processes for the determination of the risks associated with AI/ADM systems.

Although the information gathered and recommendations obtained cannot be generalized to all types of companies (and it would be important to carry out a similar complementary exercise with large and multinational companies), Open Loop Mexico primarily focused on nascent and early-stage companies to understand the situation of organizations with fewer technical, economic, and human resources, which represent the majority of the business composition in developing economies (Micro, Small and Medium Enterprises (MSMEs) represent 99.8% of companies in Mexico) (INEGI, 2019).

## Recommendations

Based on the results of the Open Loop Mexico program, including the information received from the participating companies and the group of experts, the following public policy recommendations on T&E for AI/ADM systems should be considered by policymakers:

- 1 Proactively promote Artificial Intelligence as a national priority, with a focus on operationalizing trustworthy AI principles**
- 2 Play a proactive role in AI governance in Mexico**
- 3 Build capacity for trustworthy AI in non-technical government bodies and particularly on transparency and explainability (T&E)**
- 4 Increase technical capacity for trustworthy AI in Mexico**
- 5 Invest in AI research and development for trustworthy AI**
- 6 Strengthen the capacity for responsible AI development and adoption in the Mexican workforce**
- 7 Expand civic awareness of AI in Mexico**

### 1

#### **Proactively promote Artificial Intelligence as a national priority, with a focus on operationalizing trustworthy AI principles**

- Policymakers could use existing resources and leading international practices and tools in order to create a National AI Strategy<sup>1</sup>.
- This strategy could outline the policy goals for AI in line with the OECD and UNESCO AI Principles, as well as the policies that could be needed to achieve those goals. The strategy could also include specific measures to promote transparency and explainability of AI systems<sup>2</sup>.
- This exercise should be a multi-party effort, led by national government bodies and the creation of the strategy could also include the private sector, academia, and civil society via innovative exercises.

## 2 Play a proactive role in AI governance in Mexico

- Policymakers in Mexico could take a proactive role in governing the development and use of AI in the country by: i) organizing and promoting experimental government exercises to identify and address the opportunities and challenges of AI such as public policy prototypes and regulatory sandboxes (before policies/regulations are set in place), as well as hackathons and competitions to further understand opportunities and challenges in this field; ii) developing a clear and concise normative framework for AI, based on local needs and international best practices and iii) promoting cross-sector collaborations to ensure that the AI framework is comprehensive and reflects the views of all stakeholders.

## 3 Build capacity for trustworthy AI in non-technical government bodies and particularly on transparency and explainability (T&E)

- Organize and implement capacity-building sessions and workshops on AI opportunities and risks, with a focus on T&E. Policymakers could work with civil society organizations and academia to organize and implement these, as well as massive online open courses (MOOCs) to level policymakers' and public official's knowledge of AI risks and opportunities, especially related to T&E. Improved capacity and knowledge would allow them to better participate in conversations on the topic.
- Create regular spaces for dialogue with government officials, AI developers, and other stakeholders to discuss issues related to T&E. For example, policymakers could create a task force of government officials, AI developers, and other stakeholders to discuss issues related to T&E. These dialogues could help to build consensus on best practices for AI design, development, deployment and use, and to identify areas where further guidance is needed.



## 4

### Increase technical capacity for trustworthy AI in Mexico

- Policymakers could consider developing a set of technical standards/protocols for AI systems in consultation with AI developers, businesses, and other stakeholders in the Mexican AI ecosystem, leaning on international good practices to ensure they include human-in-the-loop practices when relevant and are aligned with a human centered approach to AI.
- Explore the development of a risk management framework based on the Mexican context that is highly consistent and interoperable with international best practices and standardization efforts<sup>3</sup> for the design, development, and deployment of trustworthy AI systems and reduce their potential for unexpected negative impacts, especially relevant if companies choose to abide by T&E principles. This could be led by the regulatory institutions, in collaboration with the Mexican AI ecosystem.
- In addition to creating local resources, consider gathering existing international resources by countries, companies, and multilateral organizations on a government webpage that is regularly updated.

## 5

### Invest in AI research and development for trustworthy AI

- Policymakers could establish financial and non-financial incentives for promoting AI T&E research projects via governmental bodies and public and private universities, in collaboration with the industry and civil society to ensure a practical approach. Cross-border opportunities could also be considered. In particular, these actors could invest in research on:
  - i) techniques for making AI systems more transparent and explainable, this could include research on methods for visualizing the decision-making process of AI systems, as well as research on methods for explaining the rationale behind AI decisions; and
  - ii) research and tools for identifying and mitigating bias in AI systems, as well as overall risk management frameworks.
- In addition to fostering and funding research, the government, academia and other AI stakeholders could create spaces to share the key learnings, recommendations, and tools that result from the research activities.

6

**Strengthen the capacity for responsible AI development and adoption in the Mexican workforce**

- Promote the inclusion of courses and modules on ethical considerations in the development and adoption of AI systems in technical careers linked to data science, computer science, and artificial intelligence, among others. This could apply to formal education spaces like Universities and other learning institutions or courses, including life-long learning ones.
- Social science and humanity careers in formal and informal education spaces, including life-long learning opportunities, could also offer introductory courses and modules to AI systems, to create a more diverse workforce that can focus on responsible AI from different perspectives. This could be promoted by Certification Agencies and policymakers, through bodies like the National Institute of Transparency, Access to Information and Protection of Personal Data (INAI), in collaboration with industry, civil society, and academia, creating training programs on the importance of and how to build transparent and explainable AI systems, especially for developers.

7

**Expand civic awareness of AI in Mexico**

- Policymakers could launch a public awareness campaign about the risks and opportunities of AI systems, highlighting the importance of T&E in AI services and products. This campaign could help boost T&E practices as companies use it as a competitive advantage and consumers request them from their product and service providers
- Policymakers, and local youth and education agencies could further promote digital literacy education programs in schools and universities and as lifelong learning courses, with a focus on AI, once the digital basics have been understood, in collaboration with civil society and academia.
- Support the development and deployment of digital and AI literacy resources in Spanish, and work together with the government and AI actors to promote AI literacy.



# Introduction

AI has the potential to considerably transform society, improve social and individual well-being and the common good, and bring about progress and innovation. These systems provide several opportunities for companies in Latin America. According to the Everis and MIT Tech Review report on the use of AI in Mexico,<sup>3</sup> 47% of the companies in Mexico have an AI project, and 38% see benefits in its use but do not yet use it. This reflects a great interest in the use of these systems, which will likely result in growing development and adoption in the coming years.

This technological adoption, especially regarding AI systems, has created new challenges for the protection of people's rights and freedoms. This type of technology is increasingly integrated into our daily lives, as AI systems are making—or supporting—decisions that impact our lives. In consequence, there is a need to ensure the responsible development and use of AI models that protect users' rights and freedoms. To achieve this, the

***“AI systems must be human-centered, used to benefit humanity and the common good, with the purpose of improving people’s wellbeing and freedom.”<sup>4</sup>***

In line with this, the topic of AI governance seeks to foster an informed debate about the ethical, normative, and political implications arising from the development and use of AI. This is based on a diagnosis of this technology's challenges and opportunities, as well as paths toward these future developments. The responsible use of AI and data training in these systems has been at the heart of many AI governance debates around the world, which has resulted in the development of several proposals and guides based on ethical principles. These proposals have been created by institutions such as the Organization for Economic Cooperation and Development (OECD),<sup>5</sup> the European Parliament,<sup>6</sup> the Institute for Electrical and Electronics Engineers (IEEE)<sup>7</sup>, the United Nations Educational, Scientific and Cultural Organization (UNESCO)<sup>8</sup>, and the European Commission<sup>9</sup>, among others. The OECD's AI Principles, adopted in May 2019, in particular, offer five value-based principles:<sup>10</sup> 1) inclusive growth, sustainable development, and well-being; 2) values focused on human beings and equity; 3) transparency and explainability (T&E); 4) strength, safety, and protection; and 5) responsibility and accountability.

Based on these, several legal proposals and tools have been developed to operationalize the principles. These examples show that

ethical guidelines and AI regulations alike make T&E a core pillar. Indeed, it is a key element to generate confidence in users, guarantee the individual's right to understand a decision that impacts said individual, and promote accountability for all parties interested in developing AI systems. This principle focuses on promoting users' awareness of their interaction with AI systems, the capabilities and limitations of the system, and how particular results were reached.

Moving forward, it will be necessary to develop collaborative and dynamic learning mechanisms between regulators, companies, scholars, civil society, and the innovation ecosystem to facilitate the creation of AI governance frameworks focused on T&E that are pragmatic, inclusive, and operative.

To carry this out, and due to the complex nature of the task at hand, **public policy prototypes become particularly interesting as they** provide a secure testing ground to evaluate the suitability and the possible impacts of the public policy before it is implemented. With this in mind, a public policy prototype focusing on T&E was developed via Meta's Open Loop program and C Minds' Eon Resilience Lab, in collaboration with the Inter-American Development Bank (IDB), through its fAIr LAC initiative,



with support from Mexico's National Institute for Transparency, Access to Information and Personal Data Protection (INAI). This was the first public policy prototype in Latin America and the Caribbean.

The Open Loop program in Mexico was designed to translate findings into practical ideas, discuss the responsible use of technologies (specifically, T&E in AI/Automated Decision-Making (ADM) systems), and provide recommendations to Mexico's regulating institutions, which may also inspire other countries in the region. The following report gathers the key learnings and recommendations from this exercise.

2



**Open Loop Mexico  
and Public Policy Prototypes**

Public policy is commonly defined as the concrete actions that a government performs in the interest of the public and that arise as a result of a problem diagnosis process. Public policy may become a law or a regulation that concerns a particular concrete problem.<sup>11</sup> Given that the traditional public policy design processes, which typically involve only the government, tend to lag behind technological innovation, regulatory innovation has led to public policy prototypes, which represent a multi-party effort to create adaptable, human-focused, inclusive, and sustainable policies.<sup>12</sup>

## What is Open Loop?

Open Loop is a global experimental governance program supported by Meta. It is where regulatory and technology innovation meet through the development of public policies based on the evidence around emerging technologies, with a special emphasis on AI. Its main purpose is to generate the necessary information to create governance frameworks with better technological and public policy interaction comprehension—based on cooperation between regulators, governments, technology companies, scholars, and civil society—and implemented with experimental governance methods to cocreate public policy prototypes related to the technology to improve their development and implementation processes. The Open Loop program has been deployed worldwide several times, each time, with a different sub-theme related to the responsible design, development and use of AI. In November 2022, seven programs were conducted or were in progress ([www.openloop.org](http://www.openloop.org)).

## What is a public policy prototype?

The concept of a **prototype** is traditionally associated with industries in which an experimental process to assess and learn from a sample, model, or preliminary version of something (such as a product) is carried out before it goes to the market. In *design thinking*,<sup>13</sup> a prototype is the visible, tangible, or functional expression of an idea being tested with external parties at an early stage in its development in order to learn from it and iterate the original idea.

A **policy prototype** can be defined as a methodology to test policy efficiency by first applying it in a controlled environment. The creation of policy prototypes takes a user-focused approach to develop laws and policies<sup>14</sup> and allows researchers to test the clarity, viability, and effectiveness of potential rules or policies with a series of practices and activities before they are deployed.

## Why design public policy prototypes?

The idea of developing prototypes comes from the need to create more efficient evidence-based policies, thus preventing the social and economic costs of inadequate policies. Public policy prototypes enable the observation and testing of a policy<sup>15</sup> and invite key players to actively participate in the design process of a concrete policy.<sup>16</sup> The possible effects, strengths, weaknesses, and limitations of the legal frameworks, draft laws, and conduct codes, among others, may be analyzed in this way prior to their definitive and official application.

This methodology generally offers decision-makers the opportunity to learn about and redirect political interventions at an early stage of the process by creating a trial-and-error experimentation space to identify problems associated with the implementation of the policy, which translates into resource savings.<sup>17</sup> It may be especially useful in the case of the regulation of accelerated technology and innovation developments, the impact of which may be uncertain and difficult to foresee.

3



**Open Loop Mexico**



The partners involved in this project proposed and explored the suitability (clarity, efficiency, and viability<sup>18</sup>) of a legal framework intended to strengthen the T&E of companies' AI/Automated Decision-Making (ADM) systems in Mexico. The scope was extended from AI to all ADM systems to ensure technological neutrality considering the possible future development of new technologies. The program partners worked with 10 Mexican companies that used AI for their products or services. This exercise required these companies to implement the framework, as well as a playbook developed to guide its adoption (together referred to as the "prototype"), adjusting and testing practices in their systems and operations to fulfill the prototype's requirement. For this, companies followed a detailed working plan with missions and activities to be fulfilled approximately every two weeks, with continuous guidance and technical support.

## Responsible AI Context in Mexico

As noted in the introduction, potential regulatory frameworks include T&E considerations, as it is a key principle in global AI ethics guidelines. These discussions, however, have mainly taken place in Europe, the United States, and certain Asian countries. Since the widespread adoption of these systems is more recent in Latin America, there is less awareness in the region about what it means to use AI responsibly and the dialogue around potential frameworks for the responsible use and development of AI is still nascent. Colombia was the only country with an AI ethics framework, as of March 2023. Despite the uneven levels of global adoption, it is important for Mexico and the rest of the region (as well as low-to-middle income countries as a whole) to contribute to the international conversations so that the local perspectives, challenges, and opportunities are considered in the development of international good practices and norms.

That being said, Mexico has made progress with regard to AI in the past years. In 2018, a National Artificial Intelligence Strategy<sup>19</sup> was developed by C Minds, Oxford Insights, and the British Embassy in Mexico. It was later adopted by the National Digital Strategy Coordination, positioning Mexico as one of the first 10 countries in the world to have an AI strategy. This document addressed the country's

advantages, opportunities, and challenges with regard to AI and offered short- and medium-term recommendations for the different players in the ecosystem. Despite these initial AI governance efforts, the country must continue strengthening and creating public policy around AI. Indeed, there is insufficient mitigation strategies and tools regarding the possible negative social impacts of AI/ADM systems. According to the National Artificial Intelligence Survey conducted by IA2030Mx in 2019<sup>20</sup> in Mexico, 45% of people were a little worried about the ethical implications or the possible negative social impacts linked to the development of AI, such as bias and data privacy.

Regarding T&E in the Mexican context, some laws and regulations, such as the Federal Law for Consumer Protection, require the disclosure of information or advertising regarding goods and services in a truthful and verifiable way. While the regulation does not disclose how this should be done when it comes to AI systems, it does imply an overarching requirement for transparency in goods and services.

Given the growing impact of AI/ADM systems in our lives, it is increasingly important for Mexico and the region to carry out experiments that contribute to the trustworthiness and reliability of AI/ADM systems.

## What are transparency and explainability in the context of AI/ADM systems?

Since AI/ADM systems have a growing impact on our everyday lives and opportunities due to their widespread use in sectors such as finance, education, and health, it is important to understand how and why the decisions that affect us are being reached in order to preserve human autonomy in the face of intelligent systems.<sup>21</sup> This may represent a technical challenge, especially when these decisions are made by deep learning systems. Indeed, these may operate as black boxes, meaning their internal workings are opaque (not understandable), which makes it difficult to determine the rationale behind a decision.

The aforementioned will be key in making decisions that do not discriminate individuals or groups of the population due to mistakes and undesired bias in the training data or algorithm.<sup>22</sup> In addition, it will contribute to the long-term existence of companies, given that alignment with responsible practices reduces the possibility of crises linked to negative publicity.

While **transparency** could be mistaken for sharing a company's industrial secrets or algorithms, that is not the case. This term has several meanings. In the context of this prototype, it implies disclosing the use of an AI/ADM system (i.e., when generating a prediction, recommendation, or decision or when the user is directly interacting with an AI-driven agent, such as a chatbot). The level of disclosure should be proportional to the system's potential impact on users' rights and freedoms. Users must understand how the system is developed, trained, operated, and implemented to varying degrees depending on the application.

According to the Berkman Klein Center,<sup>23</sup> to fulfill the transparency principle, AI/ADM systems must be designed and implemented in such a way that it is possible to supervise their

operations. The principle of transparency must be applied to the whole life cycle (development and implementation) of AI/ADM systems, including the training data selection, the algorithms, and the model itself.

Furthermore, according to the AI High-Level Group created by the European Commission, **explainability** refers to the capability of people affected by the system's results to understand why a specific result was reached. To achieve this, it is necessary to know the attributes or variables that influenced the final decision. In this sense, explainability is closely linked to transparency, since the results and subprocesses must be comprehensible and traceable.<sup>24</sup>

The OECD AI principles<sup>25</sup> state that to achieve T&E, "*meaningful information relevant to the context and coherent to the state of the art must be disclosed as follows:*

- *foster general comprehension of AI systems so the interested parties are aware of their interactions with AI systems, even in the workplace;*
- *allow the parties affected by an AI system to understand the result; and*
- *allow the parties affected by an AI system to challenge their results based on simple, easy-to-understand information regarding the factors and logic that served as the basis for the prediction, recommendation, or decision".*

It is important not to confuse these terms with interpretability. Although they are complementary principles, they have different meanings.

For the purposes of this public policy prototype, the concepts were defined as follows (for more information, see Annexes A and B):

### Transparency

Transparency is primarily for the benefit of users. It enables them to decide whether to trust an AI system by giving them information about that system. The amount of information varies, but it is generally agreed product makers should give people enough information to meaningfully understand the AI system.

Another way to think about transparency is as disclosure by product makers. People who interact with an AI system will not have the same information as the people who built that system. Transparency helps bring these two sides into balance.

To be considered transparent, product makers may consider whether they are clear, open and honest about how an AI system is built, how it operates, and how it functions. It is probably not enough to simply generate this information and leave it in the open – for someone to understand whether to trust the system, they need to understand that information. That's why transparency usually also requires these disclosures are in an intelligible form.

Transparency is also a mechanism which enables accountability, by making it possible for regulators to scrutinize outputs.

### Interpretability

Interpretability ensures product makers are able to consistently predict the way an AI model makes decisions. This helps to ensure it will act as they intended, and promote a trustworthy AI system. This is not so much about understanding the 'why' of a system, but being able to predict what a given model will do. When there is a high level of consistency prediction, a model is said to be interpretable.<sup>26</sup>

However, researchers and experts generally agree that deep learning and 'black box' models can make interpretability difficult. These AI systems can be hard for humans to understand, making it difficult to predict what they will do with confidence.

One method to address the challenge of interpretability is through model cards that illustrate the internal workings of an AI system. Model cards are interpretable if the model has been analyzed using interpretability frameworks such as Captum.

### Explainability<sup>27</sup>

AI explainability is for the benefit of a product user. It helps them decide when to trust an AI-powered product by ensuring they have a level of understanding around their interactions with an AI system. The aim of AI explainability is not to ensure every person has in-depth technical knowledge about a system and how it works. Instead, it is about helping them understand how an AI system is affecting their experience, and how much control they have over an interaction with a given system. It might give someone the option to change or customize their experience, or let them know what a system cannot do.

*Table 1. Based on the TTC Labs Glossary, available at: <https://www.ttclabs.net/glossary>.*

## Goals for the Open Loop Mexico program

The Open Loop Mexico program had the following objectives:

- provide public policy recommendations for regulators on technologies such as AI/ADM systems based on evidence shared by AI/ADM developers;
- provide an opportunity for regulators and companies to get ahead of emerging issues in AI development and ethical use in Latin America;
- develop mechanisms for collaboration between regulators and innovators in T&E for agile and dynamic learning on the subject;
- contribute to the international conversation about AI ethics through a practical exercise.
- strengthen knowledge on this topic, its opportunities, and challenges, and promote the creation of functional T&E frameworks for AI/ADM systems;
- facilitate a better understanding of the design, development and deployment of AI/ADM systems;
- increase awareness of the importance of safeguarding the rights and freedoms of people in the development and implementation of AI/ADM systems;
- clarify and apply international best practices for the Mexican ecosystem to grow confidence in the development and use of autonomous and intelligent systems; and
- provide companies with an adequate group of guidelines, tools, and practices to ensure more transparent and explainable AI/ADM systems.

## Key players

The following table provides more information about the actors involved in Open Loop Mexico:

### Partners of Open Loop Mexico

#### Meta

Open Loop is a global program that connects policymakers and technology companies to help develop effective and evidence-based policies around AI and other emerging technologies. The program, supported by Meta (previously Facebook), builds on the collaboration and contributions of a consortium composed of regulators, governments, tech businesses, academics, and civil society representatives. Through experimental governance methods, Open Loop members co-create policy prototypes and test new and different approaches to laws and regulations before they are enacted, improving the quality of rulemaking processes in the field of tech policy.

**Role:** Methodology provider based on the global program, co-designer of the regulatory framework and the playbook (prototype), and co-leader of the Open Loop Mexico initiative.

#### C Minds and C Minds' Eon Resilience Lab

C Minds is a women-led organization that promotes the exploration, development, and responsible use of frontier technologies for the benefit of Latin America. In 2019, they co-published the National AI Strategy for Mexico, making it one of first 10 countries in the world to have such a strategy.

C Minds' Eon Resilience Lab is committed to preparing individuals for the future, given the accelerated changes generated by new technologies, and to generating digital inclusion strategies in collaboration with public and private actors.

**Role:** Co-designer of the regulatory framework, co-leader of the initiative, and coordinator and implementer of the program.

#### Inter-American Development Bank (IDB) - Social Sector

The IDB's Social Sector (SCL) is formed by a multidisciplinary team that believes that investing in people helps improve their lives and overcome development challenges in Latin America and the Caribbean. It creates public policy solutions to reduce poverty and the improvement the provision of public services on education, labor, social protection, and health.

**Role:** Supporting and providing input and insights, as well as identifying opportunities to strengthen the program.

## Partners of Open Loop Mexico

### **BID Lab**

The IDB’s Innovation Laboratory promotes private sector development by supporting and testing new solutions to address the problems of economic and social inclusion<sup>28</sup> in Latin America and the Caribbean.

**Role:** Supporting and providing input and insights, as well as identifying opportunities to strengthen the program.

### **IDB’s fAIr LAC Initiative**

This is an alliance (led by the IDB) between the public and private sectors, civil society, and academia to influence both public policy and the entrepreneurial ecosystem and foster responsible and ethical AI use. Four hubs have been established (ecosystems enabling the development and implementation of the fAIr LAC initiative) in Jalisco (Mexico), Costa Rica, Colombia, and Uruguay.

### **National Institute of Transparency, Access to Information and Protection of Personal Data (INAI)**

This is an autonomous constitutional body with the purpose of guaranteeing compliance regarding the right to access public information and the protection of personal data. For the former, it guarantees that any authority at the federal level, autonomous bodies, political parties, trusts, public funds, unions, or any other individual or entity that receives and exercises public resources or performs acts of authority, discloses the public information requested. For the latter, it guarantees the appropriate use of personal data, as well as the exercise and protection of access, rectification, cancellation, and opposition rights that every person has regarding their information.

**Role:** Supporting and providing input and receiving the public policy recommendations.

Table 2. Open Loop Mexico’s Partners






The program also included a group of people with expertise pertaining to different AI topics who supported the program design and execution by reviewing and complementing the documents tested. They also provided direct assistance to participating companies. The following list includes the experts who integrated this group:

- **Carla Vázquez Wallach**, Founder and Director General of Legal + Innovation in Mexico.
- **Daniel Castaño**, Founding Partner of Mokzy, Professor at the Universidad Externado de Colombia, and Researcher and Consultant focusing on AI, digital ethics, and regulation.
- **Edgar Prestes**, Researcher at the Informatics Institute of the Universidad Federal de Rio Grande do Sul, Brazil, Senior Member of the Robotics and Automation Society of the Institute for Electrical and Electronics Engineers (IEEE RAS) and of the Standards Association.
- **Guillermo Larrea**, Corporate Attorney focusing on Latin America at Jones Day.
- **Rafael Ramírez de Alba**, Professor of the Economic Environment Department of IPADE Business School.
- **Ricardo Baeza-Yates**, Research Director of the Experiential Artificial Intelligence Institute at Northeastern University in Silicon Valley.

The prototype test was tested with the following AI companies, all of which had operations in Mexico:



## Participating companies

Name	Sector	Stage	Business model	Description
 ai360 Analítica Inmobiliaria	Real estate	Scaling	B2B and B2G	A platform that estimates housing prices in a faster and more objective way, comparing several properties simultaneously. The model evaluates property attributes and state through deep learning and image recognition.
 Fincomún*	*Finance	Company	B2C and B2B2C	A financial corporation that offers credits and loans to communities not served by traditional credit institutions, typically because they are considered high risk, using models based on automatic learning for the approval, credit or loan collection processes.
 helKi	Education	Early stage	B2C	An application that guides parents or caregivers, in a professional and customized way, in daily parenting challenges, such as predicting growth and risky situations, through a conversational virtual assistant.
 hitch	Human Resources	Early stage	B2B	Platform that optimizes and eases the decision-making of human resources teams when selecting talent by complementing the process with AI-generated interviews. The AI technology uses image recognition and machine learning.
 inndot PIENSA SOLUCIONES	Communication	Scaling	B2B	A social network and digital media management and monitoring platform for companies and governments that suggests automated responses to inquiries.

Continues on next page...



Name	Sector	Stage	Business model	Description
	Health	Consolidation	B2B2C	A pregnancy control device and system that measures obstetric risks (triage) and fetal risk, making a daily assessment of the mother's health to diminish the mother-fetal death.
	*Finances and Human Resources	Company	B2B and B2C	A platform that evaluates interviews with biometric tools and specialized questions to help companies in risk assessment, either to offer financial services and products or for recruitment processes.
	Logistics	Consolidation	B2B	A platform that speeds up transport rates, arrival times, and reduces cargo container retention times, among other systems, to optimize logistics processes.
	*GovTech	Scaling	B2G	A platform that allows governments to offer administrative processes, permits, licenses, and services in a digital secure platform using AI and blockchain.
	RegTech	Company	B2B	A tool that eases the identification of liability for regulatory compliance through regulatory document consultation and analysis using natural language recognition.

Table 3. Open Loop Mexico's Participating companies

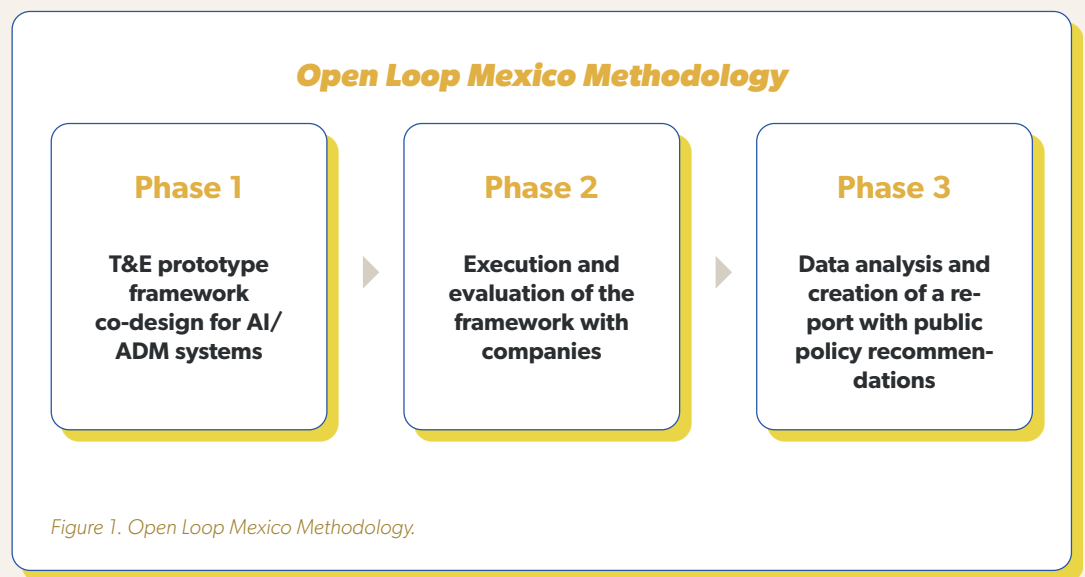
\*Due to the economic impacts of COVID-19, three of the participating companies could not continue with the program.

## Methodology

The project began with the creation of a T&E regulatory framework and a supporting playbook co-created by Meta, C Minds' Eon Resilience Lab, and the IDB. The documents were tested by 10 Mexican companies with a specialized and customized accompaniment process, which included topic exploration, constant iteration, technological development, consultation with experts, and customized training.

The purpose of the program was never to evaluate these companies' products, services, or business models. It was to evaluate the regulatory framework's applicability and to strengthen it based on the experiences of the participating companies, and produce clear public policy recommendations for regulatory institutions in Mexico.

Figure 1 illustrates the key stages in the development of the prototype.



It is important to state that the suggested framework is not a legislative initiative, nor does it replace any governing law. It is an exploration instrument used to generate the necessary information to develop potential nationwide public policies and contribute to the international conversation on AI governance. It is only a starting point and experimentation platform, not a conclusion.

### The regulatory framework and playbook prototype

While other Open Loop programs tested existing or proposed regulatory frameworks, Mexico did not have a T&E regulatory proposal for AI/ADM systems, meaning the team had to develop a proposal to be tested. For this, the Open Loop Mexico team developed two documents to carry out the public policy prototype: a regulatory framework proposal and a playbook for T&E implementation based on international good practices.

The **regulatory framework proposal** was developed upholding the values of human autonomy, determination and respect for human rights, with the aim of promoting the adoption of T&E pillars for companies' AI/ADM systems to reduce potential adverse impacts on users' and people's rights and freedoms. It was formulated and structured like any other regulatory framework, but, as

mentioned above, was not legally binding. Its sole purpose was to obtain remarks on its content and format from the participating companies in the program.

The program partners also developed a complementary **playbook** that served as a guide to comply with the requirements of the regulatory framework. This playbook included detailed explanations of each article in the framework, with more extensive definitions and practical examples, as well as tools to conduct the T&E strengthening process and additional resources for readers who wished to delve deeper into the responsible use of AI/ADM systems.

## Prototype implementation

Both the regulatory framework and the playbook were tested by the participating companies through a series of activities. Each activity examined parts of the documents and lasted between one and two weeks. Using questionnaires conducted via the mobile ethnography platform called dscout,<sup>29</sup> the companies commented on their process and experience of each activity. This stage began in February 2021 and ended in August of the same year.

The implementation cycle of the regulatory framework and playbook by the companies was divided into four key stages:

- **Stage 1. Approach to key concepts**

In this first stage, the participating companies deepened their knowledge of T&E and other AI ethics principles and received a brief introduction to public policy.

- **Stage 2. Evaluation of the regulatory framework and playbook proposal**

This stage aimed to determine how understandable the documents (framework and playbook) were, the feasibility of implementing them in each company's products and services, and how each company would choose to create and focus their T&E solution. This latter exercise was guided by an explainability scenario framework, which helped companies define their T&E solution by selecting the target audience, considering contextual factors that prompted the creation of the solution, electing the purpose of the explanation, and choosing the contents and depth of information to be shared. Box 1 below presents the various options utilized to generate these scenarios.

### Box 1. Explainability Scenarios

The following table helped companies select their explainability scenarios, facilitating the creation of a tailored T&E solution.

Elements	Possible options <sup>30</sup>
<p>Target audience: Who is this explanation for?</p>	<ul style="list-style-type: none"> <li>● Regulator (external auditor/regulatory institution).</li> <li>● Trading partner (another company, customer, supplier).</li> <li>● Consumer (user of the product or service).</li> <li>● Society (the general public).</li> <li>● The company itself (for instance: employees/ staff).</li> </ul>
<p>Context: Was there a specific reason that led to the creation of a T&amp;E solution?</p>	<ul style="list-style-type: none"> <li>● Best practices adoption.</li> <li>● Differentiation from competition.</li> <li>● Proactive approach to enhancing transparency and accountability.</li> <li>● Anticipation of customer needs and expectations.</li> </ul>
<p>Purpose: what is trying to be achieved?</p>	<ul style="list-style-type: none"> <li>● Increase user awareness of their interaction with an AI system.</li> <li>● Facilitate comprehension of the AI system's components and governance.</li> <li>● Enable feedback from consumers, users, trading partners, or regulators.</li> <li>● Engage users in the improvement of the AI system or model.</li> <li>● Provide the option for individuals or entities affected by AI system decisions, recommendations, or predictions to opt out voluntarily.</li> <li>● Influence future behavior of those affected by the AI system's decisions, recommendations, or predictions.</li> <li>● Establish accountability for a more transparent and explainable operation of the AI system or model and its governance.</li> </ul>

Continues on next page...

Elements	Possible options <sup>30</sup>
Contents: what information and what depth of information should be shared?	<ul style="list-style-type: none"><li>● Rationale (information that describes how the decision, recommendation, or selection of the AI system was made).</li><li>● Responsibility (information about who participates in the development, management, and implementation of the AI system).</li><li>● Safety and performance (information about the AI system’s accuracy, feasibility, safety, soundness of decisions, and behavior).</li><li>● Data and models (information about training datasets, algorithmic models used, etc.).</li><li>● Equity (information that guarantees that the AI system is not unfairly biased).</li><li>● Impact (information about the effects of the AI system’s use and people’s decisions).</li></ul>

● **Stage 3. Implementation of customized documents**

Customized plans were developed for each company with activities based on the information in the regulatory framework and the playbook. These plans differed depending on whether the company built its own AI/ADM system or hired a third-party provider to do so, whereby companies with in-house developments explored the quality of their system and carried out bias analyses. In turn, companies with AI/ADM systems from third-party providers explored options and, to the extent possible, the same topics as their colleagues, together with their AI/ADM system provider.

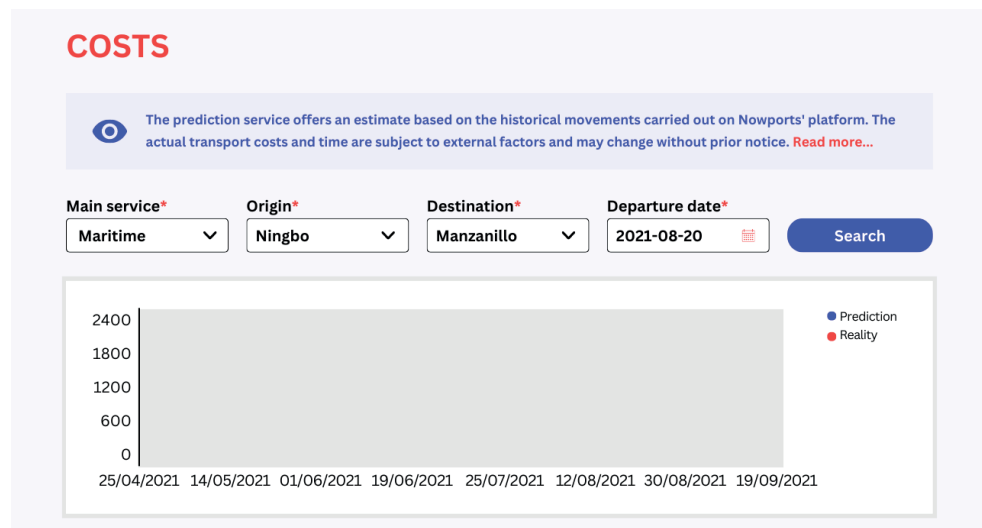
● **Stage 4. Creation of a T&E solution**

With the knowledge acquired during the program, in this last stage companies designed and presented their own T&E solution to a group of people from the partners to receive feedback, and iterate their solution. Results can be viewed in Box 2 below.

### Box 2. T&E solutions developed by the companies

Most of the solutions focused on explaining the rationale for using AI/ADM systems, as well as the data selection and preparation process. To achieve this explanation, most resorted to the use of concise language with simple words that enables end-users or subjects to understand the decision-making process of the system in an accessible and intuitive manner. They achieved that by avoiding technical terms and by using clear and concise language that can be easily understood by a non-expert audience.

This was the strategy employed by Nowports, which offered textual explanations: 1) making the nature and quality of information used by the AI/ADM system transparent, 2) emphasizing data transformation through the model, and 3) being honest about the capacities and limitations of the system. With this in mind, they implemented the solution by adding a brief text with a summary of these points, and if the user of the system wanted to know more, he or she could click on a link and a pop-up box would open.



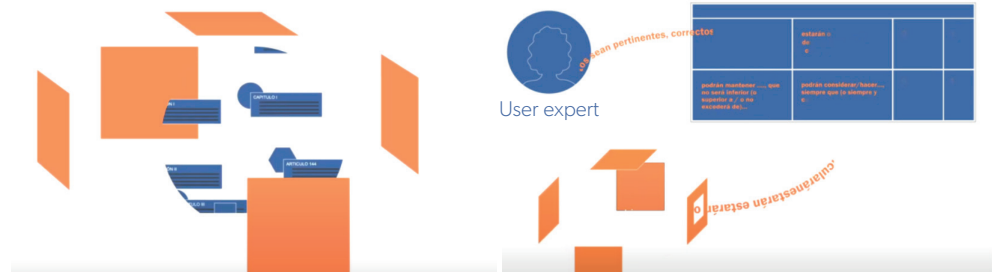
*Replica of Nowport's platform, which includes its transparency and explainability solution proposal (replicated and translated by C Minds)*

Visual representation are quite useful for explaining AI systems, as they provide a simple and intuitive way of representing data, models, and complex relations. Within these solutions, the implementation of visual material such as graphic representations, images and explanatory videos helped make the description of the AI/ADM system more transparent for the end user or subject. Although most of them focused on text, some companies implemented and expressed interest in visuals.

*Continues on next page...*

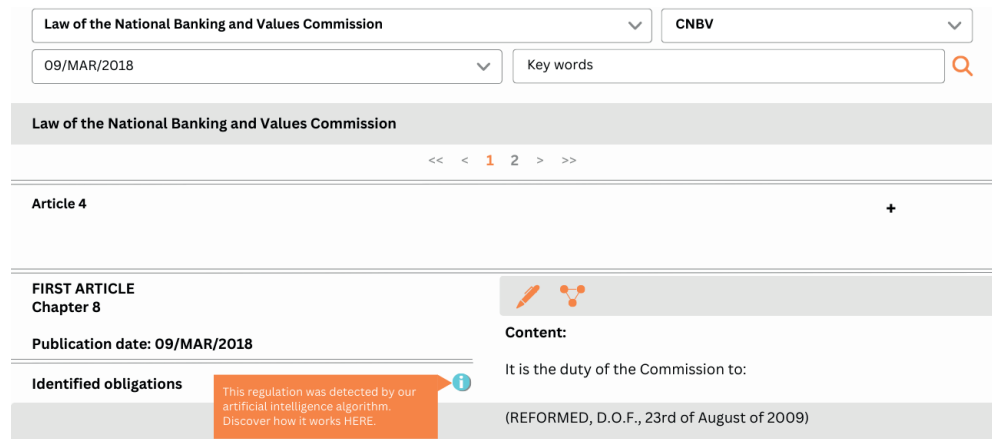
This was the case of Rhisco which decided to create a visual solution for its users and the general public to explain its model in a more understandable way. This solution consisted of two levels:

The first comprised a video, as well as a series of images and gifs, that explained how the AI/ADM system works in general terms.



Examples of visuals created by Rhisco to explain the operation of its AI/ADM system

These visuals appeared automatically when logging in as a platform user. In addition, they integrated a notification to inform the user when a recommendation was made by the AI/ADM system.



Replica of Rhisco's platform, with the notification stating that the recommendation was provided by the AI/ADM system

The second level of the solution allowed users to request more information about the rationale behind a specific recommendation made by the system.

Most participating companies included their explainability message as part of the user experience with the product or service. This user experience included activating notifications, messages, and videos when accessing the application or page. Most companies chose to build messages that could be displayed whenever necessary by the user or subject through a permanent button that displays the information.

## Evaluation criteria

Since the purpose of the program was to test the public policy prototype's efficiency, the following criteria were assessed:

- **Understanding of the regulatory framework/prototype proposal**

This referred to how clear the proposal and its requirements were to the recipients and how well they understood the requirements for compliance.

- **Effectiveness of the regulatory framework proposal**

As the purpose of public policy is to attend specific public problems competently, this criterion refers to the extent to which the proposal contributes to addressing the public challenge at hand, namely, the lack of transparent and explainable AI/ADM systems.

- **Viability of the public policy prototype**

This refers to the feasibility, in terms of technical, economic, human, and time resources, of implementing the regulatory framework, in order to avoid creating **unnecessary barriers**.

## Limitations of the exercise

While this exercise offers a series of lessons on how to improve future similar exercises and how to create public policies around T&E for AI/ADM systems, there are some limitations to consider, which are listed subsequently:

- **Limited representativeness in the size of companies**

As most participants in the program were small and medium-size startup companies (SMEs),<sup>31</sup> the findings and recommendations may not be applicable to all types of companies. The partners recommend conducting similar exercises with large and multinational companies. Nonetheless, the recommendations have the merit of representing companies with less technical, economic, and human resources (i.e. Micro, Small, and Medium Companies, known as MSMEs) in Mexico, which fosters competitiveness within the industry. Moreover, according to Mexico's National Statistics and Geography Institute (INEGI), 99.8% of all businesses in Mexico are MSMEs.<sup>32</sup>

- **Limited representativeness of companies with third-party systems**


There were initially 4 companies using third-party models, but one of them dropped out, leaving only 3, as the framework made more sense for companies that developed their own AI/ADM systems and since they had already participated in a similar exercise, they had already worked as much as possible on T&E.



4



## **Assessment of the public policy prototype**



The evaluation of the public policy prototype on T&E for AI/ADM systems indicates its overall success. The purpose of evaluating the regulatory framework adequacy and the technical manual was fulfilled, providing clear public policy recommendations to regulating institutions in Mexico. The assessment considered three criteria: policy prototype understanding, policy effectiveness, and feasibility. Each criterion is discussed in detail below.

## Framework clarity

The companies participating in the evaluation found the description and purpose of the regulatory prototype to be clear enough. This was reflected in their compliance with and understanding of the activities aimed at improving their systems' T&E. The companies recognized the benefit of aligning the lexicon and terminology between participants and the Open Loop Mexico team to ensure shared understanding.

When asked about how the program equipped them with the necessary tools to translate the framework articles into specific actions, all the companies expressed having a clear path to follow in applying the ideas and articles mentioned in the document. For example, helKi mentioned that the document was helpful in clarifying the implications of applying the recommendations and served as a benchmark for action. However, some companies suggested that the documents could be made more accessible, pointing out the lack of clarity in key concepts due to insufficient detail in the definitions provided. To address this, the companies recommended illustrating the implementation of each principle with hypothetical scenarios.

## Public policy effectiveness

Based on the results and the opinions of the participating companies, it can be concluded that, although there is room for improvement, the public policy prototype was effective. Companies that completed the program successfully fulfilled the purpose of the prototype, which was to create and strengthen their T&E

mechanisms, resulting in the implementation of solutions adapted to their specific contexts.

When asked to rate their compliance with program expectations on a scale of 1 to 10, the companies gave an average score of 9, indicating a high level of satisfaction. The companies described their experience in the program as challenging but valuable in raising awareness of the need for responsible AI tool development. They considered it a rewarding process that paved the way for continued participation in similar initiatives.

Participating companies highlighted significant contributions from the program. These include the reduction of bias, improved understanding of T&E implementation, and enhanced explanation of AI system workings to increase user confidence. Deepening bias analysis, strengthening communication with customers, and improving T&E practices through user testing were also identified as valuable outcomes. The program facilitated awareness of bias and transparency issues, understanding of the hazards of sensitive data use, and exploration of third-party T&E platforms. Basic knowledge of T&E solution development, awareness of ethical limitations, and understanding best practices for mitigating implementation problems were gained. Companies also reported implementing an ethics methodology, accessing expertise on responsible AI/ADM systems, adopting a more transparent approach, and implementing a more transparent policy on personal data use.

All the companies expressed their commitment to implementing T&E practices in all their products and services as an ongoing process. Some companies had already established permanent mechanisms by the end of the program, such as Nowports, which maintained revision and communication processes to address internal and external biases.

## Feasibility

Implementing the framework and playbook requires various resources, including financial, human, time, and technical aspects, which are crucial for feasibility assessment. During the prototype implementation, the companies considered costs, primarily in terms of time and human capital investments.

Regarding time constraints, the companies expressed that the time invested during the program was not excessively burdensome. However, due to limited workforce capacity, they faced challenges in prioritizing program activities alongside other responsibilities. In terms of technical difficulty, the average rating given by the participants was 5.6 out of 10, with 10 being considered "extremely difficult." However, this rating varied depending on the complexity and sensitivity of the models used by the companies. Some companies also acknowledged that individuals without technical skills might find the implementation more challenging, but not insurmountable.

The participating companies also faced several challenges during the prototype implementation. These included a lack of time due to competing priorities, a shortage of human resources, difficulties in resource allocation, managing time across teams, resistance from certain areas within the companies, limitations in resources and understanding of the AI/ADM system, and a lack of technical skills and documentation. Addressing these challenges and implementing the recommendations derived from the evaluation will contribute to the ongoing improvement of the regulatory framework, fostering a culture of responsible AI/ADM systems and ensuring the ethical and transparent use of AI technologies.

Based on the experiences of the participating companies, while the public policy prototype appeared to be effective overall, there were areas of opportunity identified to further strengthen the regulatory framework. For more detailed information, please refer to [Section 5](#).



5



**Specific amendments  
to the regulatory framework  
and playbook**

Based on the results and learning of the companies during prototype implementation, the following adjustments to strengthen the documents were suggested:

## Regulatory framework

**Proposal of modified texts based on remarks**  
(edits and additions in black)

**Remarks by the implementation team**

### **Adjusting the transparency definition:**

Transparency can be defined as the practice of revealing how and why a system made a concrete decision, where operation supervision is possible. The affected persons should understand how the AI/ADM system is developed, trained, operated, and deployed.

**Even though the framework has a section that focuses on T&E, during the prototype implementation, a concise definition of transparency, separate from the concept of explainability, would be quite helpful in differentiating the principles.**

### **Adjusting the explainability definition:**

The explainability of an AI/ADM system refers to letting the persons affected by the results of a system understand, in human terms, why it has been achieved. To achieve this, it is necessary to know what attributes or variables influenced the final decision.

**As above, we also recommend adding another definition of explainability to strengthen the possibility to distinguish between T&E.**

For many automated learning applications, the model is so complex that it cannot be interpreted. **Therefore, "locally interpretable" refers to an explanation of how a specific conclusion has been reached (as opposed to how the model arrives at decisions in general).**

**Because companies were confused about the different concepts and their definition, the information presented regarding local interpretability was reformulated to better differentiate these concepts.**

## Playbook

### Proposal of modified texts based on remarks

(edits and additions in black)

Interpretability **implies that a human being can understand the decision-making process, especially, how the model used in the AI/ADM system reached a decision. It is the ability to determine the model's cause and effect.**

Explainability is how the internal mechanics of an automated decision system can be explained in human terms. The difference in interpretability is subtle. **While interpretability provides a broad understanding of how a system operates, explainability provides an understanding of all attributes and variables that influence decision-making.**

- **Global interpretability (or globally interpretable models)**

**Global interpretability refers to the entire model. For decisions that require full accountability and justification, globally interpretable models are generally preferable.**

- **Local interpretability**

**Local interpretability refers to a specific result. The main challenge of black box models is that they are difficult, if not impossible, for people to understand.**

### Remarks by the implementation team

**During prototype implementation, the companies shared that the definition of interpretability was somewhat ambiguous. Therefore, to avoid confusion, the explanation was rephrased (see in blue).**

**The confusion regarding different concepts requires better and more differentiated definitions.**

**Since the companies were confused about whether their models were globally or locally interpretable, a phrase was added to emphasize this difference.**

**Step 1: Determine the risk of the decision-making process.**

**The use of AI/ADM systems entails an impact that may directly affect people's lives and the environment. AI/ADM systems have uses that could be classified as low or high impact; the main difference is how they could affect people's lives and their rights. They are considered high impact when they can damage a person's health or security or transgress fundamentally guaranteed rights.**

**Therefore, organizations must analyze any and all potential negative impacts on the rights and freedoms of users and people.**

After this risk assessment, and based on the results, organizations must determine what risks are posed by the decision-making process for individual or collective rights and freedoms.

**The importance of carrying out risk-related assessments (to mitigate possible negative impacts) was added to step 1.**

Lastly, although the public policy prototype focused on the T&E of AI/ADM systems, it would be relevant to consider the integration of information pertaining to other AI ethics principles, such as accountability and objection processes, for instance.

6



**Recommendations for  
the formulation of public  
policies focused on AI/ADM  
systems' transparency  
and explainability**



Based on the results obtained in the implementation of the public policy prototype and the remarks shared on the regulatory framework, a series of recommendations are provided in addition to the original documents (see the annexes A and B). These recommendations are aimed at regulating institutions that would like to develop policies related to AI/ADM systems' T&E.

- 1 Proactively promote Artificial Intelligence as a national priority, with a focus on operationalizing trustworthy AI principles**
- 2 Play a proactive role in AI governance in Mexico**
- 3 Build capacity for trustworthy AI in non-technical government bodies and particularly on transparency and explainability (T&E)**
- 4 Increase technical capacity for trustworthy AI in Mexico**
- 5 Invest in AI research and development for trustworthy AI**
- 6 Strengthen the capacity for responsible AI development and adoption in the Mexican workforce**
- 7 Expand civic awareness of AI in Mexico**

**1**

**Proactively promote Artificial Intelligence as a national priority, with a focus on operationalizing trustworthy AI principles**

- Policymakers could use existing resources and leading international practices and tools in order to create a National AI Strategy<sup>32</sup>.
- This strategy could outline the policy goals for AI in line with the OECD and UNESCO AI Principles, as well as the policies that could be needed to achieve those goals. The strategy could also include specific measures to promote transparency and explainability of AI systems<sup>33</sup>.
- This exercise should be a multi-party effort, led by national government bodies and the creation of the strategy could also include the private sector, academia, and civil society via innovative exercises.

## 2

### **Play a proactive role in AI governance in Mexico**

- Policymakers in Mexico could take a proactive role in governing the development and use of AI in the country by: i) organizing and promoting experimental government exercises to identify and address the opportunities and challenges of AI such as public policy prototypes and regulatory sandboxes (before policies/regulations are set in place), as well as hackathons and competitions to further understand opportunities and challenges in this field; ii) developing a clear and concise normative framework for AI, based on local needs and international best practices and iii) promoting cross-sector collaborations to ensure that the AI framework is comprehensive and reflects the views of all stakeholders.

## 3

### **Build capacity for trustworthy AI in non-technical government bodies and particularly on transparency and explainability (T&E)**

- Organize and implement capacity-building sessions and workshops on AI opportunities and risks, with a focus on T&E. Policymakers could work with civil society organizations and academia to organize and implement these, as well as massive online open courses (MOOCs) to level policymakers' and public official's knowledge of AI risks and opportunities, especially related to T&E. Improved capacity and knowledge would allow them to better participate in conversations on the topic.
- Create regular spaces for dialogue with government officials, AI developers, and other stakeholders to discuss issues related to T&E. For example, policymakers could create a task force of government officials, AI developers, and other stakeholders to discuss issues related to T&E. These dialogues could help to build consensus on best practices for AI design, development, deployment and use, and to identify areas where further guidance is needed.

## 4

### Increase technical capacity for trustworthy AI in Mexico

- Policymakers could consider developing a set of technical standards/protocols for AI systems in consultation with AI developers, businesses, and other stakeholders in the Mexican AI ecosystem, leaning on international good practices to ensure they include human-in-the-loop practices when relevant and are aligned with a human centered approach to AI.
- Explore the development of a risk management framework based on the Mexican context that is highly consistent and interoperable with international best practices and standardization efforts<sup>34</sup> for the design, development, and deployment of trustworthy AI systems and reduce their potential for unexpected negative impacts, especially relevant if companies choose to abide by T&E principles. This could be led by the regulatory institutions, in collaboration with the Mexican AI ecosystem.
- In addition to creating local resources, consider gathering existing international resources by countries, companies, and multilateral organizations on a government webpage that is regularly updated.

## 5

### Invest in AI research and development for trustworthy AI

- Policymakers could establish financial and non-financial incentives for promoting AI T&E research projects via governmental bodies and public and private universities, in collaboration with the industry and civil society to ensure a practical approach. Cross-border opportunities could also be considered. In particular, these actors could invest in research on:
  - i) techniques for making AI systems more transparent and explainable, this could include research on methods for visualizing the decision-making process of AI systems, as well as research on methods for explaining the rationale behind AI decisions; and
  - ii) research and tools for identifying and mitigating bias in AI systems, as well as overall risk management frameworks.
- In addition to fostering and funding research, the government, academia and other AI stakeholders could create spaces to share the key learnings, recommendations, and tools that result from the research activities.

**6**

**Strengthen the capacity for responsible AI development and adoption in the Mexican workforce**

- Promote the inclusion of courses and modules on ethical considerations in the development and adoption of AI systems in technical careers linked to data science, computer science, and artificial intelligence, among others. This could apply to formal education spaces like Universities and other learning institutions or courses, including life-long learning ones.
- Social science and humanity careers in formal and informal education spaces, including life-long learning opportunities, could also offer introductory courses and modules to AI systems, to create a more diverse workforce that can focus on responsible AI from different perspectives. This could be promoted by Certification Agencies and policymakers, through bodies like the National Institute of Transparency, Access to Information and Protection of Personal Data (INAI), in collaboration with industry, civil society, and academia, creating training programs on the importance of and how to build transparent and explainable AI systems, especially for developers.

**7**

**Expand civic awareness of AI in Mexico**

- Policymakers could launch a public awareness campaign about the risks and opportunities of AI systems, highlighting the importance of T&E in AI services and products. This campaign could help boost T&E practices as companies use it as a competitive advantage and consumers request them from their product and service providers
- Policymakers, and local youth and education agencies could further promote digital literacy education programs in schools and universities and as lifelong learning courses, with a focus on AI, once the digital basics have been understood, in collaboration with civil society and academia.
- Support the development and deployment of digital and AI literacy resources in Spanish, and work together with the government and AI actors to promote AI literacy.



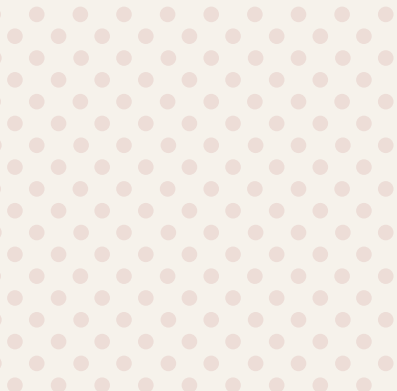
**Conclusion**

AI/ADM systems are increasingly being incorporated into different sectors in Mexico as they are becoming crucial for economic growth and development. This massive adoption is creating brand new opportunities that must be taken advantage of but, at the same time, poses new challenges. In this sense, it is key to continue promoting innovation, but this must be done responsibly, so as to not breach people's rights and freedoms.

This public policy prototype on AI/ADM systems' T&E provides valuable insight into the relevance and importance of incorporating these principles in the development and use of these systems. As observed during the public policy prototype implementation, on the one hand, transparency and explainability allow companies to develop a better understanding of the inner workings of their own AI/ADM systems, including the risks they may pose due to untreated bias, for instance. On the other hand, transparent and explainable AI/ADM systems can generate more trust from the end-user and subjects, as they have a better understanding of how the solution works and get a sense of accountability from the solution provider. The authors believe this program makes a case for putting T&E principles into practice.


The program also demonstrated the efficiency of public policy prototypes to promote the value and importance of a certain topic, in this case T&E, and to generate new insight to contribute to the global conversation on the governance of AI/ADM systems in Latin America and the Caribbean. Public policy prototypes are valuable mechanisms to foster dynamic learning and collaboration among stakeholders and establishing an open and informed dialogue in the region on a highly relevant and complex issue such as T&E. The authors and organizations involved in the program hope that it will contribute to the creation of AI governance frameworks that are practical, inclusive, operational, adapted and adaptable to different contexts, and that place social benefit at the heart of their impact, in collaboration with key stakeholders.

Finally, the authors and organizations involved hope that the lessons learned will incite the development of similar exercises in Latin America and the Caribbean.





# **Bibliography**



Ad Hoc Expert Group (AHEG) for the Preparation of a Draft text of a Recommendation the Ethics of Artificial Intelligence. (2020). Outcome document: First draft of the recommendation on the ethics of artificial intelligence. UNESCO. Retrieved from <https://unesdoc.unesco.org/ark:/48223/pf0000373434>

Brown, T., & Katz, B. (2011). Change by design. *Journal of Product Innovation Management*, 28(3), 381-383.

Buchanan, C. (2018). Prototype for policy. UK Government. Retrieved from <https://openpolicy.blog.gov.uk/2018/11/27/prototype-for-policy/>

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication No. 2020-1. Retrieved from <http://dx.doi.org/10.2139/ssrn.3518482>

Hamon, R., Junklewitz, H., & Sanchez Martin, J. I. (2020). Robustness and explainability of artificial intelligence. Retrieved from <https://publications.jrc.ec.europa.eu/repository/handle/JRC119336>

Hébert, M. (2019). A pilot is not a prototype: How to test policy ideas before scaling. *Apolitical*. (2019). Retrieved from <https://apolitical.co/solution-articles/en/a-pilot-is-not-a-prototype-how-to-test-policy-ideas-before-scaling>

Hernández, M. (2019). Estrategia Nacional de Inteligencia Artificial va por sentido ético y responsable [National Artificial Intelligence Strategy based on ethics and responsibility]. *Forbes México*. Retrieved from <https://www.forbes.com.mx/estrategia-nacional-de-inteligencia-artificial-va-por-sentido-etico-y-responsable/>

Independent High-Level Expert Group on AI. (2019). Ethics guidelines for trustworthy AI. European Commission. Retrieved from <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Institute of Electrical and Electronic Engineers. (2022). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Retrieved from <https://standards.ieee.org/industry-connections/ec/autonomous-systems/>


Instituto Nacional de Estadística y Geografía (INEGI). (2019). Micro, pequeña, mediana y gran empresa: Estratificación de los establecimientos. [Micro, small and medium-size and large companies: Stratification of institutions]. Economic Census.

Kilpatrick, D. (2000). Definitions of public policy and the law. National Violence Against Women Prevention Research Center, Medical University of South Carolina. Retrieved from <https://main-web-v.musc.edu/vawprevention/policy/definition.shtml>

Kontschieder, V. (2018). Prototype in Policy: What For?! (2018). Retrieved from <https://conferences.law.stanford.edu/prototype-for-policy/2018/10/22/prototype-in-policy-what-for/>

Organization for Economic Cooperation and Development (OECD). (n.d.). Transparency and explainability (Principle 1.3). AI Policy Observatory. Retrieved from <https://oecd.ai/dashboards/ai-principles/P7>





Organization for Economic Cooperation and Development (OECD). (2019). Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments. Retrieved from <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>

Reyes, E. (2020). Las empresas mexicanas no saben qué hacer con la Inteligencia Artificial [Mexican companies do not know what to do with Artificial Intelligence]. Expansión. Retrieved from: <https://expansion.mx/tecnologia/2020/07/30/las-empresas-mexicanas-no-saben-que-hacer-con-la-inteligencia-artificial>

Senate of the United States. (2022). Algorithmic Accountability Act. Retrieved from <https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202022%20Bill%20Text.pdf>

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2022). Recomendación sobre la ética de la inteligencia artificial [Recommendation on ethics of artificial intelligence]. Retrieved from [https://unesdoc.unesco.org/ark:/48223/pf0000381137\\_spa](https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa)

White House Office of Science and Technology Policy. (2022, June 24). Making automated systems work for the American People. The White House. Retrieved from <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>

World Economic Forum. (2018). Agile governance – Reimagining policy-making in the Fourth Industrial Revolution. Retrieved from [https://www3.weforum.org/docs/WEF\\_Agile\\_Governance\\_Reimagining\\_Policy-making\\_4IR\\_report.pdf](https://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf)

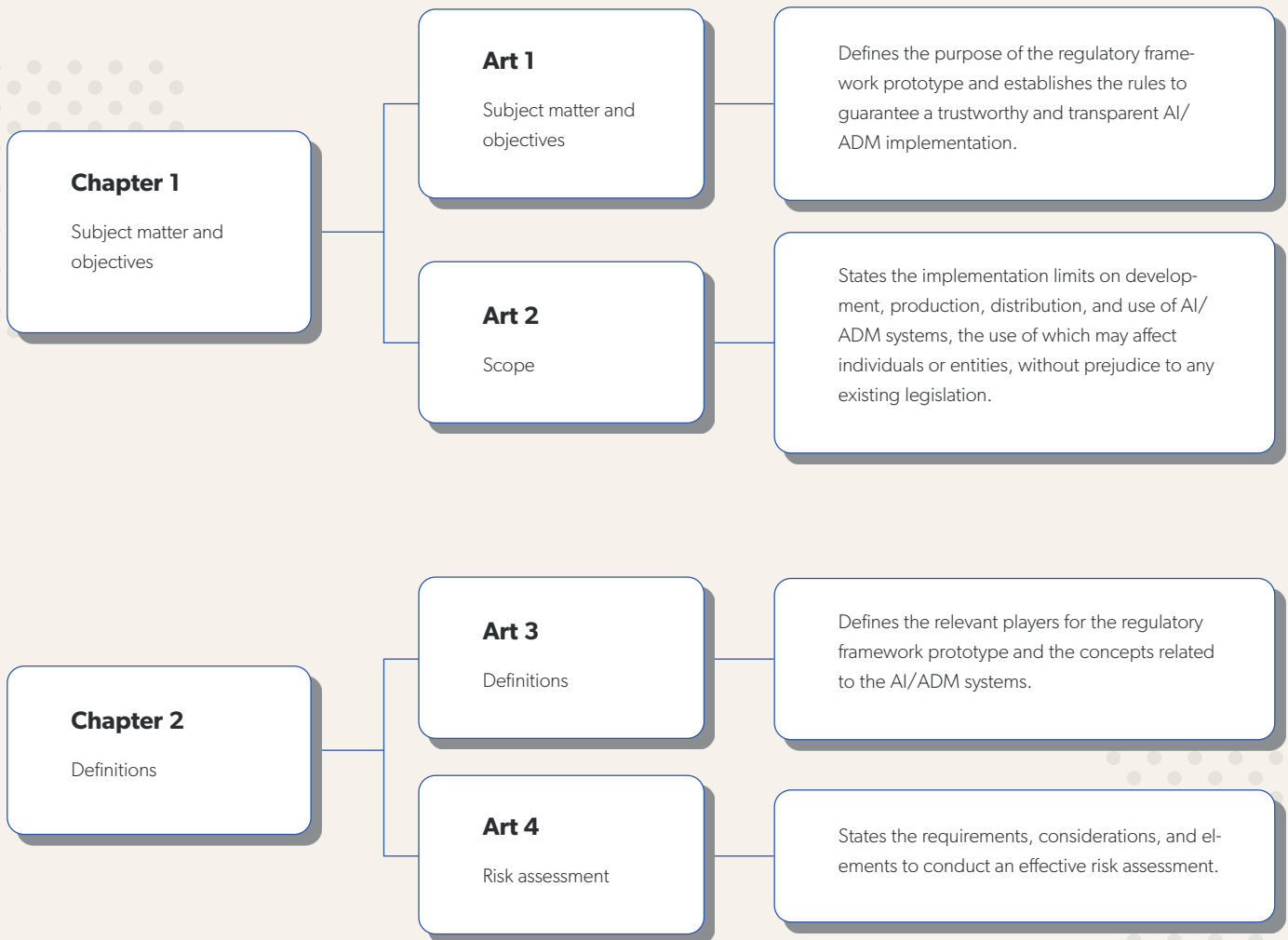


# Annexes

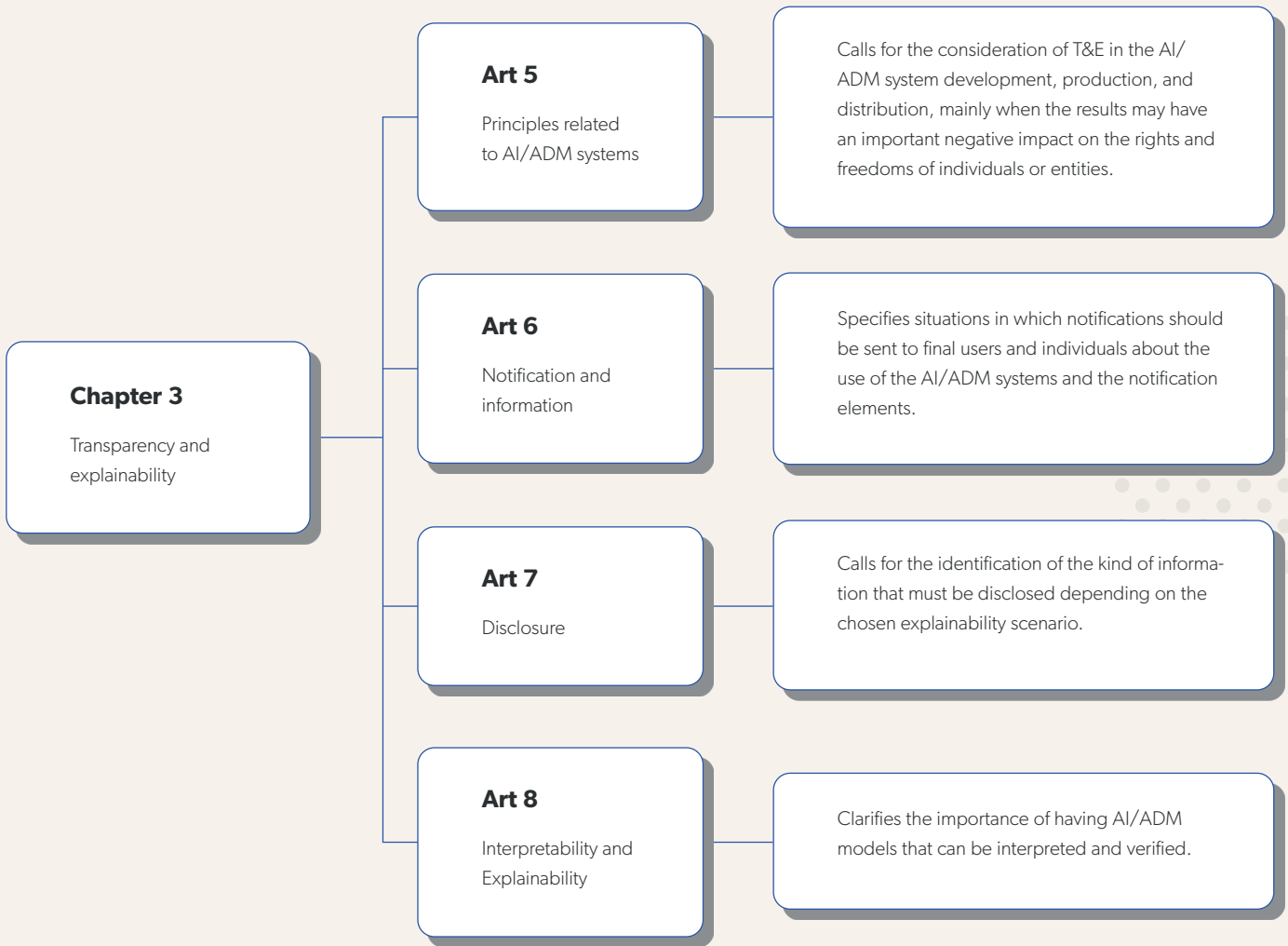
## Annex A. Proposal for a regulatory framework for AI/ADM systems' transparency and explainability.

Below is a summary of the version of the regulatory framework tested by the companies, beginning with an overview of the content before presenting the document.

### Content of the regulatory framework proposal



Continues on next page...



### The full proposal can be found below.

As the conversation around the topic continues to advance, the authors recommend updating the proposed framework if it were to be used in another context

*NOTE: This document is to be used solely for the completion of Open Loop’s Policy Prototyping Program on the transparency and explainability of ai and automated decision-making. The sole purpose of this document is to elicit feedback on its content and format from the participating companies to the Policy Prototyping program. It is a mere exploratory document devoid of any binding or legal normativity. In no way does the policy prototype replace existing laws and regulations that might apply. Requirements and other prescriptive guidance should only be interpreted in the context of this program and are not anticipated for public deployment, as more learning and development is necessary.*

*The point of departure for the prototype is that it is technology-neutral (hence the use of automated decision-making rather than only AI/Machine Learning (ML) in the text) and principle-based. In this way, the Policy Prototype does not apply to a specific technology, sector, or context.*

*This Policy Prototype makes a distinction between the implementation of automated decision-making that poses a significant negative impact on the rights and freedoms of natural and legal persons, and those that do not. For the former category, additional requirements must be met. Furthermore, the Prototype distinguishes between systems that have a human in the loop and systems that do not (fully automated decisions). For the latter category, their use is prohibited in most cases, as they could entail a significant negative impact on the subject (even if the probability thereof is low). Finally, a relevant*

*distinction is that between interpretable models and black box models. When a human is not involved and automated decision-making is used for purposes that may significantly affect the rights and freedoms of natural persons, an interpretable model needs to be used.*

## Preamble and recitals

### Subject matter and scope

**1** Rapid technological developments, especially with regard to artificial intelligence and automated decision-making, have brought new challenges for natural and legal persons' protection and accompanying values and principles. Automated decision-making systems make increasingly significant decisions regarding natural and legal persons, whereas before these decisions were made by humans. Some of these decisions are made by systems that are opaque, unexplainable, and not understandable. These features of automated decision-making systems may lead to a lack of confidence and trust in decisions made by automated decision-making systems.

**2** Trustworthy automated decision-making requires a clear, strong, and coherent regulatory framework, backed by strong and efficient oversight.

**3** This Prototype places a general care duty on all the actors involved in the development, production, distribution, and use of automated decision-making systems to ensure the transparency and explainability of automated decision-making.

**4** When automated decision-making is used for purposes that may significantly affect the rights and freedoms of natural persons, the automated decision-making system should account for lawful, ethical, fair, and trustworthy decisions. When a result is arguably unethical or unfair, reasoning for decision-making needs to be provided (e.g., fairness measurements of the model and fairness metrics used), and the users must have a right to contest the decision. Given the specific nature of automated decision-making, this Policy Prototype helps to ensure a consistent level of protection against harmful, unethical, erroneous, or unlawful decisions made by these systems. This Policy Prototype is without prejudice to existing legislation regarding decisions made about natural or legal persons.

## Definitions

**5** This Policy Prototype refers to various actors in the field of automated decision-making systems, in particular developers, users, end-users, and subjects.

**6** The developer is the natural or legal person who developed the automated decision-making system. This actor may only provide the learning algorithm, but it is more likely that they will be the person or organization responsible for selecting the (training) data and relevant learning algorithms and the subsequent creation or training of the model. Where the developer and the user jointly develop an automated decision-making system, they shall make arrangements with regard to responsibility and liability.

**7** The user is the natural or legal person deploying an automated decision-making system to achieve a particular goal. Generally, this will be an organization, such as tax authorities detecting fraud, social media platforms providing automated personalized recommendations, or banks assessing the creditworthiness of a client. The automated decision system deployed can be a stand-alone system or an integral part of the delivery of a product or service.

**8** The end-user is the natural or legal person who is intended to use the automated decision-making system, as opposed to actors involved in developing or determining its use. The end-user is the actor informed by the decision of the automated decision-making system or deciding based on the outcome of the automated decision—for example, a doctor getting advice on a treatment from an automated decision-making system. The end-user can be an employee of the user or independent of the user, using the automated decision-making system as a product or service of the user.

**9** The subject is a natural or legal person subjected to an automated decision-making system. The decision that this system makes directly or indirectly affects the subject. For example, in the case of the doctor example (see 8), it would be the patient who is the subject of the decision or, in the judge example, the alleged or accused person that is being subject to a legal procedure; or it could be the subjects affected by actions of a hedge fund that uses ADM to choose what options to call or put.

**10** While each actor has a discrete role, these roles might coincide in practice. For example, with autonomous cars, the driver could be simultaneously considered a user, end-user, and subject. At the same time, the car manufacturer may also be considered a user. Designations of roles need to be done on a case-by-case basis, determining which role is associated with an actor as well as their associated responsibilities.

**11** Automated decision-making systems are often too complex to be fully understood. This Policy Prototype differentiates between transparency (in the sense of disclosure and information) and the explainability of automated decision-making models and automated decisions. The explainability of a decision-making model and its decisions depends on the interpretability of the model. Simple decision-making models (e.g., simple linear regression, decision trees) are considered “globally interpretable” and can be fully understood. For many machine learning applications, the model is so complex that it cannot be interpreted. When an individual decision of such a model can be interpreted, the decision is considered “locally interpretable.”

**12** Whether an automated decision has a significant negative impact on the rights and freedoms of natural and legal persons, the context, nature, purpose, and scope of the implementation must be judged. In assessing the impact on natural and legal persons, the fact that a decision-making model cannot be interpreted, or individual decisions cannot be explained, shall specifically be taken into account.

**13** Effects with significant negative impact on the rights and freedoms of natural and legal persons may include loss of life or injury, financial or property harm, professional reputational damage that interferes with a person’s livelihood, and interference with fundamental rights such as the right to equal treatment, right to privacy, the right to personal data protection, and the right to freedom of speech. In the context of automated decision-making, particular attention should be given to economic, psychological, and societal harms that may follow from automated decision-making. Significance should consider both individual and collective rights and freedoms. Regarding the latter, one should note that a negative impact on the rights and freedoms at the individual level could be deemed a significant negative impact as it could be repeated in a group of collectivity, meaning it is no longer an individual issue, but a collective one.

**14**

The impact of an automated decision-making system depends on the context in which it is used. For each context, those who deploy automated decision-making systems should assess which individual and collective harms are relevant to consider at the individual or collective level. These include inter alia (legal) effects that lead to a loss of economic opportunity, namely (but not necessarily) price discrimination, employment discrimination, or unfair commercial practices; effects that lead to psychological harm such as self-censorship, loss of self-worth, and loss of personal autonomy; and collective harms such as a loss of liberty and economic or political instability.

### ***Transparency and explainability***

**1**

Humans should be aware that they are interacting with a machine, especially when this is not apparent from the interaction, given the intelligence of the system. Furthermore, subjects should be aware that a decision about them is being made by a machine rather than by a human or by a combination of both. Therefore, both end-users and subjects should be informed of the existence of an automated decision-making system when they interact with such a system or when decisions are made by such a system. This disclosure should inform them about the existence of the automated decision-making system, its purpose, its capabilities and known limitations, and the user deploying the system. In addition, users and subjects should be allowed to enforce their privacy and data protection rights when their personal data is processed as part of a decision-making process.

**2**

To enable an understanding of the decision-making process and accountability, relevant aspects of the automated decision-making process should be documented internally and may need to be disclosed upon materiality. This includes, the training data, the learning algorithms, and the model itself, as well as processes such as data selection, model creation, model validation, testing, and model governance (e.g., creator, creation date, last retraining, etc.). The scope and depth of the disclosure, as well as the relevant audience for the disclosure, depend on the goal of the disclosure. Goals may include explainability, accountability, and internal and external oversight. To protect the intellectual property rights and other legitimate interests of the developer or user, the audience to which the information is disclosed may be bound to secrecy.



**3**

Decisions made by automated decision-making systems should generally be comprehensible and verifiable. For automated decision-making systems that may have a significant negative impact on the rights and freedoms of natural and legal persons, these requirements are mandatory. In these cases, the general rule is to only use models that can be fully interpreted by humans (globally or inherently interpretable models). Only in those cases in which the user can show that more complex models (e.g., non-inherently deep neural networks) will provide better accuracy and performance may a non-globally or non-inherently interpretable model be used. This is only allowed if the individual decisions can be explained with a significant degree of accuracy, and if other risk-reducing measures have been taken.

**4**

Existing standards for the explanation or justification of decisions should apply to automated decision-making. For instance, if in a legal verdict it is required that the judge explain how different pieces of evidence were weighed and contributed to the final decision, the same standard should be met by an automated decision-making system if it is used in court. In other words, if the automated decision is replacing a human decision, and the human decision used to require a specific level of information, the automated decision should require the same level of explanation, if not greater.

## Chapter 1: Subject matter and objectives

### **Article 1: Subject matter and objectives**

- 1.1 This Policy Prototype lays down rules to ensure a trustworthy and transparent implementation of automated decision-making.

### **Article 2: Scope**

- 2.1 This Policy Prototype shall apply to the development, production, distribution, and use of automated decision-making systems whose use may have an effect on natural and legal persons.
- 2.2 This Policy Prototype is without prejudice to any existing legislation, in particular in the area of fundamental rights, data protection, and unfair commercial practices.

## Definitions

### **Article 3: Definitions**

- (a) "Actors" refers to the developers, users, end-users, subjects, and any other party that contributes to the design, development, production, distribution, training, or deployment of automated decision-making systems or is affected by such a system or its decisions.
- (b) "Algorithm" refers to a finite sequence of instructions or set of rules designed to complete a task or solve a problem.
- (c) "Model" refers to the result of using a machine-learning algorithm with specific training data. This model is a mathematical representation of the learned domain and is used to map inputs to outputs. The model is the primary component of an automated decision-making system and is used by an algorithm to generate a decision.
- (d) "Fully automated decision" refers to a decision made by an automated decision-making system that is implemented without any meaningful human intervention.
- (e) "Automated decision-making system" refers to a computational process such as one derived from machine learning, statistics, or other data processing techniques that makes a decision or facilitates human decision making.
- (f) "User" refers to the natural or legal person deploying an automated decision-making system to achieve a particular goal.
- (g) "End-user" refers to the natural or legal person using the automated decision-making system for the purposes intended by the user.
- (h) "Subject" refers to the natural or legal person subjected directly or indirectly to a decision of an automated decision-making system.
- (i) "Interpretability" refers to the level of understanding of the decision-making process, in particular understanding of the model used in automated decision-making.
- (j) "Global interpretability" refers to the capability of an individual to comprehend the entire model at once and all the different automated decisions that can be made by the model. This level of interpretability involves understanding how the model makes decisions based on a holistic view of its features and each of the learned components, such as weights, other parameters, and structures.
- (k) "Local interpretability" refers to the extent to which an individual automated decision can be understood or explained by determining how a particular input led to a particular output.

- (a) “Significant negative impact on rights and freedoms of natural and legal persons” may include loss of life or injury; financial, property, or psychological harm; and interference with fundamental rights such as the right to equal treatment, the right to privacy, the right to personal data protection, and the right to freedom of speech. Significance should consider the impact to rights and freedom on an individual but also collective level. A negative impact on the rights and freedoms at the individual level could be deemed a significant negative impact as it could be repeated in a group of by way of repetition in a collectivity, meaning it is no longer an individual issue, but a collective one.

#### **Article 4: Risk assessment**

- 4.1 Prior to the deployment of an automated decision-making system, the user shall assess the risks of the envisaged automated decision-making system and its implementation on the rights and freedoms of natural and legal persons.
- 4.2 In those cases where the implementation of an automated decision-making system is likely to result in a significant negative impact on the rights and freedoms of natural or legal persons, the user shall carry out an automated decision-making system impact assessment prior to deployment.
- 4.3 An automated decision-making system impact assessment referred to in paragraph 4.2 shall in any case be required in case of:
  - (a) potential unfair bias or discrimination toward subjects, including but not limited to price discrimination, employment discrimination, or unfair differential access to services;
  - (b) potential loss of control or agency for the subject, including economic or psychological manipulation;
  - (c) large-scale implementation of automated decision-making that may affect communities or society as a whole; or
  - (d) systematic and extensive or large-scale data processing that presents a high risk to a subject’s data protection rights, including profiling and systematic monitoring.

- 4.4 An automated decision-making system impact assessment should contain at least:
- (a) a detailed description of the automated decision-making system, its design, its training, data, and its purpose;
  - (b) an assessment of the quality, integrity, and representativeness of the data used to train the underlying model;
  - (c) an assessment of the risks involved for natural and legal persons, with a specific focus on subjects and end-users; and
  - (d) the measures envisaged to address the risks, including safeguards, security measures, periodicity of revisions, and mechanisms protecting the rights and freedoms of end-users and subjects, and to demonstrate compliance with this Prototype, taking into account the rights and legitimate interests of those concerned.
- 4.5 In those cases in which the automated decision-making impact assessment indicates that the implementation may result in a high risk to the natural rights and freedoms of natural and legal persons, and these risks can or will not be mitigated, the user shall seek the approval of the supervisory authority prior to deployment.

## Transparency and explainability

### ***Article 5: Principles related to the development and use of automated decision-making systems***

- 5.1 In the development, production, distribution, and use of automated decision-making systems, actors shall consider, taking into account the context, scope, purpose, and nature of the implementation, the transparency and explainability of their automated decision-making systems.
- 5.2 Where an automated decision-making system makes decisions that are likely to result in a significant negative impact on the rights and freedoms of natural or legal persons, the user shall take the necessary technical and organizational measures to ensure that the use of an automated decision-making system is transparent and that the outcomes of automated decision-making are explainable.

**Article 6: Notification and information**

- 6.1 Users and developers, when applicable, shall notify end-users and subjects of the use of automated decision-making systems in those instances in which:
- (a) their use may have a significant negative impact on their rights and freedoms, or when the automated decision-making system interacts with end-users or subjects as a human would.
  - (b) subjects as a human would.
- 6.2 Users shall provide meaningful information to end-users and subjects about:
- (a) the purpose(s) of the automated decision-making system and, where appropriate, the justification for its use over human decision-making;
  - (b) the possible impact of the decision-making system on the rights and freedoms of end-users and subjects;
  - (c) the logic of the decision-making process in line with the requirements in Article 7; and
  - (d) their right to contest automated decisions and the way in which these rights can be exercised.
- 6.3 The information described in paragraph 6.2 shall be clear, concise, accessible, and easily readable. The information provided to end-users may be more detailed and adapted as necessary to the particular intended group to meaningfully facilitate its understanding.

### **Article 7: Disclosure**

- 7.1 Taking into account the context, nature, and scope of the implementation and the risks the automated decision-making system may pose, developers and users of automated decision-making systems shall disclose relevant elements of their development, operation, and use.
- 7.2 The relevant elements mentioned in paragraph 1 include, but are not limited to:
- (a) the rationale for automated decision-making use;
  - (b) the training, testing, and validation data;
  - (c) the algorithms used;
  - (d) the decision-making model;
  - (e) the process of data selection and preparation;
  - (f) the process of training, selecting, validating, and testing the model; and
  - (g) the process of managing and maintaining the model in operation.
- 7.3 The disclosure of the relevant elements of an automated decision-making system should consider the audience, means (messaging medium), and goals of the disclosure. This could include:
- (a) the subject, to provide insight into the logic of the decision-making process;
  - (b) the internal units within the organization responsible for risk management, compliance, or internal audits as part of the exercise of their functions;
  - (c) external auditors for the purposes of auditing or third-party verification; and
  - (d) supervisory authorities for the purpose of compliance, investigations, and general oversight.

**Article 8: Interpretability and explainability**

- 8.1 Decisions of an automated decision-making system shall be interpretable and verifiable, taking into account the nature, scope, context, and purpose of the automated decision-making system.
- 8.2 When fully automated decisions may have a significant negative impact on the rights and freedoms of natural and legal persons subjected to the decision-making process or affected by it, a globally interpretable model must be used.
- 8.3 Paragraph 2 shall not apply when:
  - (a) The user can argue that a non-globally interpretable model is strictly necessary for the purpose of automated decision-making; only in those cases where the user can show that more complex models (e.g., deep neural networks) will provide better accuracy and performance may a non-globally interpretable model be used. This is only allowed if the individual decisions can be explained with a significant degree of accuracy, and if other risk-reducing measures have been taken.
  - (b) The user can provide explanations for individual decisions with a sufficient degree of accuracy, for instance, using local interpretation methods; and
  - (c) The user has taken the necessary technical and organizational measures to protect the rights and interests of subjects.
- 8.4 End-users and subjects shall be given explanations for individual decisions in a clear, concise, accessible, and understandable format.

## **Annex B. Proposal for a practical manual for the adoption of principles of AI/ADM systems' transparency and explainability**

*The playbook would complement forthcoming legislation and could be the basis of co-regulatory and soft law instruments: codes of conduct, codes of practice, standards, certifications, industry guidelines, etc.*

*In this section, we set out ways to comply with the proposed prototype law. An organization that implements the elements from the playbook is in a good position to comply with the prototype law.*

### **AI transparency and explainability: Why do it?**

Based on the assessment of the AI literature, we can conclude that there are no universally accepted definitions of requirements such as transparency, interpretability, auditability, explainability, comprehensibility, and traceability, and they are often used in different ways. Furthermore, in most of the documents analyzed, it is not clearly defined what purpose a requirement specifically serves in a given context.

It is worthwhile to determine (within a given context) how transparency and explainability requirements should be interpreted. In deciding how legal and ethical requirements for transparency, interpretability, etc., should be met, the following questions should be considered:

- Should there be certain restricted intended uses for AI use? Why?
- What potential problems can be caused by opaque/inscrutable automated decision-making?
- In a given context, what are the impacts of these potential problem?
- Are these impacts felt at the individual level, the collective level, or both?
- How will understanding of the automated decision-making process through transparency/interpretability, etc., reduce these impacts—i.e., what goals are we trying to achieve by increasing understanding of the decision-making process?
- What are the requirements for achieving our overarching goals?
- Are there alternative options that would be more effective at realizing these goals?
- What is the best way to identify and take into account the relevant audience?
- How can one explain to an end-user or subject how automated decision-making works in a given context?

We capture these questions and possible answers in Table 1.



## What potential problem are you trying to solve with your explanation?

Potential problems	Goals	Requirements	Possible solutions	Focus of transparency/explainability	Main target audience
It is unknown whether an automated decision-making system is used, or it is unknown what it does	Notify and inform users, end-users, and subjects of the use of an automated decision-making (ADM) system	Notify users, end-users and subjects that they are using an ADM, or are subject to decisions made by an ADM, and provide them with (high-level) information on its functioning	Information statements, notifications, and warnings	ADM system as a whole	Users, end-users, subjects, supervisory authorities, society at large
It is unclear if a model is working correctly (to the extent that it is possible to know) and makes accurate/fair/timely/ proportionate decisions in a real-world environment	Account for the correct operation of the model in terms of safety, fairness, etc.	Provide insight into the operation of a model (transparency, interpretability, auditability, traceability)	Model assessment, global interpretability, local interpretability	Training data, learning algorithms, trained model, model outcomes, input data	Users, end-users, subjects, supervisory authorities, auditors, society at large
It is unclear how a particular decision was made	To enhance understanding of the scope of automated decision-making and the reasons for a particular decision	Explain why a particular decision was made in a particular case (explainability)	Global interpretability, local interpretability	Input data, model outcomes	Users, end-users, subjects, supervisory authorities, auditors, society at large
It is unclear if the decision that was made meets the (legal or otherwise) thresholds for a justified decision	Provide justification for a particular decision	Provide an explanation for the decision and proof that the decision was made following the proper (legal or otherwise) procedure/standard for this decision-making process (explainability, traceability)	Model assessment, global interpretability, local interpretability	Training data, learning algorithms, trained model, model outcomes, input data	Users, end-users, subjects, supervisory authorities, society at large

Continues on next page...

Potential problems	Goals	Requirements	Possible solutions	Focus of transparency/explainability	Main target audience
The reasons for coming to a particular decision are unknown, making it difficult to contest or challenge the decision	To help contest a decision	Understanding why a particular decision was made (explainability)	Local interpretability	Input data, model outcomes	Subjects
The reasons for coming to a particular decision are unknown, making it difficult to determine how to change the outcome of the decision in the future	To alter future behavior to potentially receive a preferred outcome	Understanding why a particular decision was made (explainability)	Local interpretability, counterfactual explanations	Input data, model outcomes	Subjects
It is unclear how the whole AI system works and why users should trust it	Have people understand that a particular product or experience is powered by AI and how that AI works	Societal reliance/ acceptance: This type of explanation is designed to generate trust and acceptance by society. For example, if an unexpected output is provided by the system, the explanation may help users understand why it was generated. It may also provide an increased sense of comfort with the system if a rationale can be provided	Model-, system-, and process-level explainability to elucidate how internal cost benefit analyses are conducted	An ADM system, including model inputs, final results, potential impacts	End-users, subjects, supervisory authorities, society at large

Table 1. Understanding automated decision-making – A guide to explainability based on the problem you are solving

Another way to address these questions, and get to the corresponding answers is to contextualize the transparency and explainability requirement into its fundamental components: audience, context, purpose and content.

## Transparency and Explainability elements

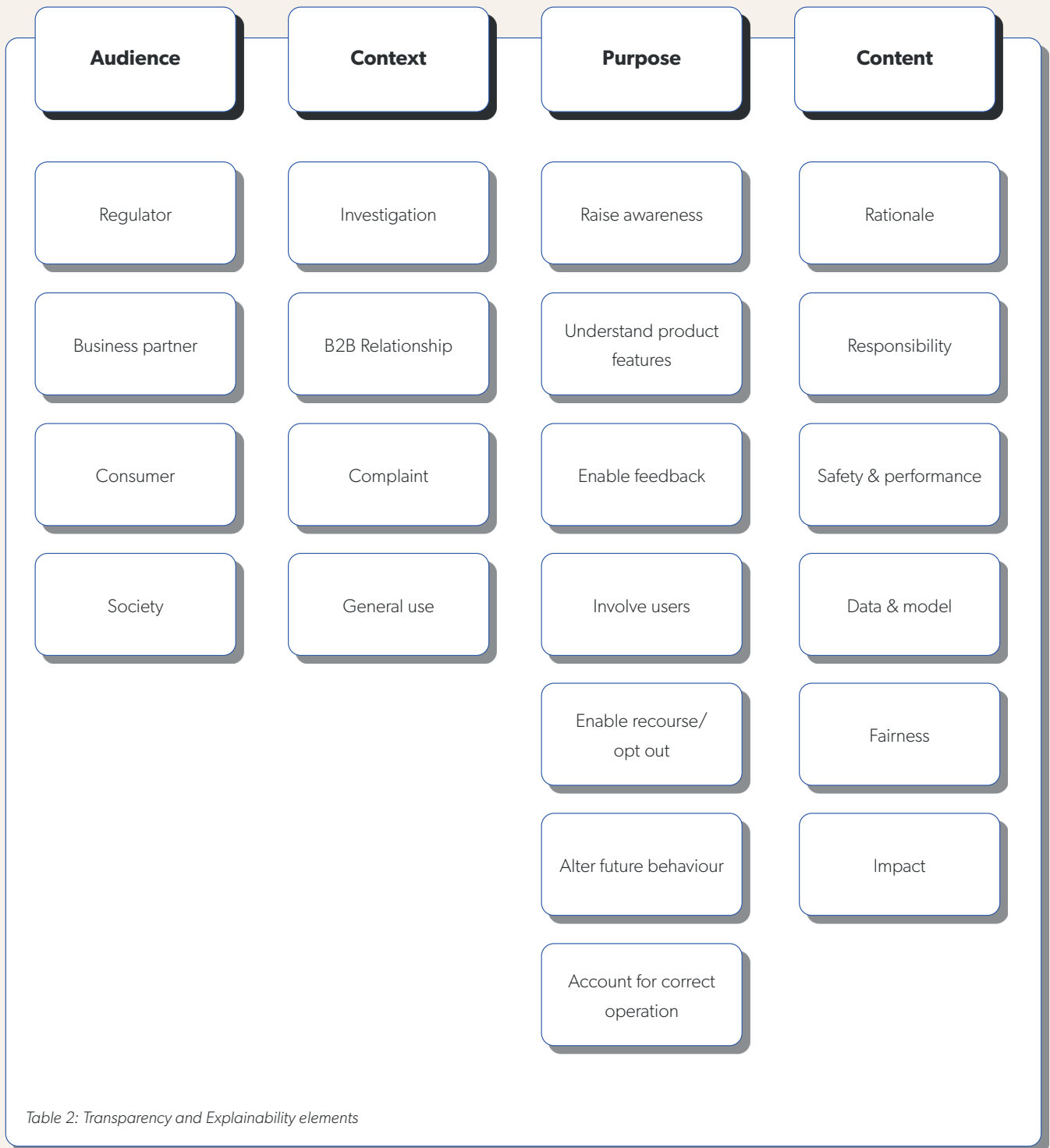


Table 2: Transparency and Explainability elements

**Audience:** The recipients/addressees of the explanation.

**Context:** The reasons why an explanation is (or is not) important and is being requested.

**Purpose:** What the explanation is trying to achieve and the main reason why an AI explanation is provided.

**Content:** The information given to the recipients in the explanation and what to focus on as information to be provided.

## Requirements of the public policy prototype

Based on the potential problems and goals described in Table 1, we see that different requirements and possible solutions emerge. Based on these, three separate but related legal and ethical obligations were described in the Policy Prototype: 1) notification and information (Article 6), 2) interpretability/explainability (Article 8), and 3) disclosure (Article 7).

### Notification and information (Article 6)

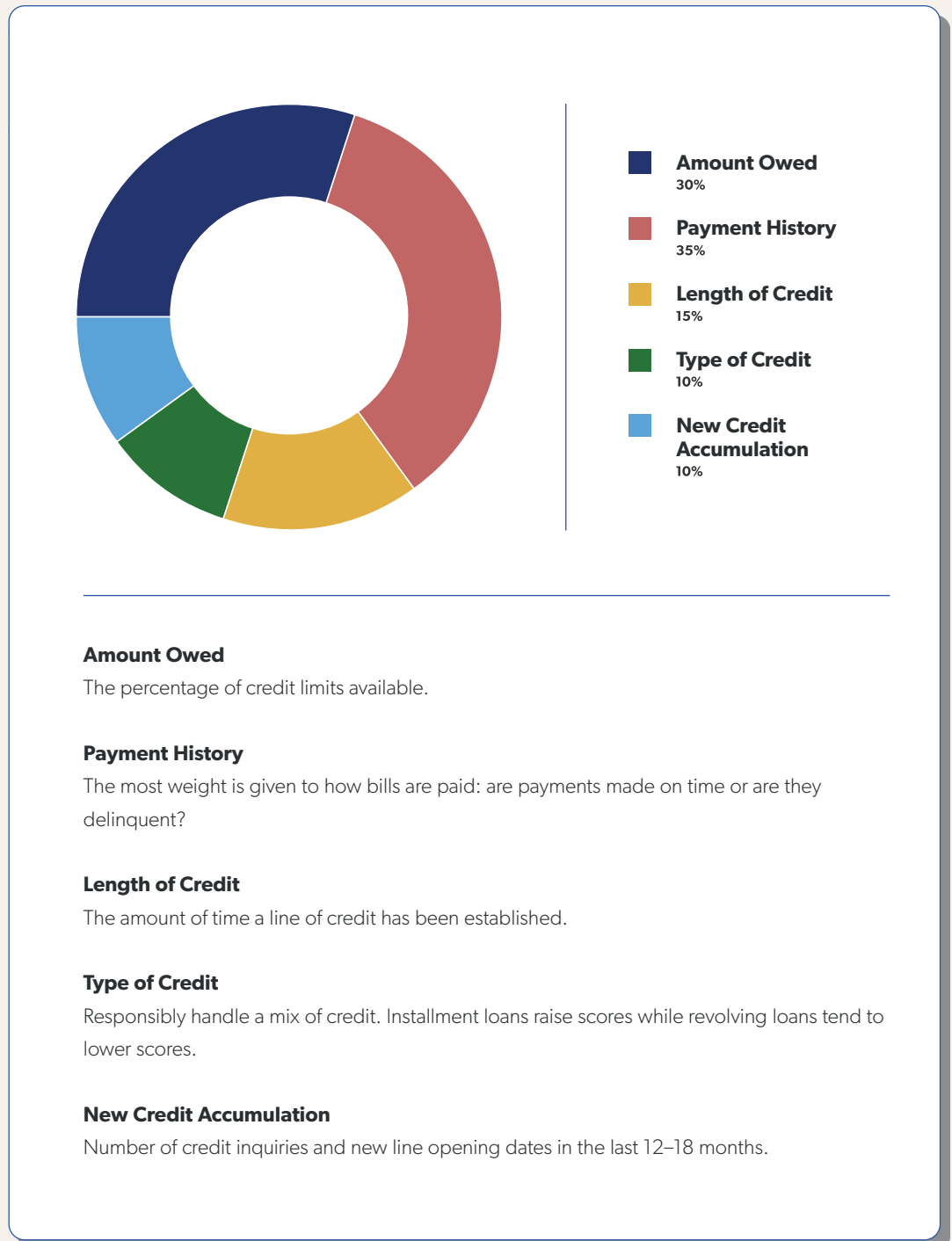
The Policy Prototype requires users to notify users, end-users, and subjects about the use of an automated decision-making system in two specific situations:

- when the use of the automated decision-making system may have a significant impact on rights and freedoms or perceptions (e.g., echo chamber effect, distortions of reality), and
- when the automated decision-making system interacts with end-users or subjects as a human would.

Users (and, depending on the case, developers) can fulfil the notification obligation in various ways—for instance, by providing a small amount of text or (generally accepted) logo to signal the existence of automated decision-making (ADM).

The content of the information requirement is dependent on the context, and ties in with the disclosure requirement of Article 7 and the explainability requirement of Article 8. The Policy Prototype sets four requirements. To meet the information requirement, full disclosure of algorithms, etc., to subjects is not necessary in all cases, as it will not help them understand the automated decision-making process. Rather, subjects should understand why automated decision-making is taking place, how it can impact them, what the logic of the decision-making is, what data is used to make decisions, how the decision can be contested, and how they can opt out of automatic decision-making or be subject to other rights enforcement. For the logic of the decision-making, it is important that subjects can understand how a decision is made and how their particular situations are impacted by the algorithm/model.

The explanation given by Dupaco Credit Union on how a credit score is calculated serves as a good analogy for the features that could inform an automated decision:



What having a particular score means in practice (impact/consequences) is described as follows:

Score	Credit Union
<b>720 and up</b>	Score 720 or higher and you'll have a good chance of obtaining loans at the best interest rates. These loans may require less documentation and paperwork, and potentially less—or ven no—down payment or collateral.
<b>680–720</b>	Score in this range, and you'll usually be able to negotiate good terms.
<b>620–680</b>	Landing in this range will place you under "standard" company rules, giving you less flexibility in choosing better loans or services.
<b>580–620</b>	You'll be reviewed with a critical eye and will need compensating factors to be approved by companies for most loans or services.
<b>Under 500</b>	Not a happy place to find yourself. You'll typically be required to provide a substantial down payment/collateral and/or pay a higher interest rate.

Another option is to use “data nutrition labels.” The Data Nutrition project provides a prototype on their website: <https://datanutrition.org/>.

### Interpretability and explainability (Article 8)

Article 8 of the Policy Prototype concerns itself with the interpretability and explainability of automated decision making.

Let us start by defining each of these concepts:

- Interpretability is the extent to which you are able to predict what is going to happen given a change in input or algorithmic parameters. Or, to put it another way, it is about being able to discern changes in a decision when the input changes without necessarily knowing why.
- Explainability, meanwhile, is the extent to which the internal mechanics of an automated decision system can be explained in human terms. It is easy to miss the subtle difference with interpretability, but consider it like this: interpretability is about being able to discern the mechanics without necessarily knowing why. Explainability is being able to explain quite literally what is happening.

This article specifically aims to address the “black box problem” of automated decision-making. The biggest criticism of high-impact automated decision-making based on AI/ML is that oftentimes the

results cannot be explained and verified because the model cannot be interpreted due to its inherent complexity.

Therefore, users should first consider, based on the context and their implementation, whether complex machine learning models are warranted.

#### Global interpretability (or inherently interpretable models)

The most effective way of avoiding the pitfalls of black box models is not to use them. For decisions that require full accountability and justification, globally interpretable models are generally preferable. A major drawback is that the automated decision-making process for globally interpretable models is limited to using rule-based systems and simple predictive models. This may entail, in certain cases, a trade-off between interpretability on the one hand, and accuracy, effectiveness, and efficiency on the other. Nevertheless, for certain types of decisions (e.g., determining whether somebody is guilty of a crime or automatically selecting a medical treatment), global interpretability may be a mandatory requirement.

In the literature, such “red lines” have not yet been drawn, though requirements set by different lawmakers (e.g., the United States Congress, the European Commission, and the Council of Europe) may entail such a ban in practice. Furthermore, in the context of administrative law and criminal law, it is a general principle that decisions that cannot be explained or are not accompanied by a justification are invalid.

#### *Local interpretability*

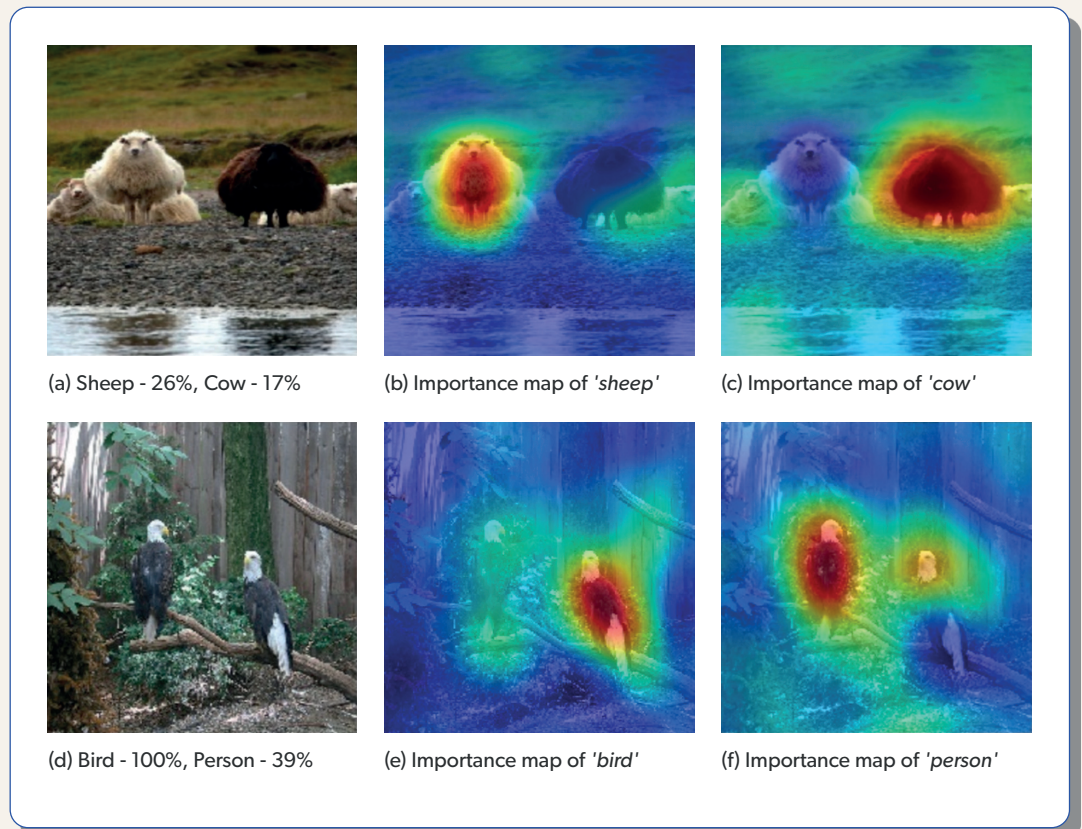
The main criticism, and challenge, of black box models is that they are difficult, if not impossible, to comprehend by humans. The nascent field of explainable AI (xAI) is concerned with the explainability of automated decision-making, more specifically that of black box models. This field of research is rapidly expanding, and it is important to keep up with the various approaches being developed and presented to render black box models explainable. For an extensive (though not exhaustive) overview of these approaches, see Annex 1. Based on our literature review, there are different methods to achieve the goal of explainability, such as by creating a *proxy model* that behaves similarly to the original model, but in a way that is easier to explain, by creating a *salience map* to highlight a small portion of the computation that is most relevant, or through *automatic rule extraction*. What these methods have in common is that they provide local interpretability; in other words, they do not explain the whole model, but rather how a particular conclusion was reached.

#### *Proxy models*

Proxy models, also called local surrogate models, are interpretable models used to explain individual predictions of black box machine learning models (Molnar, 2019). The most well-known example of such a proxy model approach is LIME (Locally Interpretable Model-agnostic Explanation; Ribeiro et al., n.d.). LIME generates a new dataset consisting of permuted samples and the corresponding predictions of the black box model. On this new dataset, LIME then trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest (Molnar, 2019). Decision trees can also be used to approximate neural networks.<sup>33</sup> Finally, SHAP (SHapley Additive exPlanations) also provides a local explanation for a model outcome (Lundberg & Lee, 2017).

*Saliency mapping*

In image recognition, the concept of a saliency map contributes to improving understanding. For a given classification of a picture, a saliency map can be drawn up to show which distinguishing parts of the image were used to arrive at a particular conclusion. This helps users gain an understanding of which parts of the picture are most relevant for the classification. An example of this approach is RISE (randomized input sampling for explanation of black box models), which shows which parts of an image were most relevant in classification (Petsiuk et al., 2018).



*Automatic rule extraction*

Automatic rule extraction is a method that, given a trained neural network and the data on which it was trained, produces a description of the network’s hypothesis that is comprehensible yet closely approximates the network’s predictive behavior (Craven, 1996; Gilpin et al., 2018). Through rule extraction, a simpler model is built that mimics the behavior of the complex, deep neural network (Jacobsson & Ziemke, 2005). Rule extraction techniques can be categorized as decompositional, pedagogical, and eclectic. Decompositional techniques involve analyzing the weights between units and the activation function (“looks inside” the network) to extract rules. Pedagogical techniques treat the network as a black box and extract rules by examining the relationship between inputs and outputs. Eclectic approaches incorporate both decompositional and pedagogical techniques (Biswas et al., 2017). An example of an automatic rule extraction technique is the TREPAN algorithm, which derives a decision tree from a neural network (Craven & Shavlik, 1996).



### Counterfactual explanations

Wachter et al. (2017) argued that an explanation for an automated decision is in many cases insufficient, in particular for subjects of a decision-making process. Furthermore, they argue that local interpretability may be difficult to implement in practice. They propose a novel way of increasing understanding and trust in automated decision-making: counterfactual explanations. A counterfactual explanation gives the subject a view of how her situation would have to be different for a desirable outcome to occur. Multiple counterfactuals are possible, as multiple desirable outcomes can exist (Wachter et al., 2017). A counterfactual decision could look something like this:

*"You were denied a loan because your income (30k euros) was too low. If your income had been 50k euros, your loan would have been approved."*

This approach provides data subjects with meaningful explanations to understand a given decision, grounds to contest it, and advice on how they could change their behavior or situation to possibly receive a desired decision (e.g., loan approval).

### Disclosure (Article 7)

Finally, the actual disclosure of the elements of an automated decision-making system and the processes and procedures for creating the system is required. The relevant elements mentioned in the Policy Prototype include, but are not limited to:

- the rationale for automated decision-making;
- the training data, which may include description, origin, consent (when appropriate), type (personal, sensitive), update level, security mechanisms for its protection, and safeguarding;
- the machine learning algorithms used;
- the decision-making model;
- the process of data selection and preparation;
- the process of training, selecting, validating, and testing the model; and
- the process of managing the model in operation.

When it comes to disclosure, it is of particular importance what the disclosure goal is and what audience the content is disclosed to. Therefore, to meet the requirement of Article 8, a user must first establish what might be expected of them in terms of disclosure. Tables 1 and 2 should be used to assess what the disclosure requirement is and what the relevant audience is.

Generally speaking, disclosure is only relevant for ADM systems that can have a significant negative impact on the rights and freedoms of natural and legal persons. In these areas in particular, the user must be able to account for the results. A third party (e.g., an auditor or supervisory authority) may wish to assess the operation of the ADM and the processes for developing, deploying, and main-

taining it. In other words, a system operator should consider having a third party assess the operation in those particular cases. Therefore, the disclosure requirement ties in closely with requirements in the area of auditability.

## Process for increasing transparency and explainability

Based on our analysis of the literature regarding transparency and explainability, we suggest the following process for dealing with requirements related to improving the understanding of decision-making.

### **Step 1. Identify the risks associated with the decision-making process.**

Based on the outcomes of your risk assessment, determine what risks the decision-making process poses for individual or collective values.

### **Step 2: Determine whether global interpretability is a requirement.**

Based on this assessment, determine what level of information, disclosure, and interpretability is required. If the decision-making process has a significant negative impact and requires strong justification (e.g., because it affects the legal position of the subject), consider whether a globally interpretable model is required.

### **Step 3: Determine the goals for understanding and the associated target audience.**

Determine the goals of information provision and disclosure. Based on the different goals, determine which target audience must be addressed (see Table 2 for audience types).

**Step 4: Determine the appropriate disclosure for the target audience.**

For each target audience, determine what form of disclosure or explainability is required and desirable. When it comes to interpretability and explainability, tailor the explanation and message to the target audience’s goals and their capability to understand and assess the information that you provide them. Take into account factors such as expertise and time limitations.

- **Identifying and studying the audience**

It is essential to answer the question “Who is directly affected by the AI-based product or service?” so that communication and support materials are legible, comprehensible, and custom-made.

Similarly, the normative/legal and human contexts of different geographical regions should be taken into consideration, especially for global products or services, to avoid generic and empty communication.

- **Defining the level of detail for each audience**

Each group will likely have a different technical and digital literacy level based on educational level or personal interest. Therefore, it is important to define the amount of information and complexity that will be presented to end-users. It is important to agree on the information to be communicated (data used, model information, level of human involvement, situations in which AI is used, impacts of AI decisions or predictions, etc.) and how that information will be communicated, which in turn must be aligned with communication purposes and the ethical AI use principles.

**Step 5: Implement technical and organizational measures to increase understanding.**

Finally, implement the required technical and organizational measures. Choose methods for interpretability and explainability (see Annex 1) based on the impact of the decision making, the required type and level of transparency and explainability, and the relevant audience as described in this document.

Consult Tables 1 and 2 to identify potential issues and what audience is to be addressed, along with the different types of contexts, purpose, and content underlying the transparency and explainability requirement.

*Suggestions to complement and go beyond transparency and explainability toward ethical AI*

**Include the practice of ethics in the use of AI-based systems.**

Implement workshops, courses, and training on the ethical use of AI in the different areas of the company or institution involved in the design, development, implementation, maintenance, support, improvement, or any activity related to the product or service that relies on the use of AI. The entirety of the team, in a transversal manner, must be informed on the state of the art of responsible development and use of AI, as well as transparency and explainability best practices. Furthermore, consider making tools that comply with ethical standards so that the very development tools engineers use help enforce best practices.

**Preserve a culture based on the ethical use of AI.**

This process is not over once the steps described in this playbook are completed; it is a continuous process that allows us to preserve a culture that fosters ethical principles on the use of AI, which also responds to changes in the technology, improvements or changes to products or services, new business objectives, new user and market expectations, organizational changes, new laws and regulations, and social and cultural changes.

## Additional resources

### Section 1

#### Existing guides

- [Explaining decisions made with AI - UK's Information Commissioner's Office, 2019](#)
- [Companion to the Model AI Governance Framework: Implementation and Self-Assessment Guide for Organizations – World Economic Forum, Info-communications Media Development Authority of Singapore\(2020\)](#)
- [The NIST AI Risk Management Framework \(AI RMF\) and Playbook on interpretability and explainability.](#)

### Section 2

#### When your AI systems are provided by third parties

When the organization uses AI technology provided by third parties, it is recommended to obtain an explainability model from the supplier; companies offering ML platforms such as Facebook, Google, IBM, Amazon and Microsoft, are starting to include explainability tools.

### Section 3

#### Recommended tool

- [“People-centric Approaches to Notice, Consent, and Disclosure”](#) from TTC Labs and Singapore Infocomm Media Development Authority.

#### Helpful tools for implementing governance

- “Explaining Decisions with AI. Part 3: What explaining AI means for your organisation” UK’s Information Commissioner’s Office (2019)

#### Recommended courses

- Big Data, Artificial Intelligence and Ethics, University of California, Davis.
- Ethics and Big Data, The Linux Foundation.

#### Impact and risk assessment

This should include but not limited to:

- describing the nature, scope, context, and purposes of the process;
- evaluating necessity, proportionality, and compliance measures;
- identifying and assessing risk to people; and
- identifying any additional measures to mitigate those risks.

### **Recommendations to comply with the accountability principle**

- **Developing mechanisms, challenging processes, and decision reviews**

It is essential, especially for a human-in-the-loop system, that there exist processes and tools for end-users who feel unsatisfied with the decision made by the AI system to be able to go through a challenge process or to request a human reviewer to audit the process. In the case of a result with a significant negative impact on an individual's life, it is recommended that the final decision not be automated.

- **Communicating human review mechanisms and challenging decisions**

Users must be clear about how they can contest decisions made by AI that have disadvantaged them and that they believe to be a mistake. These mechanisms must always be easily accessible, visible, and present so that a user can decide whether to challenge a decision and request an explanation or human review.

- **Open feedback channels**

Finally, no user-focused design exercise would be complete without a continuous feedback process; users must have clear and always available mechanisms to issue comments, suggestions, or complaints about the performance or service provided by the AI-based system.

### **Existing impact analysis methodologies**

- Data Protection Impact Assessments (DPIAs), UK's Information Commissioner's Office
- Human rights impact assessments (HRIAs)
- Human Rights Due Diligence, United Nations
- "AI Impact Assessment: A Policy Prototyping Experiment" (2021), at [https://openloop.org/wp-content/uploads/2021/01/AI\\_Impact\\_Assessment\\_A\\_Policy\\_Prototyping\\_Experiment.pdf](https://openloop.org/wp-content/uploads/2021/01/AI_Impact_Assessment_A_Policy_Prototyping_Experiment.pdf)
- [ICO: AI and Data Protection Risk Mitigation and Management Toolkit](#)

### **Recommended technical guides (in addition to Annex 1)**

“Individual Explanations in Machine Learning Models: A Survey” (Alfredo Carrillo, Luis F. Cantú, & Alejandro Noriega, 2020).

Over the past few years, the use of sophisticated statistical models that influence decisions in areas of great social relevance has been increasing. In real-world applications, mainly in domains where decisions could have social impacts, model interpretation is desired. This survey reviews the most relevant and groundbreaking methods to tackle explaining individual instances problems in machine learning. In particular, it aims to provide a guide for data scientists on the research for appropriate methods to solve the need for explainability models.

“Individual Explanations in Machine Learning Models: Case Study” with Prosperia.ai (Alfredo Carrillo, Luis F. Cantú, & Alejandro Noriega, 2020).

The case study had two main objectives. The first was to expose the challenges technical teams face in the real world and how they use relevant and novel methods of explanation. The second was to present a set of strategies that mitigate such challenges.

### **Explainability techniques:**

Strategies and tools like surrogate models, partial dependence plots, global variable importance/interaction, sensitivity analysis, counterfactual explanations, or self-explaining and attention-based systems are examples of explainability techniques. Adding documentation on how an AI system was built, trained, and tested will help improve its explainability. The technical team might also consider building predictive models that mimic real conditions or train simpler versions of the model (e.g., linear regressions or decision trees instead of neural networks) to simplify the explanation. Additionally, these options could prove useful:

- a** using visualizations to explain predictions or individual decisions;
- b** explaining the features or weights of each input; and
- c** taking into account the research and methods developed by third parties in academic and scientific fields that allow for the analysis of AI models from multiple perspectives—for example, by scope (global or local), by generality (model-agnostic or model-specific), or by level of interpretability (post-hoc or ad-hoc—specifically, intrinsically interpretable methods).

### **Design techniques (user interfaces and experiences)<sup>2</sup>**

It is important to remember that the end-user will always interact with the AI through some kind of interface (be it graphical, audio-based, or some other type). Thus, it is important to create a user experience that not only is user-friendly and intuitive but also adequately and elegantly communi-

cates all of the information, mechanisms, and processes that come with using AI. Because of this, it is recommended to use co-design methodologies and user-focused design to maximize the quality of the end result and follow these design principles:

- a Transparency is not opposed to simplicity; it is possible to communicate generously provided that it is done at the proper time and place, meaning in a progressive way throughout the journey of ADM use.
- b An organization should be aware of the different levels of technical literacy on ADM and adapt accordingly. When users deliver data, they should perceive it as a value exchange so that the relationship between the input of their data and the outcomes from that data input are understood.
- c Trust is built by being proactive, contextual, and transparent, allowing people to understand at all times how their data is being used, what their privacy options are, and how and where to update them.
- d Transparency alone is not enough but should be accompanied by “controls” (where applicable) that allow end-users to make decisions about ADM, the use of their data, and give them the option to—depending on the level of impact of the ADM in question—challenge, appeal, or opt out of ADM.

## Additional recommendations:

### ***Communicate and control AI retraining***

In cases where the end-user provides new data or performs actions that feed a constant training process for the AI, it is critical to communicate with end-users: on the one hand, from the point of view of data usage, and on the other, from the point of view of AI safety—that is, to keep the AI system safe (to the extent possible) from malicious users and contents. Because of this, organizations must consider covering these topics in their terms and conditions of use.

One of the key tasks of this process is converting the technical complexity that entails understanding how an AI-based system works, and achieving clear and concise communication. It is recommended that the contents used for this include clear and specific benefits<sup>3</sup> to end-users from using an AI-supported system, as well as the potential negative consequences.

## Possible tensions that may emerge from the implementation of this playbook

The implementation of transparency and explainability (T&E) practices also involves an ongoing process to balance its opportunities and risks (i.e., tensions) for the organization, for example:

- 1 T&E vs. enabling bad actors to act more effectively, game the system, and manipulate algorithms for their own purposes.
- 2 T&E vs. effectiveness/accuracy: Modern AI methods, especially non-inherently interpretable deep learning, may—in some cases—become more effective but harder to understand.



- 3 T&E vs. disclosure of potential IP issues: Full transparency of algorithms, namely disclosure of source code, raises important legal problems from intellectual property and trade secrecy perspectives, just like the disclosure of other types of proprietary information (e.g., software, patents).
- 4 T&E vs. meaningfulness/actual understanding: Overly detailed T&E may not be meaningful to users and may not advance their understanding of how their data is being handled or how decisions/recommendations/predictions are made.
- 5 T&E vs. multiple actors: Disclosure obligations are very different from those of developers and users.
- 6 T&E vs. data protection: Modern AI may act against some of the principles of data protection (minimization, purpose specification, etc.).



# Endnotes

1. The term automated decision-making (ADM) is understood and used in different ways by engineers and lawmakers. In addition, the term ADM is used in different ways according to the specific legislation in which it is employed. While engineers use the term broadly, describing it as a system that leverages machine learning to produce an outcome (decision), policymakers and regulators use the term in a more granular manner, narrowing it to specific types of outcomes (decisions) that observe narrowly defined criteria. The narrow legal understanding of ADM can mostly be found in privacy regulation. The broader legal understanding of ADM is sometimes used in legislation that elsewhere would be called "AI legislation." Note: For the purposes of this policy prototyping exercise, and regarding the actual testing of the prototype, the term ADM is used in the broadest sense, detached from any qualifying criteria. The focus was on the operationalization of transparency and explainability concepts into concrete practices, and not on the specific type, relevance, or impact of the decisions produced or supported by ADM systems.
2. See e.g. the basis for a National AI Strategy in Mexico developed by C Minds, Oxford Insights and the British Embassy ([English - Spanish](#)); the Mexican National AI Agenda developed by IA2030Mx ([English - Spanish](#)), among other resources.
3. Such as recommending that AI developers provide documentation about how their systems work, among other T&E practices.
4. See e.g. the NIST Risk Management Framework (RMF) created by the Computer Security Resource Center "provides a comprehensive, flexible, repeatable, and measurable 7-step process that any organization can use to manage information security and privacy risk for organizations and systems" More information: <https://www.nist.gov/itl/ai-risk-management-framework>
5. Reyes, E. (2020). Las empresas mexicanas no saben qué hacer con la Inteligencia Artificial [Mexican companies do not know what to do with Artificial Intelligence]. Expansión. Retrieved from <https://expansion.mx/tecnologia/2020/07/30/las-empresas-mexicanas-no-saben-que-hacer-con-la-inteligencia-artificia>
6. United Nations Educational, Scientific and Cultural Organization (UNESCO). (2022). Recomendación sobre la ética de la inteligencia artificial [Recommendation on ethics of artificial intelligence]. Retrieved from [https://unesdoc.unesco.org/ark:/48223/pf0000381137\\_spa](https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa)
7. Organization for Economic Cooperation and Development (OECD). (2019). Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments. Retrieved from <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
8. European Parliament. (2023). DRAFT Compromise Amendments on the Draft Report Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. European Parliament. Retrieved from: <https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf>
9. Institute of Electrical and Electronic Engineers. (2022). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Retrieved from <https://standards.ieee.org/industry-connections/ec/autonomous-systems/>
10. United Nations Educational, Scientific and Cultural Organization (UNESCO). (2022). Recomendación sobre la ética de la inteligencia artificial [Recommendation on ethics of artificial intelligence]. Retrieved from [https://unesdoc.unesco.org/ark:/48223/pf0000381137\\_spa](https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa)
11. Independent High-Level Expert Group on AI. (2019). Ethics guidelines for trustworthy AI. European Commission. Retrieved from <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
12. Supra note 7

13. Kilpatrick, D. (2000). Definitions of public policy and the law. National Violence Against Women Prevention Research Center, Medical University of South Carolina. Retrieved from <https://mainweb-v.musc.edu/vawprevention/policy/definition.shtml>
14. World Economic Forum. (2018). Agile governance – Reimagining policy-making in the Fourth Industrial Revolution. Retrieved from [https://www3.weforum.org/docs/WEF\\_Agile\\_Governance\\_Reimagining\\_Policy-making\\_4IR\\_report.pdf](https://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf)
15. Design thinking is a process used to solve problems by prioritizing consumers’ needs. It uses tests that determine how consumers relate to products or services, as well as a repeating and practical approach to create innovative solutions. For more information, visit <https://designthinking.ideo.com/>.
16. Brown, T., & Katz, B. (2011). Change by design. *Journal of Product Innovation Management*, 28(3), 381-383.; Villa Alvarez, D., Auricchio, V., & Mortati, M. (2020). Design prototype for policymaking. Retrieved from <https://dl.designresearchsociety.org/cgi/viewcontent.cgi?article=1168&context=drs-conference-papers>; Kontschieder, V. (2018). Prototype in policy: What for?! (2018). Retrieved from <https://conferences.law.stanford.edu/prototype-for-policy/2018/10/22/prototype-in-policy-what-for/>
17. Buchanan, C. (2018). Prototype for policy. UK Government. Retrieved from <https://openpolicy.blog.gov.uk/2018/11/27/prototype-for-policy/>
18. Hébert, M. (2019). A pilot is not a prototype: How to test policy ideas before scaling. *Apolitical*. (2019). Retrieved from <https://apolitical.co/solution-articles/en/a-pilot-is-not-a-prototype-how-to-test-policy-ideas-before-scaling>; see also [https://openloop.org/reports/2022/12/Experimental\\_governance\\_emerging\\_technologies\\_Chapter1.pdf](https://openloop.org/reports/2022/12/Experimental_governance_emerging_technologies_Chapter1.pdf).
19. Kontschieder, V. (2018). Prototype in policy: What for?! (2018). Retrieved from <https://conferences.law.stanford.edu/prototype-for-policy/2018/10/22/prototype-in-policy-what-for/>
20. In this case, we measure clarity (to what extent do those who are subject to the requirements of the policy understand what is required of them?), efficiency (how much does the public policy prototype help achieve the overall policy objective?), and viability (how feasible are the requirements?). For more information on these definitions, visit [https://openloop.org/wp-content/uploads/2021/01/AI\\_Impact\\_Assessment\\_A\\_Policy\\_Prototype\\_Experiment.pdf](https://openloop.org/wp-content/uploads/2021/01/AI_Impact_Assessment_A_Policy_Prototype_Experiment.pdf).
21. A multi-sector coalition composed of professionals, academic institutions, companies, startups, public agencies, and other digital and AI ecosystem players in Mexico with the purpose of developing a framework adequate to promote AI recommendations and best practices. See Hernández, M. (2019). *Estrategia Nacional de Inteligencia Artificial va por sentido ético y responsable* [National Artificial Intelligence Strategy based on ethics and responsibility]. *Forbes México*. Retrieved from <https://www.forbes.com.mx/estrategia-nacional-de-inteligencia-artificial-va-por-sentido-etico-y-responsable/>
22. Prior to the construction of a Comprehensive and Inclusive artificial intelligence agenda for Mexico, the IA2030Mx Coalition launched a National Artificial Intelligence Consultation to better understand the main digital transformation challenges in Mexico and the different proposals in favour of the AI revolution for the benefit of all. For more information, visit <https://www.ia2030.mx/consulta>.
23. Independent High-Level Expert Group on AI. (2019). Ethics guidelines for trustworthy AI. European Commission. Retrieved from <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
24. The training data are those used to train a model. In ML systems, algorithms learn from data they are fed.
25. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Sri Kumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication No. 2020-1. Retrieved from <http://dx.doi.org/10.2139/ssrn.3518482>

26. Ad Hoc Expert Group (AHEG) for the Preparation of a Draft text of a Recommendation the Ethics of Artificial Intelligence. (2020). Outcome document: First draft of the recommendation on the ethics of artificial intelligence. UNESCO. Retrieved from <https://unesdoc.unesco.org/ark:/48223/pf0000373434>
27. Organization for Economic Cooperation and Development (OECD). (n.d.). Transparency and explainability (Principle 1.3). AI Policy Observatory. Retrieved from <https://oecd.ai/dashboards/ai-principles/P7>
28. See Interpretable AI. (2022, September). What is Interpretability?
29. See also TTC Labs. People-Centric Approaches to Algorithmic Explainability, available at <https://www.ttclabs.net/report/people-centric-approaches-to-algorithmic-explainability>
30. A process that ensures that those at risk of poverty and social exclusion have the opportunities and resources necessary to participate fully in life (OAS, 2019).
31. Dscout is a mobile qualitative research ethnographic platform that can be used to obtain information on users' experiences. For more information, visit: <https://help.dscout.com/hc/en-us/articles/360038171372-dscout-Overview-for-Researchers/>.
32. For public and context selection, only one could be chosen, however, more than one option could be chosen for purpose and contents.
33. Among the biggest companies we consulted, the main reasons provided for not participating in the program were 1) the considerable investment of resources and time that would be required and 2) the lack of existing regulation that mandated strengthening the T&E of their AI/ADM systems.
34. Instituto Nacional de Estadística y Geografía (INEGI). (2019). Micro, pequeña, mediana y gran empresa: Estratificación de los establecimientos. [Micro, small and medium-size and large companies: Stratification of institutions]. Economic Census, p. 20.
35. Supra Note 2
36. Supra Note 3
37. Supra Note 4