


POLICY RECOMMENDATIONS

Preventing and mitigating gender bias in AI-based early alert systems for education





The policy recommendations were developed by Itad, PIT Policy Lab, Women in Digital Transformation, and Athena Infonomics in the context of the Equitable AI initiative. The project was enabled by the support of the United States Agency for International Development (USAID). The contents are the responsibility of the aforementioned organizations and do not necessarily reflect the views of USAID or the United States Government.

Publication date: July, 2023

TABLE OF CONTENTS

Overview of Use Case: Guanajuato, Mexico	1
Introduction	2
Recommendations	
Phase 1: Self assessment & reflection	3
Phase 2: Paving the way	4
Phase 3: Involvement and transparency	5
Phase 4: Strengthening existing instruments	6
Phase 5: Broader sensitivity	7

OVERVIEW OF USE CASE: GUANAJUATO, MEXICO

Under the current administration (2018-2024), the Ministry of Education of Guanajuato (MEG) has been working to ensure valid, reliable, and up-to-date information about the state of the local education system to inform decision making. Thus, they have been developing and harmonizing databases with information regarding student attainment and overall school performance under the initiative **Educational Trajectories**. In addition, to identify students at risk of dropping out, the Ministry has partnered with the World Bank to create an AI based early alert system, the **Early Action System for School Permanence (SATPE**, by its acronym in Spanish: Sistema de Actuación Temprana para la Permanencia Escolar), which aims to provide at-risk students with support and improve retention and graduation rates through focalized preventive actions.

Educational Trajectories brings together internal and external operational data sets, aiming to be an instrument for the monitoring of students, as well as to guide tailored actions from the Ministry. This consolidated data framework allows for the identification of a set of variables that have an impact on student retention and on the quality of education, allowing for powerful and effective interventions. It has also made it possible to develop predictive analysis on the individual trajectories of students in Guanajuato, for a close follow-up together with schools and parents.

The Educational Trajectories initiative serves as the foundation for the SATPE. The development of an AI-driven early alert system to identify students at risk of dropping out arose from a former collaboration with the World Bank, to leverage existing data from the Educational Trajectories initiative within the vision of the State's "Pact for Education".



INTRODUCTION

Both the Educational Trajectories and the Early Action System for School Permanence (SATPE) projects are set to be key to informing Guanajuato's educational policy. At the beginning of 2023, the first list of students at risk of dropping out was identified leveraging data analysis and the Ministry of Education of Guanajuato (MEG) team is evaluating actionable frameworks for piloted interventions. At a technical level, the last details for the deployment of the information system are being adjusted, as well as the mechanisms to feed it with more records within the framework of Educational Trajectories initiative. As demonstrated throughout the Equitable AI Use Case, there have been positive iterations to the SATPE design and in the policy interventions it aims to inform. However, there are some critical aspects that need to be further considered and addressed:



ROLE OF COMMUNICATION

towards the main impacted parties (students, parents, teachers) regarding the use of AI for decision-making in the educational field, prior to the implementation and deployment of the SATPE.



DEEPEN RISKS CONCEPTION

associated with the use of AI, beyond the potential misuse of personal data (for example, biases in the data and the potential of these systems to reproduce and expand them).



PUBLIC CONSULTATION

Around the design and deployment of an AI-driven public policy, and/or feedback mechanisms with different stakeholders.

These challenges are common to governments beginning to implement AI projects and developing internal ethical and regulatory frameworks for such initiatives. In addition, they reflect areas of opportunity shared by subnational governments in Mexico, the Latin American region, and the Global South.

With the intention of promoting practices aiming towards a more equitable and responsible use of AI, this report presents a series of actionable policy recommendations to mitigate biases and to mainstream gender perspective within AI projects. The policy recommendations are tailored to the Ministry of Education of Guanajuato's SATPE and Educational Trajectories projects, and were thought as a guide towards any other future project they develop that involves the use of data and AI. Nevertheless, these recommendations are general enough to be adopted by other local governments that are also beginning to explore policies that use data and AI, including initiatives outside the education sector.

RECOMMENDATIONS

Phase 1: Self assessment & reflection

Throughout the development of the Equitable AI initiative, many questions arose from the MEG team around how to capture aspects that could have gender implications in the data beyond the disaggregation by sex, or if there are other realities not captured in the information. The scope of planned interventions takes on new implications when human rights, ethics, and other design criteria are included. This discussion is not idle and makes it possible to identify areas of opportunity, as well as challenges yet to be resolved.

The following is suggested:



Document the processes and learnings from the design and deployment of current initiatives, particularly related to risks posed by introducing an AI-based system, identifying the relevant questions that have been raised along the development of the initiative and how they could be resolved or answered in different time horizons.



Weigh strengths and weaknesses in an analysis of institutional capacities, especially in relation to ethical and responsible use of technology, including the different teams and areas involved in projects with data and AI, to identify specific needs on issues such as data governance, cybersecurity, gender, and ethics of AI.



Develop a risk analysis methodology, or adapt an existing one, that includes some of the aspects addressed during some of the workshops conducted within the project (Human Rights, gender, AI ethics, etc.), involving additional key stakeholders (teachers, students, parents, civil society, other Ministries).



From the design stage of interventions, teams should ask themselves what could possibly go wrong, mapping potential stakeholders involved to take their concerns in consideration. **The Self-Assessment AI Checklist published within the Equitable AI initiative can be a useful input for the identification of risks and areas of opportunity;** however, it is not a substitute for consultation with the people impacted by the intervention.

PHASE 2: PAVING THE WAY

In order to standardize best practices when working with data and AI, the area/organization of the education sector involved in a use case must establish decision criteria that aligns with ethical and Human Rights practices. Based on the questions raised and systematized in the previous step, this stage suggests that clear guidelines should be established to guide action. In addition, to address concerns about the management of private information within and outside the Ministry, as well as to mitigate potential risks of misuse of data, cybersecurity and personal data protection criteria must also be taken into account.

The following is suggested:



Set or adopt internal principles to guide action within the Education Ministry in line with the principles that the country has already pledged towards. Create a working group that is in charge of systematizing and designing the ethical criteria for interventions with data and AI within the area/organization. **The AI Ethics Guide published within the Equitable AI initiative can be a useful input in this regard, as it provides examples of such practices.**



Create a data governance protocol that allows establishing rules and criteria on how information is stored, consumed and used within the Ministry. Such a protocol must establish different levels of data security and rules for its transfer within and outside the organization. It must also contain criteria for the quality of the data, ensuring that there is complete, accurate, and representative information for the population that the use cases addresses. This should also consider criteria for interoperability between other Ministries of Guanajuato's Government.

PHASE 3: INVOLVEMENT AND TRANSPARENCY

Interested stakeholders, such as parents, students, community leaders, educational authorities, and other key actores must be included from the design and implementation phases within the Responsible Technology Cycle. Their role goes beyond an initial discussion of possible risks, as to be considered in a pilot stage and to be able to provide feedback throughout the life cycle of the project. In addition, the right to audit the results of the model must be considered within the framework of the attributions of each party involved, as well as a mechanism for the explainability of the results.



Create consultation mechanisms open to civil society and other specific interest groups (both large companies and SMEs, citizens, and professional bodies). From the design stage, they must ensure that the parties involved understand how a decision impacts them and to what extent it is based on the predictions of an AI system. This broad discussion should be encouraged whenever there are new uses of data with potential implications for the well-being of certain groups or the population in general.



Design an explainability flow, which allows end users and impacted groups to understand the reason for AI's decision-making, from the identification of students at risk who will require support, to the implementation of interventions. Clearly delineate to what extent the model suggests each component of the specific action and what is the reasoning for the policy intervention. This is important to identify where a certain course of action originates from, when or not it is an automated recommendation, and who has implemented a control or verification measure on such a decision.



Design a clear communication strategy for the parties impacted by the SATPE (students, teachers, school administrators, parents, community leaders, other interested stakeholders). This should prioritize understanding the strategy in the context of the target audience; for example, boys and girls, students from different socioeconomic levels, rural and urban population, etc. In the design of the communication strategy, the Ministry should question which discourses can be stigmatizing (for example, "risk of dropping out of school" versus "risk of interrupting the educational path") and how it complements the inclusion strategy of the interventions.



Offer awareness campaigns and incentives for the ethical use of data and AI for SMEs, data scientists, and start-ups involved in the provision of services; as well as training to consider such aspects in procurement, both for potential/actual suppliers and for government personnel.



Other actions include: These projects are frequently an iterative process that can benefit from revisiting the design and implementation. The ability to engage in national and international discussions, share learnings widely, and to conduct peer to peer regional, national, state, and/or city level knowledge exchange can ensure a more agile approach to programs that can change and adapt as new evidence becomes available.

PHASE 4: STRENGTHENING EXISTING INSTRUMENTS

There are some precedents that guide different practices of the Ministry in relation to specific aspects, such as ethics in the public service or the protection of personal data (for example, in Mexico there is the Federal Law on Personal Data Held by Private Parties). Being able to coordinate existing efforts and to incorporate an ethical perspective of AI and the protection of personal data as criteria allows us to build a solid foundation for ethical, responsible, and trustworthy policies powered by AI.

The following is suggested:



Include basic guidelines for the protection of personal data, AI ethics and Human Rights in the existing codes of conduct for public officials.



Consolidate an Ethics Committee to establish and monitor compliance with mechanisms to ensure the deployment of an ethical and responsible AI in all the interventions conducted by the Ministry. This committee may be linked to in house Innovation policies or teams (for instance, the MEG-UNESCO's Educational Innovation Laboratory).



Promote transversal AI literacy within the Ministry and other State Government agencies in order to generate specific capacities in decision makers who will interact with AI systems; for instance, people directly interacting with the results of the SATPE to ensure that their predictions are correctly interpreted.

PHASE 5: BROADER SENSITIVITY

The SATPE is a first exercise that is creating awareness around ethical implications and risks associated with the use of AI for decision-making in public policy. Increasing the precision of diagnosis and the effectiveness of interventions can only be done by measuring and evaluating the results. Contemplating verification mechanisms beyond the data and the algorithm is essential to ensure the human component as ultimate decision-makers.

The following is suggested:



Contrast the abandonment risk alerts with alternative indicators and qualitative information about the project's impact. A case analysis exercise on a representative random sample can shed light on what other elements of the context complement or contradict the predictions of the SATPE.



Consider data fairness metrics, such as those provided by the **AI Fairness 360 toolkit**, to identify when biases may be present in the data and correct for them, before classification/prediction.[1] Revisit these metrics often during the use of AI systems, as these can change over time as new data is added to the system.



Develop broad sensitivity to calibrate AI-enabled decision making processes. It will also depend on how sensitive the results and interventions are to stakeholder feedback. For example, whether or not to incorporate mechanisms to include additional information from identified students, teachers or parents.



Generate adequate metrics. Although the measurement and evaluation of results is part of the project's life-cycle at the Ministry, the use of appropriate schemes fit to the reality of the phenomenon is a priority. The effectiveness of the results will have to be weighted with equity and inclusion criteria, and they must be able to separate the results of the prediction (precision) and the interventions (effectiveness) in the analysis.

[1] Other tools and resources available are: Google's People + AI Research (PAIR), Carnegie Mellon University's Aequitas audit and bias toolkit, Fairlearn, Microsoft's Datasheets for Datasets, and Tensorflow's Model Card.

POLICY RECOMMENDATIONS

Preventing and mitigating gender bias in AI-based early alert systems for education

