FRAMEWORK FOR FUNDAMENTAL RIGHTS RISK ASSESSMENT AND TRANSPARENCY EVALUATION

USE OF ARTIFICIAL INTELLIGENCE BY PUBLIC AUTHORITIES

TRANSPARÊNCIA BRASIL

BY

CHIEF EXECUTIVE
Manoel Galdino

CHIEF OPERATING OFFICER
Juliana Sakai

PROJECT COORDINATOR
Tamara Burg

RESEARCH
Jonas Coelho
Tamara Burg

TEXT
Juliana Sakai
Manoel Galdino
Tamara Burg

TEXT FORMATTING
Marina Atoji

www.transparencia.org.br

FUNDING

PARTNERSHIP
Northwestern University

ACKNOWLEDGEMENT
Controladoria-Geral da União (CGU)
Ministério da Ciência, Tecnologia e Inovação (MCTI)
Centro de Estudos sobre Tecnologias Web (Ceweb.br)
Bruno Kunzler
Enrico Roberto (InternetLab)
Nathalie Fragoso (internetLab)
Nazareno Andrade

TABLE OF CONTENTS

INTRODUCTION

This document presents a proposed framework for risk assessment in the use Artificial Intelligence (AI) algorithms by the government, seeking a convergence between innovation, technology promotion and public accountability, transparency.

We aim to present a simplified approach that allows for risk assessment concerning real and possible threats to rights and public spaces. The assessment results can make it possible for the government to understand the necessary transparency in AI development, acquisition and implementation procedures so that the proper democratic social accountability is guaranteed. For this reason, the framework includes risk assessment on fundamental rights violations and a separate transparency evaluation.

The risk assessment framework is organized in four tracks: i) risks related to the nature of the tool; ii) algorithmic discrimination risks; iii) privacy rights violation risks; and iv) potential abuse of authority in civic spaces.

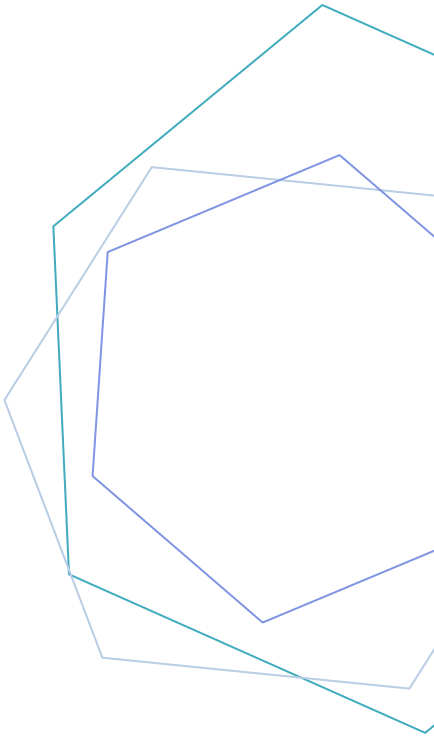
The transparency assessment framework is described in a fifth chapter (item v). In this context, we do not consider transparency as a threatened right, rather as an essential tool to ensure public oversight and accountability. Therefore, the existing levels of transparency are assessed as an essential tool for following up on the full procedures of the use of AI by public sector and accounting for the risks they might pose to fundamental rights.

This document offers guidance on assessing the use of AI by government agencies as a mean to identify possible threats to fundamental rights and civic engagement. The framework provides the means for elaborating and proposing specific recommendations for the use of AI tools by government and governance, as well as for general guidelines on AI development, acquisition and implementation.

A multisectoral analysis, covering a wide range of fields, is an important requirement for the risk assessment procedure. Thus, contributions from several civil society organizations are fundamental in applying this framework. Given they act on promoting distinct causes and defending different rights, their participation greatly enriches the analysis.

RISK ASSESSMENT FRAMEWORK

1. Tool type-related risk assessment



The aim is to assess the possible impact of certain AI tools on fundamental rights in specific cases, based on their output or on their results – that is, based on what a specific AI tool has been designed to deliver.

AI algorithms may be used by governments for different goals. Some of these tools are designed to accelerate internal management procedures, as natural language processing algorithms that help with automated screening of documents. Others may be used to estimate the chance of criminal recidivism of convicted felons, impacting on future guilty verdicts[1].

Regarding the first type of tool, any mistake or system malfunction would only lead to a loss of agility in the workflow, as the task would have to be reassigned to a human – not a serious risk to fundamental rights, then.

In the second case, however, mistakes or malfunctions may have critical impacts: wrongful results might compromise civil rights and liberties, contribute to stigmatization of already marginalized groups and violate the very premises and legal requirements of due process and the principle of fairness. Given

[1] As ProPublica revealed, the COMPAS system, used in the US for criminal recidivism risk assessment https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

its own nature, this type of AI tool poses serious risks to fundamental rights.

To address and mitigate these risks, a proper risk assessment should take place early on, preferably during design and implementation. Adopting follow-up procedures on their operation and results is indispensable.

The minimum requirements for a proper risk assessment are:

Frameworks for tool type-related risk assessment

FRAMEWORK 1. GENERAL DATA ON THE TOOL

Nr	INFORMATION TO BE REGISTERED
I	Government agency
II	Tool name
III	Usage category *Image classification (except facial recognition), facial recognition, recommendation systems, chatbot, risk estimation (including fraud detection), sentiment analysis, and others.*
IV	Statistical model *Logistic regression, linear regression or variations, methods based on decision trees (including random forests and XGBoost), neural networks, natural language processing, AutoML, others.*
V	*Inputs* *Description of input variables*
VI	*Output* *For example, probability of a certain case being fraudulent; or credit approval or denial.*
VII	Level of support provided by the tool *Diagnosis and decision making; diagnosis and action recommendation; diagnosis not including action recommendation.*

FRAMEWORK 2. ASSESSMENT CHECKLIST

Nr	ASSESSMENT TO BE PERFORMED
I	When using the tool, is there human supervision for all decisions recommended or made by the algorithm?
II	In case of algorithm error corrected by a human, is this information used to improve the algorithm?
III	Considering its nature, can the tool directly or indirectly impact on fundamental rights, either due to mistakes or to the algorithm design itself? If so, on which rights?
IV	Which groups or populations will be affected by the algorithm? Have these segments been considered during the machine learning procedures?
V	Are there civil servants in the agency who are able to provide information on its use to competent authorities?
VI	Is the estimated negative impact created or intensified by the algorithm?
VII	Is this algorithm essential to reach the elected goal? If it has the ability to affect fundamental human rights or intervene on accessing them, are there any alternatives for the exercise of such rights? If so, which ones?
VIII	Is there evidence that this algorithm will work in the environment it is being used? Are the evidence based on relevant scientific experiments?
IX	Is there specific regulation on the usage of this algorithm in this specific area? Which is it? If not, is there a specialized legal team to ensure legal support?
X	Is there a technical team that follows-up on and monitors the implementation of this algorithm? Does this team include public agency's civil servants that are able to carry out critical analysis over the choices made by the tool?

CONTINUES

CONTINUATION

Nr	ASSESSMENT TO BE PERFORMED
XI	Does the AI development team include specialists in the field in which the algorithm will be used?
XII	Does an ethics committee oversee or has overseen the algorithm development and the processes of data collection and usage?
	ASSESSMENT: HIGH, MODERATE OR LOW IMPACT

2. Algorithmic discrimination risk assessment

Algorithm discrimination commonly arises from insufficiently representative training databases. Given the models are trained from available data, databases that do not comprise proportionally all affected groups tend to generate discriminatory models. Thus, the representativeness of a database will impact directly on an algorithm's results. Facial recognition algorithms trained from a database containing mostly images of white people, for instance, will be much less accurate in recognizing faces that deviate from this standard, such as black or brown people's faces[2].

As this algorithmic bias might reflect the lack of representativeness for certain groups, a tool could further emphasize social inequality and the oppression which marginalized groups are subjected to[3].

Apart from data representativeness, bias might also arise from data validity or from the algorithm design itself[4]. The available data used in training a model may not be the most adequate choice for a specific design, and its results might end up generating further discrimination among groups.

A study published in Science magazine[5] showed the example of an algorithm used in a hospital in the United States designed to guide medical decisions and sort patients by estimating which of

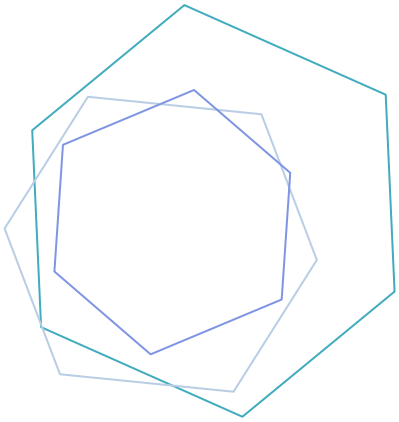
2 https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/

3 InternetLab's contributions to the Brazilian National Strategy for Artificial Intelligence: https://www.internetlab.org.br/pt/privacidade-e-vigilancia/as-contribuicoes-do-internetlab-para-a-estrategia-nacional-de-inteligencia-artificial/

4 https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

5 https://science.sciencemag.org/content/366/6464/447

them needed more urgent care. The algorithm favored white patients, wrongfully concluding that black patients were healthier than the white ones. This bias was a result of the usage of health insurance bills data as a proxy for the algorithm to assess the patients' medical condition, overlooking the fact that white people have more access to health insurance than black people.



Several activists, journalists, researchers and tech employees have been warning about the dangers of bias in AI systems for at least one decade. They have been doing thorough and rigorous researching efforts and have detected, proved and exposed the existence of algorithm discrimination in facial recognition, social media advertisement targeting, credit granting procedures, pension plan systems, and in algorithms used in criminal sentencing[6].

Furthermore, certain social classes and affluent regions tend to participate more actively in the technological community, influencing the design of automated classification and recommendation models, which might imply a certain degree of self-selection bias.

This risk assessment requires answering the following questions:

[6] AI Now Report, 2019. Available at https://ainowinstitute.org/discriminatingsystems.pdf

Framework for algorithmic discrimination risk assessment

Nr	ASSESSMENT TO BE PERFORMED
I	Have the possible biases on the tool's performance been considered during its development, acquisition and/or implementation?
II	If so, were biases corrected or mitigated by the code? How?
III	Have tests been run before and during implementation to estimate if error rates are the same or lower for minority groups?
IV	Is the training data sample diverse and representative enough to ensure good results for the different groups the tool is used on?
V	If the tool has not been developed internally, has it been designed specifically for the Brazilian people or for its target audience? If not, has its accuracy been tested for these specific groups?
VI	Is there a team that periodically monitors and evaluates the algorithm's performance in relation to these groups? If so, have retraining routines been planned during the algorithm's implementation?
VII	In case of tools that interact with the external audiences, such as chatbots, is there anyone responsible for handling complaints about algorithm discrimination?
ASSESSMENT: HIGH, MODERATE OR LOW IMPACT	

3. Violation of privacy rights risk assessment

Privacy violations might arise from the creation and/or availability of massive databases, since developing and employing AI technologies demand a great amount of data processing.

Government agencies have been using massive monitoring technological devices capable of collecting personal data from citizens, such as mobile phone tracking to enforce the lockdown imposed by COVID[7], and even tools for facial recognition[8].

Algorithms that aim at collecting strategic information and improving public services and policies are also on the rise. For example, the city of San Diego has installed thousands of cameras in street light poles, in an effort to study traffic conditions. Although the collected data have been proven unhelpful to improve traffic conditions, the police use these images without oversight nor accountability[9].

Discussions on the limits for personal data use requires a better understanding of what are personal data. The Law 13.709/2018, called Lei Geral de Proteção de Dados no Brasil (LGPD, General Law on Data Protection) defines personal data as all the information related to a natural person, given they are identified or identifiable. This includes personal traits, personal qualification, genetic data and so on.

7 https://www.bbc.com/portuguese/brasil-52357879
8 https://theintercept.com/2020/02/11/metro-sao-paulo-reconhecimento-facial/
9 https://www.latimes.com/california/story/2019-08-05/san-diego-police-ramp-up-use-of-streetlamp-cameras-to-crack-cases-privacy-groups-raise-concerns

LGPD prescribes that data processing and sharing by the public administration is restricted to necessary data for public policy making and implementation (art. 7º, III). This means that the data collection must be done strictly for the provision or improvement of specific public goods or services – having adequate and well-defined purposes and criteria.

For that matter, the public administration must always be transparent regarding to collecting data from identifiable or identified individuals, as well as on data sharing and usage.

This risk assessment requires answering the following questions:

Framework for violation to privacy rights risk assessment

Nr	ASSESSMENT TO BE PERFORMED
I	Is it necessary to use personal data to train the tool?
II	If so, are people notified that their personal data is being used to train the tool?
III	Is it necessary to use sensitive or confidential data to train the tool? If so, what is the legal basis and what additional security layers are used to protect this data?
IV	Can people choose to remove their data from the tool training or exercise other rights over their personal data, such as data access, portability, etc.?
V	Are the personal data collected/produced by the government agency or are they outsourced? Are these other sources legally permitted to share these data with the government agency?
VI	Are the personal data shared with third parties? Is there a legal basis for this sharing? If so, have data anonymization techniques been used prior to sharing?
VII	If personal/confidential data are used by the algorithm: have anonymization/pseudonymization techniques been used during the data pre-processing?
VIII	If personal/confidential data are used for the algorithm training: have privacy protection techniques been used during the model training?

CONTINUES

CONTINUATION

Nr	ASSESSMENT
IX	If the tool has been developed by a third party: is there any document that regulates the data use and sharing by this entity?
X	Can citizens choose not to have their data analyzed by an algorithm?
XI	Have people expressly authorized the use of their data for training this algorithm?
XII	Are the training data and the data collected by this algorithm stored in a cloud server (either in Brazil or abroad) or in a local server? Are there security protocols developed for accessing these data?
ASSESSMENT: HIGH, MODERATE OR LOW IMPACT	

4. Assessment of potential abuse of authority and civic space restrains

AI tools that collect and cross-reference personal information may be useful to some public policies – notably, public security ones –, but they might also represent a threat to civil society and become a powerful weapon in the hands of authoritarian governments. These governments might use such data to implement a surveillance state, chasing opponents and restraining civic space.

Sensitive data are information about personal aspects of the data owner's life that have a higher potential of being abused in this context. For instance, they might be used to persecute minorities or political opponents morally and politically.

Due to their nature, they are granted special protections under the law. LGPD defines sensitive data as 'personal data about racial or ethnic origin; religious beliefs; political opinion; trade union affiliation; religious, philosophical or political affiliation; data on health or sexual practices; genetic or biometric data, when linked to a natural person (art. 5°, II). To this category of personal data, LGPD provides a higher level of protection and establishes even more restrictions for handling them.

The goal of this section is to evaluate if a given AI tool brings risks to civil rights and liberties, and the civic space itself, due to possible authoritarian use.

This risk assessment requires answering the following questions:

Framework for civic space risk assessment

Nr	ASSESSMENT TO BE PERFORMED
I	Does the AI tool create or collect information that could be used to monitor individuals or political, ethnic or religious groups, as well as activists? If so, which tools are employed to avoid this excessive data use?
II	Does the algorithm use sensitive or potentially discriminatory data? If so, what are the additional security layers used to protect these data?
ASSESSMENT: HIGH, MODERATE OR LOW IMPACT	

TRANSPARENCY EVALUATION FRAMEWORK

Transparency is essential to enable risk assessments and, therefore, to defend fundamental rights and the civic space.

Thus, the usage of AI systems by the state must comply with public transparency principles and regulations, in order to ensure public oversight and accountability of AI tools.

In addition to the risk assessment frames provided in the previous sections, other questions need to be answered for a proper transparency evaluation.

Nr	QUESTIONS THAT MUST BE ANSWERED
I	Prior to employing the tool, is it possible to access previous impact reports containing the tests carried out to assess possible biases, as well as the measures taken to avoid or mitigate discriminatory behaviors of the AI tool?
II	Prior to employing the tool, is it possible to access reports disclosing information on the chosen model, the reasons behind its development, its purpose and affected populations, as well as the fundamental rights that could be affected by it and what measures were taken to prevent this?
III	Is it possible to access information on the input variables of the AI system?

CONTINUES

CONTINUATION

Nr	QUESTIONS THAT MUST BE ANSWERED
IV	Are there metrics to assess the tool's accuracy?
V	Is it possible to access the algorithm developed/used by the AI tool?
VI	After employing the tool, are there regular impact reports with updated bias and accuracy tests, tool corrections and improvements, as well as accountability reports regarding impacts on individuals and populations affected by the tool?
VII	Are the ones responsible for implementing the tool able to explain to the affected person(s) the reasons for certain results?
VIII	Is there anyone responsible for monitoring how the algorithm affects human decision-making? Are there any studies, reports or researches analyzing the interaction between humans and machines?
ASSESSMENT: HIGH, MODERATE OR LOW TRANSPARENCY	

CONCLUSION

Employing Artificial Intelligence systems to support public policies, services and goods provision requires attentiveness to its effects on fundamental rights and liberties. It also requires special attention to the enforcement of laws and regulations applicable to the public administration, especially when involving the principles of transparency and accountability.

It is essential to demand from government agencies employing AI tools documents such as impact reports – past and updated ones. These documents must be publicly available, preferably on the website of the referred agency.

This fundamental rights risk assessment framework and transparency evaluation intends to support public oversight efforts on monitoring AI tools and systems employed by government agencies. It works as a guide to identify critical points and elaborate recommendations to the government, to demand more public transparency, corrections or tests to ensure non-discrimination and prevent algorithm errors, or even to demand the discontinuation of any tool plagued by risks that cannot be controlled or mitigated.

TRANSPARÊNCIA
BRASIL

transparencia.org.br