# DATA PRIVACY BRASIL RESEARCH ASSOCIATION RESPONSE TO THE PUBLIC CONSULTATION ON THE OECD FRAMEWORK FOR CLASSIFYING AI SYSTEMS - COMMENTS ON THE DRAFT FRAMEWORK.

**Authors: Bruno Bioni, Thaís Helena Aguiar, Helena Secaf, Nathan Paschoalini, Marina Garrote, Júlia Mendonça.**

June 2021.

---

**The following are considerations about key dimensions 1 (Context) and 2 (Data and Input) of the draft Framework made available by the OECD.AI Policy Observatory for public consultation.**

## QUESTION 1. Should core and non-core criteria be distinguished? I.e. should there be a core classification framework for information that is generally accessible and additional, more complex or technical, considerations?

We understand this classification as useful, provided that the distinction between core and non-core criteria is based on risk. The main criteria would be those that are able to indicate the degree of risk that the use of AI systems represents in terms of violations of human or fundamental rights. In this sense, there are some criteria that need to be moved to the core group, which is explained in the following questions.

In addition, we also suggest modifying the term "optional" to avoid the impression that such criteria are "disposable", considering that, according to the question itself, the difference between the groups of criteria in helping policy makers to assess AI systems opportunities and challenges is not related to their relevance, but rather would involve the degree of accessibility of information linked to each criterion. In this case, this should also be made clear in the AI framework report.

## QUESTION 2. Which characteristics should be core criteria and which 'optional'?

**Question 2, Dimension 1 - Context:** In accordance with a risk-based classification framework, in which what makes a characteristic a core criteria is the ability to help identifying the severity of the risk posed by AI, supporting a more detailed and cautious regulation, we highlight as context core criteria the following:

A. The Industrial Sector: we agree that different sectors raise different policy implications, since some of those sectors' activities are, by nature, more likely to generate more serious risks of rights' violation. This implies that decision errors through IA can generate very serious consequences. In this sense, there should be a higher regulatory scrutiny for the use of AI in this sector. For instance: hospitals, arrest warrants, immigration services, education are sectors in which fundamental human rights are at stake and an error could result in serious harm. On the other hand, from the entertainment sector, for instance, the risks of rights being violated is lower. [principles: 1.1 and 2.4]

B. Business Function: in terms of a risk analysis, this criteria is an important complement for the Industrial Sector criteria. While the sector gives us an overview of the potential sensibilities of AI when used in a particular area, the business function provides a more detailed view on the actual AI function, confirming or displacing the first general impression of risks and sensibilities based on the sector. For instance, the health sector is a very sensitive one, because AI usage can impact fundamental rights broadly related to life and death issues. Nonetheless, if the AI system used in a hospital is used for monitoring something unrelated to health and patients, the implications are not related to the sectors' core activity and the regulation can be less strict. [principles: 1.1, 1.2, 1.4., 2.4]

C. Impacts critical functions/activities: this criteria should be a core one because it highlights which sectors and functions present serious consequences in case of interruption or disruption. In other words, this criteria is a risk analysis by itself: it is a risk classification of the sector and business function criteria based on an assessment not limited to fundamental rights. Critical sectors and critical functions carried out by AI deserve even more attention due to the impact it may have in critical areas. Among the list already presented[1], we would add to "the effective functioning of services essential to the economy and society, and of the government" the effective functioning of democracy. [principle: 1.4]

D. Scale of deployment: The breath of the development can also help perceive the impact of the AI system, since the number of users gives us a dimension of the technology impact. This is not to say that AI technologies that violate few people's rights and negatively impact in few people's lives are to be tolerated, but rather that, for policy impact issues, the more people affected by the technology, the more regulatory attention it should draw (keeping in mind that this criterion should not be

[1] "1) the health, safety, and security of citizens; 2) the effective functioning of services essential to the economy and society, and of the government; or 3) economic and social prosperity more broadly" (OECD, 2019);"

considered alone). Moreover, in terms of risk assessment and policy making, we believe that measuring the security of an AI system is more important than measuring its maturity. This is why we do not understand the AI system's maturity as a core criterion, but as a constituent criterion of a security analysis, along with other security standards analysis. [principles: 1.4, 1.5, 2.1]

E. Users of AI Systems: apart from links to accountability, transparency and explainability and security, as already stated by the framework report, AI users and people affected by the technology are at the center of the risk-based approach to classify AI systems. In this sense, we understand this criterion as part of the core group.

F. Impacted stakeholders, optionality and business model: We agree that identifying to which social group the AI systems' users belong should be a core criteria, because it helps understanding the risks involved. The situation is more sensible – and therefore demands a more careful regulation, if it impacts the lives of already vulnerable groups, for instance: children, elderly, members of minority/already excluded and discriminated against. [principles: 1.3, 2.2]. We also agree that optionality/ dependence should be a core criteria, because of a defense/escape logic. If one can opt-out or correct the AI system, the violation of rights are still worrying, but there is more space for personal defense and, borderline, one can choose not to be a part of the system. At the same time, the lack of an opt-out possibility (or even the lack of a correction possibility) represents a situation in which violation of rights is more serious and harmful, since it happens in a compulsory situation.

G. Benefits and risks to human rights and democratic values: this is, in our view, the basis of the core criteria. The main goal of regulating AI systems should be to avoid human rights violations. If there is a risk of a fundamental right being violated by an AI system implemented by the state, then this risk should be identified before the development and acquisition of the system, prior to use and on an ongoing basis[2]. If it is a private actor, the human rights due-diligence is required: identifying potential discriminatory outcomes, taking effective action to prevent and mitigate discrimination and track responses and to be transparent about those efforts[3]. The risk based approach means the need to identify which fundamental rights can be violated and how it can happen, in order to avoid it. The other core-criteria gives us standards to analyze the severity of violations if they were to happen, allowing for an ex-ante regulation to pay attention to the more sensitive AI contexts, anticipating possible harms and taking measures to avoid it/mitigate it. The regulation "[...] must consider a kind of impact assessment capable of covering this plethora of fundamental rights at stake and, above all, more focused on a systemic-collective

[2] Toronto Declaration, 2018, paragraph 31.

[3] Toronto Declaration, 2018, paragraph 43 and 44.

dimension and not just an individual one."[4] [principles 1.3 and 1.5]. We do not believe that benefits and well-being have to be a core criteria, but, in our view, there is no harm in doing so. It is a variable that does not measure rights violation directly, but the well-being of individuals and the collective is something to try to foster - avoiding and correcting AI systems that mitigate or harm this well being. It might be a lighter variable, but we believe it's an extra coverage of concern for those affected.

**Question 2, Dimension 2 - Data quality:**

Data quality and appropriateness should be a core criteria to evaluate AI systems. Data is the main element in any AI system and its quality and appropriateness impacts directly its outcomes in terms of accuracy, equity, bias, discrimination, privacy and security. It is a common risk to AI systems the existence of training systems with incomplete or non-representative data or databases with historical and/or systemic bias, which then has the potential to produce unlawful outcomes violating human rights [Principles 1.2 and 1.4].

At its simplest level, a machine learning algorithm receives data as an input and uses the algorithm (model) to produce an output[5]. It is especially relevant that the data used to train the AI system must be of sufficient quality as to allow the AI system to produce outcomes that are predictable, reliable and optimized. As described in the Berkman Klein Center for Internet and Society Report "Artificial Intelligence & Human Rights: Opportunities & Risks", the relation between the quality of training data and the resulting data system is one of "garbage in, garbage out" problem, meaning that if the data used is biased the system will reflect or aggravate this biases, due to its learning process.[6]

As data sets usually lack diverse representation, there are systemic biases in data models. For example, if a company uses personal data of job candidates for recruiting purposes and the algorithm used was trained with a data set using biased data, for instance, data from a tech company that does not hire women or people of color, the algorithm outcome will be to conclude that only male and white are suitable candidates for the job

[4] Data Privacy Brasil. "Contribuição à consulta pública da Estratégia Brasileira de Inteligência Artificial". São Paulo, 1st ed., Reticências Creative Design Studio, April 2020, p. 33. Available at: https://www.dataprivacybr.org/wp-content/uploads/2020/06/E-BOOK-CONTRIBUIC%CC%A7A%CC%83O-DPBR-INTELIGE%CC%82NCIA-ARTIFICIAL-FINAL.pdf. Last visited: Jun. 29, 2021.

[5] Rovatsos, Michael, Brent Mittelstadt, and Ansgar Koene. "Landscape Summary: Bias In Algorithmic Decision-Making: what is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it?." (2019). CDEI. The Centre for Data Ethics and Innovation. p. 15-16. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819055/Landscape_Summary_-_Bias_in_Algorithmic_Decision-Making.pdf. Last visited: Jun. 29, 2021.

[6] Raso, Filippo A., Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. "Artificial intelligence & human rights: Opportunities & risks." Berkman Klein Center Research Publication 2018-6 (2018), p. 15. Available at: https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf. Last visited: Jun. 29, 2021.

position. That means that the biased data training data (data that was not appropriate for its purpose nor representative) has resulted in a biased outcome.[7]

Another example of a biased outcome, now regarding ethnic concerns, is the Compas Risk assessment. As demonstrated by investigation carried by the organisation ProPublica, Compas was 77 percent more likely to flag a black defendant as a higher risk of commiting a future violent crime and 45 percent more likely to commit a future crime of any kind. In addition to the racial bias, the algorithm had a low level of accuracy: from the people classified as likely to reoffend, only 61 percent were arrested for any crime in the two years following the risk assessment.[8]

## QUESTION 3. Can AI systems be classified consistently & reliably with the core criteria?

Consistent and reliable classification is likely to always remain a challenge in an AI policy framework regardless of the attempts towards a robust systematization; this, however, does not mean that such classification should not be pursued, but rather only reinforces the importance of continuous efforts in this direction. Nevertheless, as they currently are in the framework, the core criteria still need further explanation and/or inclusion of other categories in order to help policymakers to assess the impact of AI in public policy areas, which is the case of data quality (Framework dimension 2, D).

In this sense, we reinforce that risk assessment and management should be used as a guidance for the application of artificial intelligence systems to determine which risks are acceptable and which represent unacceptable harm, and to measure what type of regulatory effort is appropriate to mitigate those risks. This is necessary so that the analysis of impact on Human Rights and democratic values are considered core and fundamental criteria for this framework for the classification of AI systems.

The usage of a cost-benefit methodology as a parameter to the criteria to be applied by agencies means a better selection method of the AI approaches that represent the greatest economic, environmental, health, public safety and other public policy areas benefit, considering the minimization of social costs and distributional effects related to the regulation and deployment of AI applications.

Thus, the categorization of AI systems based on the weighting between risk and benefit to Human Rights and democratic values allows the public authorities to optimize

[7] Ebert, Isabel L., and Thorsten Busch. "Systemic bias in data models is a human rights issue". Open Global Rights, 2020. Available at: https://www.openglobalrights.org/systemic-bias-in-data-models-is-a-human-rights-issue/. Last visited: Jun. 29, 2021.

[8] Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks". ProPublica, 2016. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Last visited: Jun. 29, 2021.

their regulatory intervention, as well as allows an informed decision-making process by the agents when defining safeguards for the use of AI. However, it is necessary to consider the development of a regulatory framework that is based on the idea of hybrid regulation, in which there is active participation of both governmental and non-governmental actors, where both are regulators and regulated, working in a network and in a polycentric way.

Therefore, risks and benefits to human rights and democratic freedoms should guide the development of artificial intelligence systems, thus constituting a core criterion of this framework, in order to ensure the respect for the OECD AI principles of Inclusive growth, sustainable development and well-being and Human-centered values and fairness.

Moreover, understanding this as a core criteria allows a solid classification of artificial intelligence systems, since it is based on the weighting between risks and benefits, which implies on the consideration of impacts to fundamental rights and democratic values, and social gains when implementing an AI system.

## QUESTION 4. Which criteria should be in a more detailed, technically-oriented framework?

### Question 4, Dimension 1 - Benefits and risks to human rights and democratic values:

When discussing the development of a classification framework for artificial intelligence systems, it is necessary to choose a framework capable of understanding the variety of safeguards that may be required when implementing an AI system. In this sense, the precautionary principle becomes a useful tool for classifying such safeguards.

In consonance with the movement that advocated evidence-based public policymaking, the precautionary principle was developed to deal with scenarios in which there is, due to lack of scientific certainty, indeterminacy and ambiguity about the effects of something to be released into the environment. It is precisely this tension that is present in discussions about artificial intelligence, especially when the possible effects related to the automation, partial or total, of the decision-making process are unknown.

In this context, the precautionary principle plays the role of helping to articulate a benchmark capable of measuring the possible safeguards to be established, in order to catalog them based on a taxonomy that considers which actions and inactions should be taken in face of an eventual imbalance between risk and benefits with the use of AI.

Considering that data protection is not the only right affected by the implementation of artificial intelligence systems and assuming the potential impact on human rights, the Human Rights Impact Assessment and the Data Protection Impact Assessment emerge as fundamental tools for an implementation that reflects the principles established by the OECD, such as Accountability, Human-centred values and fairness, Transparency and explainability principles.

In this context, impact assessments emerge as tools to address possible negative consequences of an AI initiative on one or more relevant social interests, aiming to inform a decision on its formulation, as well as its continuity. Since more than one social interest can be affected by a given AI initiative, different types of impact assessment may coexist. Despite this variety, impact assessments share a common logic and structure.

Thus, considering the complexity in the elaboration of Impact Assessments on Human Rights and on Data Protection, we believe, according to the document made available by the OECD, that the criteria referring to risks and benefits to human rights and democratic values must be treated in a specific and technically oriented framework, in order to enable a more accurate use of these assessment mechanisms, avoiding the risk of these documents becoming merely a procedure.

**Question 4, Dimension 2 - Structure and format of data and input:**

In this section, the format of data and metadata are considered optional criteria. The explanation in the draft report sheds light on benefits of standardisation of data formats, such as facilitates in terms of interoperability, data re-use across applications, accessibility and system's robustness and security. In addition to these points, we suggest deepening considerations about data portability and interoperability, since these unfoldings of data and input have practical and extensive implications in terms of competition and consumers' rights.[9] Examples of this would be the degree to which an approach for enhancing access to and sharing of data puts users in control, as well as cross-agency regulatory and enforcement cooperation.[10] Therefore, this analysis has to be taken into account by policymakers in AI systems risk assessments.

**Question 4, Dimension 2 - Rights and 'identifiability':**

We consider that this core criterion deserves a deeper explanation regarding each concern raised by the categories of data domains. It could also benefit from making it clear that such issues can affect other key actors in data and input.

[9] As noted by OECD's workshop on the subject, data portability and interoperability present different opportunities and challenges to benefit competition; therefore, both remain important parts of the analysis. See: OECD (2021), Data portability, interoperability and competition. Available at: https://www.oecd.org/daf/competition/data-portability-interoperability-and-competition.htm. Last visited: Jun. 29, 2021.

[10] OECD (2015), "Drawing value from data as an infrastructure" In "Data-Driven Innovation: Big Data for Growth and Well-Being", OECD Publishing, Paris, https://doi.org/10.1787/9789264229358-8-en. Last visited: Jun. 29, 2021.

For instance, the framework states that proprietary data raise issues such as transparency and explainability (Principle 1.3), besides bias in AI systems (Principle 1.2) and considerations of business scale-up (Principle 2.2). While, in fact, these are concerns even more common in data that are privately held, they can affect other key actors in data collection and processing, such as public entities. For instance, considering transparency and explainability alone, and in light of the framework's aim to help policymakers and others to classify AI systems according to their potential impact on public policy according to OECD's AI principles, it should be noted that this principle breaks down into concrete obligations by states. According to the Toronto Declaration:

> "States must ensure and require accountability and maximum possible transparency around public sector use of machine learning systems. This must include explainability and intelligibility in the use of these technologies so that the impact on affected individuals and groups can be effectively scrutinised by independent entities, responsibilities established, and actors held to account. States should:
>
> a) Publicly disclose where machine learning systems are used in the public sphere, provide information that explains in clear and accessible terms how automated and machine learning decision-making processes are reached, and document actions taken to identify, document and mitigate against discriminatory or other rights-harming impacts;
>
> b) Enable independent analysis and oversight by using systems that are auditable;
>
> c) Avoid using 'black box systems' that cannot be subjected to meaningful standards of accountability and transparency, and refrain from using these systems at all in high-risk contexts."[11]

In terms of enforcing oversight, said Declaration, in article 33, makes it clear that States must take steps to ensure public officials are aware of and sensitive to the risks of discrimination and other rights harms in machine learning systems. Therefore, States should:

> a) Proactively adopt diverse hiring practices and engage in consultations to assure diverse perspectives so that those involved in the design, implementation, and review of machine learning represent a range of backgrounds and identities;
>
> b) Ensure that public bodies carry out training in human rights and data analysis for officials involved in the procurement, development, use and review of machine learning tools;
>
> c) Create mechanisms for independent oversight, including by judicial authorities when necessary;

[11] Article 32 of the "Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems", launched on May 16, 2018 at RightsCon Toronto. Available at: https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf. Last visited: 29 jun. 2021.

d) Ensure that machine learning-supported decisions meet international accepted standards for due process.

Still in terms of transparency, it is noteworthy that a UNESCO report on "Artificial Intelligence in Education: challenges and opportunities for sustainable development" (2019) identified not only many positive applications of AI, but also social and ethical concerns that must be addressed. Using the example of educational institutions using Machine Learning algorithms, the report found potential problems including lack of explainability and unfair discrimination and concluded that governments must clearly communicate the scope and purpose of any data collection exercise: what type of data will be collected, for what purpose the data will be used, and what consequences, intentional or not, may occur in the model of data.[12] Considering that areas of social policies (such as education, health and transport) are particularly sensitive, a basic due diligence procedure is needed to assess discriminatory effects. This procedure basically consists of (i) identifying potential discriminatory results, (ii) taking effective actions to prevent and mitigate discrimination in machine learning systems, and (iii) being transparent about efforts to identify, prevent and mitigate discrimination in machine learning systems.

Likewise, actors from private and other sectors should also pay attention to concrete unfoldings of the principles. An example of this comes from ICO's guidance on AI and data protection, which contains recommendations on best practice and technical measures that organisations can use to mitigate risks caused or exacerbated by the use of this technology from a data protection perspective, explaining how the principles corrected by the GDPR apply to AI projects, without losing sight of the benefits they can offer.[13] The guidance is reflective of current AI practices and is practically applicable, for instance, to proprietary data.

It must be kept in mind that the above are only examples of one principle (1.3) affected by one category of data domain, but the same applies to all categories and key actors in item C of dimension 2 of the framework, a whole section to which we suggest more details in accordance with the Toronto Declaration.[14]

[12] PEDRÓ, Francesc, et al. Artificial intelligence in education: challenges and opportunities for sustainable development. Paris: UNESCO, 2019, p. 32-33. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000366994. Last visited: Jun 29, 2021.

[13] Information Commissioner's Office (ICO). ICO launches guidance on AI and data protection. 2020. Available at: https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2020/07/ico-launches-guidance-on-ai-and-data-protection/. Last visited: Jun 29, 2021.

[14] Another example would be the risk of bias (proprietary data, 1.2), mentioned by the Toronto Declaration in its article 46: "when mapping risks, private sector actors should take into account risks commonly associated with machine learning systems – for example, training systems on incomplete or unrepresentative data, or datasets representing historic or systemic bias. Private actors should consult with relevant stakeholders in an inclusive manner, including affected groups, organizations that work on human rights, equality and discrimination, as well as independent human rights and machine learning experts."

## 5. Should there be industry or application domain specific criteria and classifications?

States and private sector actors should be able to promote the development and use of machine learning and related technologies when these systems help to exercise and enjoyment of Human Rights by individuals. Especially when the implementation of AI systems is carried out by the public sector, there is a need for a proper classification that holds as one of its core criteria the analysis of risks and benefits to human rights and democratic values, since the public sector has specific duties inherent to its nature, such as transparency, accountability and respect for fundamental rights, so that the violation of these rights and duties can endanger their citizens[15].

[15] "There are numerous other human rights that may be adversely affected through the use and misuse of machine learning systems, including the right to privacy and data protection, the right to freedom of expression and association, to participation in cultural life, equality before the law, and access to effective remedy. Systems that make decisions and process data can also undermine economic, social, and cultural rights; for example, they can impact the provision of vital services, such as healthcare and education, and limit access to opportunities like employment." (The Toronto Declaration, paragraph 6, 2018)