Tool 1: Robust and Responsible AI Checklist

This tool consolidates the main concerns by risk dimension of the AI lifecycle. The checklist must be reviewed continuously by the technical team accompanied by the decision-maker (Fritzler 2015; drivendata 2019).

Planning and Design

☐ **Correct definition of the problem and the public policy response**

- (Qualitative) Is the public policy problem clearly defined?
- (Qualitative) Describe how this problem is currently being addressed – considering responses by related institutions – and how the use of AI would improve the government response to this problem.
- (Qualitative) Were the protected groups or protected attributes identified within the project (e.g., age, gender, education level, race, level of marginalization, etc.)?
- (Qualitative) Were the actions or interventions to be carried out based on the result of the AI system defined?

☐ **AI Principles**

- (Quantitative) Has the need for an AI system been justified, considering other possible solutions that do not require the use of personal data and automated decisions?
- (Quantitative) Is there evidence that both public policy action and the recommendation of the AI system will result in a benefit to people and the planet by driving inclusive growth, sustainable development, and well-being?
- (Qualitative) For the implementation of these technologies, have there been similar previous projects, and have they been reviewed?
- (Quantitative) Have you considered minimizing the exposure of personally identifiable information (e.g., by anonymizing or not collecting information not relevant to the analysis)?

Lifecycle

☐ **Data Collection and Processing**

Data quality and relevance of the available data

- (Qualitative) Discuss possible historical social inequalities in the use case with specialists in the field.
- (Quantitative) Perform an exploratory analysis of the available data with which the model will be trained to identify possible historical biases or undesirable states.

Poor Correspondence between Ideal and Available Variables

- (Qualitative) The ideal target variables should be clearly stated. The collected/available variables must be analyzed to understand how suitable they are to substitute for the target variable. Systematic biases or validity of the proxy metric should be identified.
- (Qualitative) Has the use of the selected response variable been clearly justified for the purposes of the intervention?

☐ **Data qualification and completeness for the target population**

Probabilistic and Natural Samples

- (Qualitative) Have the possible differences between the database and the population for which the AI system is being developed been analyzed? (Use literature related to the topic and information from experts. Study in particular unmeasured selection biases.)
- (Quantitative) Although models can be built with various data sources, designed or natural, validation should ideally be carried out with a sample that allows statistical inference to the target population. The validation sample must appropriately cover the target population and sub-populations of interest.

Missing or incomplete attributes

- (Qualitative) Has an analysis of missing values and variables been performed?
- (Qualitative) Has it been determined whether there are important omitted variables for which there are no associated measurements? (If any)
- (Qualitative) Have the reasons for the missing observations been identified? (If any)

☐ **Causal comparison**

- (Qualitative) Understand and describe the reasons why the response variable is correlated with known and unknown variables. Describe possible biases based on expert knowledge and analysis.
- (Qualitative) In the event that no work was done to ensure causality in the results, were the limitations of the results explicitly communicated to the public policy decision-maker?

Model Building and Validation

☐ **Absence or inappropriate use of validation samples**

- (Quantitative) Were the validation and test samples constructed properly, considering an appropriate size, covering subgroups of interest and protected subgroups, and avoiding information leaks during its implementation?

☐ **Data leakage**

Training-Validation Data Leak

- (Quantitative) Any processing and preparation of training data should avoid using the validation or test data in any way. A solid barrier must be maintained between training versus validation and testing. This includes data recoding, normalizations, selection of variables, identification of outliers, and any other type of preparation of any variable to be included in the models. This also includes sample weights or balances based on oversampling/undersampling.

Target Leakage

- The validation scheme should replicate as closely as possible the scheme under which the predictions will be applied.

☐ **Probabilities and classes**

Imbalanced Data

- (Quantitative) Make probability predictions instead of class predictions. These probabilities can be incorporated into the subsequent decision process as such.
- Quantitative) When the absolute number of minority cases is very small, it can be very difficult to find appropriate information to discriminate that class. More data need to be collected from the minority class.
- (Quantitative) Sub-sampling the dominant class (weighting the cases up to avoid losing calibration) can be a successful strategy to reduce data size and training time without affecting predictive performance.
- (Quantitative) Replicate the minority class to better balance the classes (over-sampling).

Arbitrary Cut-off Point

- (Quantitative) Using probabilistic classification algorithms is more suitable for decision-making to incorporate uncertainty regarding the classification.
- (Quantitative) Avoid standard probability cut-off points such as 0.5. Choose an optimal interpretation of the predicted probabilities using the receiving operating characteristic curve and other measures to analyze errors.

Adequateness of assessment metrics

- (Qualitative) Were the implications of the different types of errors for the specific use case, as well as the correct way to evaluate them, questioned?
- (Qualitative) Were the limitations of the model clearly explained? This implies identifying both false positives and false negatives and the implications that a system decision would have on the life of the target population.
- (Quantitative) Was a cost-benefit analysis of the system conducted and compared with the status quo or with the use of other decision-making or decision support strategies? (When possible)

Underfit and Overfit Checklist

- (Quantitative) Overfitting: If necessary, methods should be refined to moderate the overfitting, including such methods as regularization, restricting the functional space of possible models, using more training data, or disturbing the training data (Hastie, Tibshirani, and Friedman 2017).
- (Quantitative) Underfit: Data on protected groups or other sensitive variables should be reviewed to verify that there are no undesirable systematic errors.

☐ **Unquantified errors and human evaluation**

Failures Not Measured by the Model

- (Qualitative) Was a human assessment conducted with use-case experts to look for known biases or errors? Establishing monitoring schemes that allow for the identification of unmeasured errors or biases is recommended. For example, panels of reviewers can be used to examine particular predictions and consider whether they are reasonable. These panels must be balanced in terms of user type and expertise, including decision-makers if necessary.

☐ **Fairness and differential performance**

Algorithmic Fairness and Inequality

- (Qualitative) Was the algorithmic fairness criterion to be used in the model defined with experts and decision-makers?
- (Quantitative) When protected attributes exist, an assessment must be made of how far predictions deviate from the chosen algorithmic fairness definition. (e.g., tested for disparate error rates)?
- (Quantitative) In the case of classification models, cut-off points for different subgroups can be adjusted to achieve the chosen algorithmic fairness criterion.

Deployment and Monitoring

☐ **Performance degradation:**

- (Qualitative) Is there a plan to monitor the performance of the model and the collection of information over time?
- (Quantitative) Monitor various metrics associated with predictions in predefined subgroups (including protected variables).
- (Quantitative) Monitor drift in variable distributions with respect to the training set.
- (Quantitative) Monitor changes in the data collection and processing methodology that may reduce the quality of predictions.
- (Quantitative) When possible, plan to assign randomized (or status quo) treatments to some units under experimental designs. Make performance and behavior comparisons between this sample and the results under the algorithmic regime.
- (Quantitative) Identify unobserved variables and seek ways to measure them. If possible, re-fit the model and evaluate model performance using this information.

Accountability

☐ **Interpretability and explanation of predictions**

Explainability of Individual Predictions

- (Qualitative) Were the legal and ethical explainability requirements in the project's context analyzed?
- (Qualitative) Is there a process in place to provide explanations to particular individuals about why a decision was made?
- (Qualitative) Were the pros and cons of the algorithms discussed according to their level of interpretability and explainability to choose the most appropriate one?

Parsimonious Models

- (Qualitative) Including all available features to build and train a model may increase the risk of disproportionately affecting users. The variables to be included in the learning process must have some theoretical support or explanation of why they can help in the prediction task.
- (Quantitative) More parsimonious methods that use fewer, but relevant, features are preferable to models that use many, but less relevant, features.
- (Quantitative) Methods such as partial dependence plots (Friedman 2001) or permutations-based importance (Breiman 2001; Molnar 2019) can point to problematic variables that are heavily weighted in prediction against past observations or expert knowledge.

☐ **Traceability**

- (Quantitative) Is the AI lifecycle well documented (including data collection, infrastructure used, dependencies, code, metrics, and interpretation of results)?
- (Qualitative) Have the deficiencies, limitations, and biases of the model been communicated to stakeholders so that they are considered in decision-making/decision support?

(Qualitative) Has the technical team completed the Data Profile (see Tool 2) and the Model Card (see Tool 3), and has a process for continuous updating of these tools been defined?