Responsible AI Development for Data Scientists

The aim of this document is to ensure that ethical and legal and basic security standards are maintained. Each section relevant to the project in question should theoretically be performed as needed while each section is conducted in the project development process. If a project is following proper agile processes this implies that all checks should be conducted at each Sprint Review.

Failure to consider the questions in this document suggests that your project is a possible danger to your organisation's reputation and/or is a danger directly to persons.

This is a process which requires continuous improvement. If you have queries or contributions to make, please contact:
Pelonomi Moiloa (pelonomi.moiloa@gmail.com)

Contributors:
Pelonomi Moiloa
Alessandra Maraschin

This document was last updated on: 11 January 2022

Table of Contents

Introduction

A danger to human kind

The possibilities of AI are ever expanding. With increased opportunities to do good however there are just as many opportunities to do harm and even in the case that good intentions are at the helm, opportunities for good may eventually prove to be the opposite. We have many examples of this. In 2020 alone advanced technologies were behind: britain delaying effective Covid19 spread prevention methods, a Telegram bot app removed the clothing off pictures of women, "universal" technical tools like Twitter and Zoom perpetuated racial discrimination and language models were praised for their ability to smash certain academic benchmarks when their real life applications continuously prove to be a threat to already discriminated, marginalised, oppressed and vulnerable communities (read: bot suggests suicide to imitation mentally ill patient).

The greatest threat to what machine learning is able to achieve is two fold: Machine learning created with harmful biases built into its core intentionally and machine learning that does not reflect the diversity of the users it serves. It is important that we are not so pre-occupied with advancing humankind into the future that we have not taken the time to invest the energy into engineering the security measures this future requires. It is important to realise that humans are inherently biased and loaded with prejudice and that these traits find themselves in the machines we create and increasingly allow to run our lives.

International movements toward AI regulation

Measures have been put into place to protect personal information. GDPR in Europe was implemented in 2018 affecting all institutions that interact with the union. South Africa followed behind with our POPI Act, officially coming into being 5 years later. Though personal data is protected, how it is used has no protective measures in place currently that speak to technology specifically rather than considering indictments on human rights. The need for regulating data and data applications is increasingly recognised internationally. With ,officially, The European Commission releasing a Proposal for a Regulation of the European Parliament and of the Council with their document *Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence ACT) and Amending Certain Union Legislative Acts* in April of 2021. The drive to put in place formal regulations for AI has begun. The aim of the proposal is to encourage responsible uptake of AI technologies. Regulation in South Africa too is inevitable and we may as well put into place strategies to ensure the appropriate development of AI technologies now. Not only does this help you to align your practise with international standards but also to ensure the protection of the people you aim to serve.

AI regulation proposal

The regulation proposed applies to all machine learning techniques including supervised, unsupervised and reinforcement learning. It also applies to logic and knowledge based systems as well as to statistical approaches such as Bayesian estimation and optimisation methods. The proposal refers to high risk applications. Risk applications include Biometric Identification, Management of critical infrastructure, Education, Employment, Essential services access e.g Creditworthiness, Law enforcement, Migration management and Administration of justice.

The proposal then goes on to include technical documentation to accompany analytical projects such that their particulars are adequately recorded. The proposal contains many of the recommendations derived by the FACT (Fair)conference. This includes:

1. A general description of the AI system

2. Detailed description of the elements, this refers Datasheets as well as to the Analysis, Modelling of the use case which may be representee by a Fact Sheet

The results of the checks and considerations stipulated in the proposal are summarised in an EU Declaration of Conformity, a type of certificate that states certain details of the application and confirms adherence to regulations. It is also required for a list of these systems to be recorded along with their particulars along with a list of procedures in place to ensure the quality management of these systems with more particular regulations around how the quality management system is managed itself.

Areas Covered in this screening

There are three main areas concerned with regulating AI. The first is the machine conceptualisation phase, the second the machine development phase and the third the machine management phase.

Machine Conceptualisation: This phase is primarily concerned with the ethics of a particular use case to ensure that the intentions of the use case are pure. Even in the case that the intensions of a use case are pure, the metrics used to determine the use cases success may be ill suited resulting in unethical models. Some use cases need to be pursued and others by virtue of their principles should not. High risk type models generally teeter on the edge of the accepted ethical landscape. Application areas that fall into this category include: biometric identification and surveillance, employment screening processes, essential services access such as loans and credit risk, law enforcement and administration of justice as well as cases that infringe on the regulations set in place to protect privacy of information and persons. Both Section B and Section C of this document deal with this concern.

Machine Development: This phase deals primarily with algorithmic fairness and addressing algorithmic bias which is largely a problem due to the black box nature of machine learning models, the overwhelming mathematical complexity of non machine learning models but also due to the lack of consideration in screening input data and the subsets and subpopulations that exist within it. There are many technical tools that have been built to assist in addressing these issues as can be found in this Github repo. Many of these focus on transparency and interpretability. Section C and Section D of this document deal with these areas.

Machine Management: This phase primarily deals with the assurance that the model system developed is robust and secure and that the relevant measures are put into place to ensure accountability of the systems and its creators. Under this section replicability, deployment and system security is covered in Section E and F. Also very important here are considerations of visualisation and interpretation of results. It can be quite easy to mislead through visualisation this is covered in Section G.

Projects Eligible for screening

In the case of any of the project categories below wherein the project results in an insight that informs any kind of decision making process, the relevant sections should be completed. The purpose of the project screening is not to deduce high from low risk models as this step is to be conducted within Model Risk Management. The aim of the screening process is to ensure developmental integrity of Data Engineering, Modelling and Analysis.

Project categories:

- Machine learning (supervised, unsupervised, reinforcement learning)
- Logic and Knowledge based analysis and calculations
- Statistical projects e.g Bayesian Estimation, optimisation methods and other summary statistics

Use Case/ Project Screening

Section A: Use Case/Project Particulars

Name:
Project members:
Overview:
Purpose/objectives:
Intended domain:
Stakeholders:

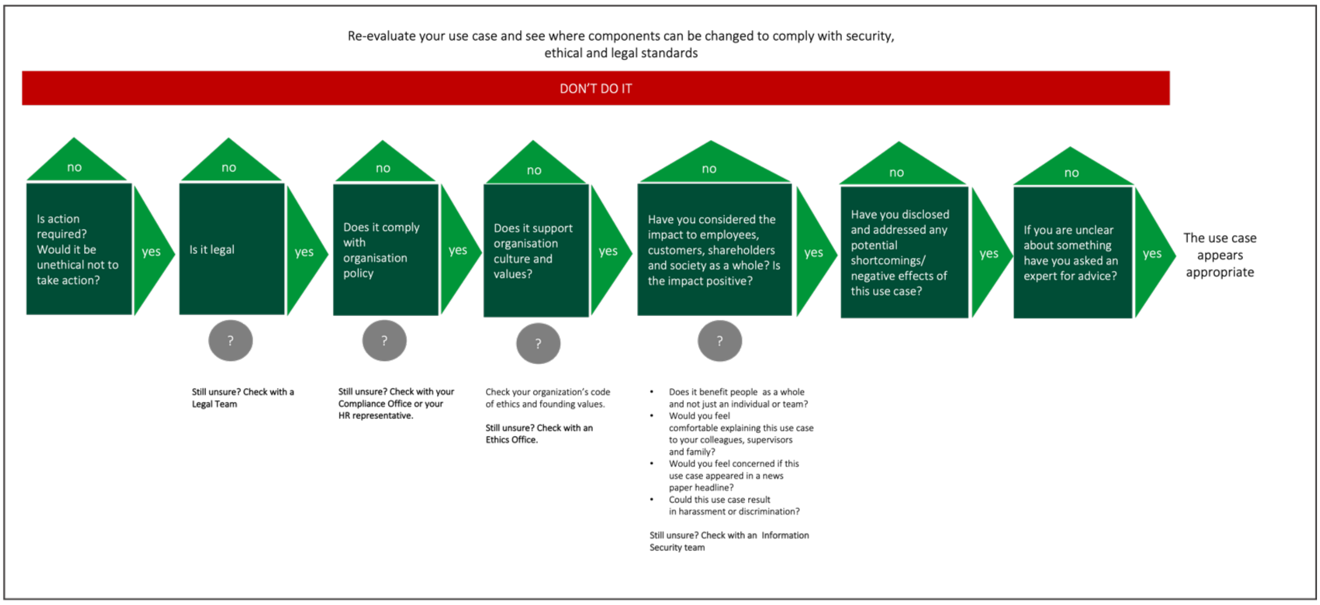
Section B: Use Case/Project Integrity Check

The aim of the Use Case Integrity check is to ensure that the project in question is:

- Legal
- Ethical
- Secure

Only when a resounding yes can be answered for the questions in the pipeline below, should the use case be considered.

1. Is action required?
2. Is it legal?
3. Does it comply with organisational policy?
4. Does it support organisational culture and values?
5. Have you considered the impact to employees, customers, shareholders and society as a whole? Is the impact positive?
6. Have you disclosed and addressed any potential shortcomings/negative effects of this use case?
7. Have you engaged an Architect and an Information Security person to highlight any issues?
8. If you are unclear about something have you asked an expert for advice?


Re-evaluate your use case and see where components can be changed to comply with security, ethical and legal standards

DON'T DO IT

Is action required? Would it be unethical not to take action?

Is it legal

Does it comply with organisation policy

Does it support organisation culture and values?

Have you considered the impact to employees, customers, shareholders and society as a whole? Is the impact positive?

Have you disclosed and addressed any potential shortcomings/ negative effects of this use case?

If you are unclear about something have you asked an expert for advice?

The use case appears appropriate

Still unsure? Check with a Legal Team

Still unsure? Check with your Compliance Office or your HR representative.

Check your organization's code of ethics and founding values.

Still unsure? Check with an Ethics Office.

- Does it benefit people as a whole and not just an individual or team?
- Would you feel comfortable explaining this use case to your colleagues, supervisors and family?
- Would you feel concerned if this use case appeared in a news paper headline?
- Could this use case result in harassment or discrimination?

Still unsure? Check with an Information Security team

Action Item: Generate a scope document for your use case with answers to the questions above

Section C: Projects involving sourcing/processing of data

Data integrity is the maintenance of, and the assurance of the accuracy and consistency of data over its entire life-cycle. As the saying goes, “Garbage in garbage out”. Ensuring data quality ensures that models can produce the best results possible.

Also included here is checks to see where implicit bias may creep into data resulting in unfair discrimination or infringements on human rights related to personal information, security and automation.

The following questions are adapted from:
Datasheets for Datasets
Timnit et al
19 March 2020
https://arxiv.org/pdf/1803.09010.pdf

A. Motivation

- □ **Purpose**: For what purpose was this dataset created?
- □ **Creator**: Who created the dataset?
- □ **Funding**: Which party required the creation of this dataset?

B. Composition

- □ **Data format**: What data mediums are present in the dataset e.g. documents, photos, relational data?
- □ **Volume:** How much data is there?
- □ **Sample percentage:** Is the data a sample of a larger set and is it sampled at random, is there information missing?
- □ **Instance composition:** Is the data raw or is it features?
- □ **Labels:** Is there a label or target and is there a recommended test and train split?
- □ **Relationship:** Are there relationships between data collections?
- □ **Noise:** Is there noise in the dataset?
- □ **Source:** Is the data self-sourced or does it rely on external resources (e.g. twitter) and will data not change?
- □ **Personal Information:** Does the dataset contain confidential information?
- □ **Offensive information:** Does the dataset contain disturbing information?
- □ **People:** Does the dataset relate to people?
    - ▪ **Subpopulation:** Does the dataset identify subpopulations?
    - ▪ **Privacy:** Is it possible to identify individuals and their persona sensitive information (e.g. data that reveals race, ethnicity, sexual orientation, religion)?
    - ▪ **Limit PII exposure**: Have we considered ways to minimize exposure of Personal Identifiable Information (PII) for example through anonymization or not collecting information that isn't relevant for analysis?

C. Collection Process

- □ **Acquisition:** How was data acquired (e.g. directly observable, reported by subjects, indirectly from other data)?
- □ **Collection mechanism:** What collection mechanism was used to collect data (e.g. sensor, manual human curation, software program, API)
- □ **Sampling:** If the dataset is a subset of a larger population, what was the sampling process?
- □ **Timeframe:** What was the timeframe over which data was collected?
- □ **Ethical considerations:** Was ethical approval required for data collection, what kind?
- □ **People:** Does the dataset relate to people?
    - ▪ **Source:** Did you collect data from individuals directly or from a third party?
    - ▪ **Consent:** Did the individuals give consent for their data to be collected and were they given the option to revoke consent?
    - ▪ **Potential Impact:** Has a potential impact analysis of this dataset on these people been conducted?

D. Pre-processing

- □ **Pre-processing:** Was any prepossessing conducted?
- □ **Raw Data availability:** Is the raw data available?
- □ **Method availability:** Is the pre-processing method/software available?

E. Uses

- □ **Previous usage:** Has the dataset been used previously, provide access points if available?
- □ **Future use:** Is there anything about the dataset that may impact future uses?
- □ **Prohibited use:** Are there tasks for which the dataset should not be used?

F. Distribution

- □ **Third parties:** Will/can the dataset be distributed to third parties how?
    - ▪ **How?**
    - ▪ **When?**
    - ▪ **Under copyright?**
- □ **Restrictions:** Are there distribution restrictions?

G. Maintenance

- □ **Who:** Who is maintaining the dataset if it is being maintained?
- □ **Contact:** How can the maintainer be contacted
- □ **Updates:** Will the dataset be updated

H. Data management/protection

- □ **Data security**: Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?
- □ **Data Access**: Have you limited access to the relevant databases and put definitive access control measures in place?
- □ **Right to be forgotten**: Do we have a mechanism through which an individual can request their personal information be removed?
- □ **Data retention plan**: Is there a schedule or plan to delete the data after it is no longer needed? Are you managing memory limitations effectively?

Action item: Create a Datasheet for the data based on the questions above

Section D: Projects that involve the analysis/modelling of data

The aim of screening the analysis/modelling procedure is to gain understanding of developed models such that we can be sure that the model is in fact performing per requirements and as we believe without hidden consequences. It is exceptionally important to check that these points have been considered before models go into production to avoid harm to persons.

The below questions are derived from a combination of the guidelines set by the Deon ethics checklist for data scientists as well as from Annex IV of the EU Commission Proposal for regulating AI. The below encompasses a process I have dubbed EMA (exploratory model analysis).

I. A. Analysis

- ☐ **Missing perspectives**: Have we sought to address blind spots in the analysis through engagement with relevant stakeholders (e.g., checking assumptions and discussing implications with affected communities and subject matter experts)?
- ☐ **Dataset bias**: Have we examined the data for possible sources of bias and taken steps to mitigate or address these biases (e.g., stereotype perpetuation, confirmation bias, imbalanced classes, or omitted confounding variables)?

  Notes: Autostrat – for subsets analysis before model development, Bias scan – for subsets after model development. Comparing Autostrat results and Bias Scan results can give an indication of where bias has been amplified
- ☐ **Honest representation**: Are our visualizations, summary statistics, and reports designed to honestly represent the underlying data?
- ☐ **Privacy in analysis**: Have we ensured that data with PII are not used or displayed unless necessary for the analysis?
- ☐ **Auditability**: Is the process of generating the analysis well documented and reproducible if we discover issues in the future?
- ☐ **Outliers and Anomalies**: Does your analysis have to deal with anomalies and outliers? How is this addressed?

J. B. Modelling

- ☐ **Proxy discrimination**: Have we ensured that the model does not rely on variables or proxies for variables that are unfairly discriminatory?
- ☐ **Fairness across groups**: Have we tested model results for fairness with respect to different affected groups (e.g., tested for disparate error rates)?
- ☐ **Metric selection**: Have we selected the correct metric relevant to the problem to be solved and considered the effects of optimizing for our defined metrics and considered additional metrics?
- ☐ **Explicable**: Can we explain in understandable terms a decision the model made in cases where a justification is needed?
- ☐ **Communicate bias**: Have we communicated the shortcomings, limitations, and biases of the model to relevant stakeholders in ways that can be generally understood?
- ☐ **Validation and testing**: What kind of validation and testing was used to confirm results?

Action Item: Generate an analysis summary based on the questions above

Section E: Projects that involve the deployment of a model

The intention of a FactSheet is to provide transparency for a model that exists in production in order to ensure proper use. It should answer any questions about how the model was made such that it can be replicated if need be and contains details of the analysis/modelling check to illustrate what performance requirements were taken into account for any consumer or persons affected by the model. The questions and details required in the Fact Sheet below are adapted from Factsheets as proposed by IBM as well as the EU Commission Proposal for regulating AI.

A. Model Facts

- ☐ **Data Transform**: What are the particulars in the data used to build the model
    - Link to dataset or generative script
    - Motivations for selected features
    - Motivations for engineered features
    - Dataset distribution (Test/Train/Validate)
- ☐ Model Results:
    - Data
        - Problem type (regression, classification etc)
        - Input type (structured and unstructured)
        - Training data path
    - Model output path (access point)
    - Training algorithm
        - Optimal conditions
        - Poor conditions
    - Feature list
        - Fairness feature list
    - Fairness level (AIF360 score?)
    - Explainability level (AIX360 score?)
    - Adversarial Robustness (ART score?)
    - Quality/Performance Metrics
        - Accuracy/Area under ROC/F1/MCC or BLEU/CIDER etc
        - Speed
        - Compute
        - Climate impact
        - Others
    - Model limitations:
        - Degrees of performance for different groups
        - Specific persons of models intended use
        - Foreseeable unintended outcomes/sources of risk
- ☐ Results from Sections A through D that were implemented prior to model development/deployment

Action Item: If you have put a model into production generate a fact sheet for your model deployment with the above information

Section F: Projects that involve deployment

The below questions are derived from a combination of the guidelines set by Factsheets as proposed by IBM, ANNEX IV of the EU Commission Proposal for regulating AI as well as possible organisational specific protocols and functioning.

A. Code Quality

- ☐ **Logging:** Have you implemented logging?
- ☐ **Exception Handling:** Have you implemented exception handling?
- ☐ **Coding Standards**: Have you adhered to coding standards, PEP8 etc?
- ☐ **Version Control:** Have you applied adequate version control measures? Does your project come packaged with a README/ Requirements.txt with dependencies and their versions?
- ☐ **Config:** Has all confidential information been kept from public file sharing systems/ encrypted in non-accessible config files, such as API keys, database passwords etc
- ☐ **Requirements:** All libraries used within the project should be put in a requirements file along with the versions and dependencies used for that project.

B. Deployment

- ☐ **Redress**: Have we discussed with our organization a plan for response if users are harmed by the results (e.g., how does the data science team evaluate these cases and update analysis and models to prevent future harm)?
- ☐ **Change Management**: Have you implemented change management processes for when changes need to be made in production?
- ☐ **Roll back**: Is there a way to turn off or roll back the model in production if necessary?
- ☐ **Concept drift**: Do we test and monitor for concept drift to ensure the model remains fair over time?
- ☐ **Unintended use**: Have we taken steps to identify and prevent unintended uses and abuse of the model and do we have a plan to monitor these once the model is deployed
- ☐ **App registration**: Does your app require registration with a particular entity? Have you registered your app on this entity/submitted the model to your model inventory?
- ☐ **Up/Down time:** What is the considered availability of the application and the considered app amendment time frame for fixing malfunctions
- ☐ **Architect/Security team clearance**: Have you cleared your use case implementation with the relevant technical experts?
- ☐ **Environments**: Does the app require you to have Dev, QA, ETE, and PROD? Do they need to be hosted on separate servers/nodes?
- ☐ **Load testing**: If the app is an API have you done load testing?
- ☐ **Human Oversight:** How is the model monitored? Is decisioning regulated by a human or does the application have full control (full control is not recommended read Cathy Oneil: Weapons of Math Destruction)
- ☐ **Deployment location and access:** <details here>

C. System Security

- ☐ **System Access**: Have we ensured that system access has been limited to only those who should have access
- ☐ **Pipeline Interception Protection**: Have measures been taken to ensure that all possible points of data interception in the deployment pipeline have been secured

- □ **Database Security**: Has data storage been secured?
- □ **Application Security:** Has the application been sufficiently secured?

Action item: Create a deployment summary with the results of the above

Section G: Projects that require reporting/visualisation

A. Visualisation:

- □ Are any ambiguities in presentation results expressly communicated in the case that alternative visualisations cannot be chosen
- □ Have you expressed uncertainties
- □ Have you provided context where necessary

B. Selective Reporting:

- □ Hiding relevant data
- □ Presenting too much data
- □ Describing the data inaccurately in annotations, title, or within the visualization itself

Action item: generate a disclaimer on the above questions

Conclusion:

It is recommended that all the necessary sections are completed in writing in a metadata store with a private version available to project administrators and a public version (within the confines of IP, security and privacy limitations) that can be made public for data science artefact consumers.