AI Fairness Guide

AI Fairness Guide is also available tailor-made to your needs. Are you interested in professional support to tackle your ethical challenge? Contact us at info@ethix.ch.

Made possible by MIGROS Pioneer Fund

ethix

ETHICS AT THE HEART OF ALGORITHMS

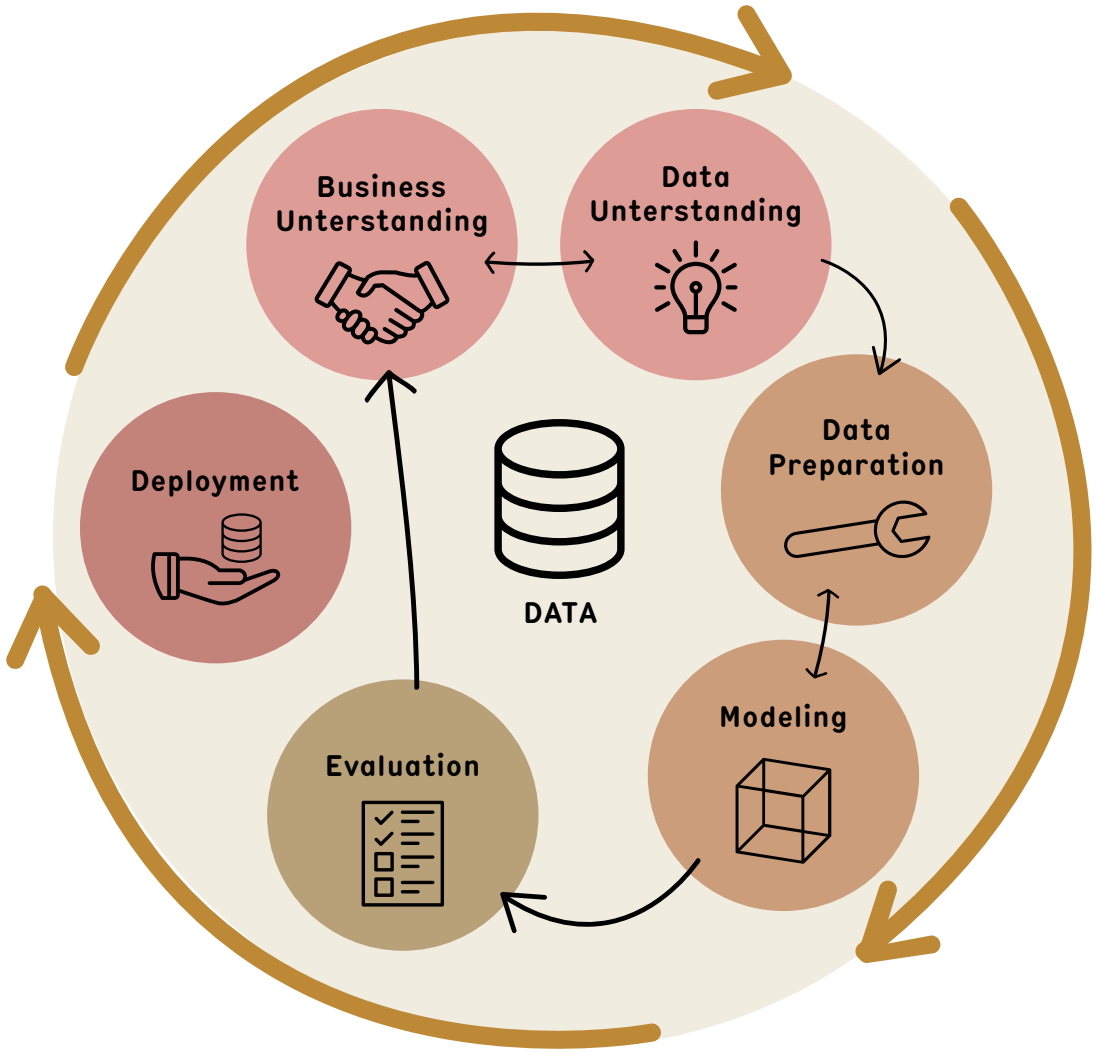
Automated data analysis opens up new economic opportunities. Beyond the hype of applying the term “artificial intelligence” to all sorts of things, data analysis is useful in all lines of business. AI allows us to learn more about our customers, our markets and our own performance. AI also makes new products and services possible. Businesses, but also public institutions and civil society organisations, can make great use of AI data analysis.

This document is intended for all those who programme, supervise or initiate data analysis projects. Data scientists, who engineer and configure these tools, embark on a long journey that starts with data collection and ends with the final implementation of the „intelligent“ system – passing through the stages of data cleansing, choice of algorithm, system parameterisation, etc.

In this document, we offer a checklist for artisans - data scientists who create these systems, as well as their supervisors, to help them identify ethical issues at various stages of their work. This checklist can be used independently as a guideline for evaluating the entire process. It is the result of discussions with our colleagues at Swiss Statistical Design & Innovation (Swiss SDI) and Datastory, two Swiss companies active in the field of data science. It is also based on the „Algo. Rules“ project of the Bertelsmann Foundation (Germany).

AI Fairness Guide

ethix



This diagram summarises, in a simplified way, the different steps of a data analysis process. It follows a chronological approach to the project. However, the different parts are self-contained and can be used independently.

You can use this document as a map, to navigate the different ethical challenges to be addressed at each of these stages. This document is of particular interest to data teams.

Further reading:

- *The project „Algo.Rules“ by the Bertelsmann Foundation: https://algorules.org/de/startseite*
- *The article by Raji et al., „Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing“ (2020)*
- *The scientific paper „Getting into the engine room: a blueprint to investigate the shadowy steps of AI ethics“, published by J. Rochel and F. Evéquoz in „AI & Society“ (2020) https://link.springer.com/article/10.1007/s00146-020- 01069-w*

AI Fairness Guide

ethix

Step 1
BUSINESS UNDERSTANDING

ACTIVITY CLARIFY PROJECT OBJECTIVES

Description

- Understand the client's needs and business objectives:
  What is the client aiming for with this project? What are the motivations behind these objectives?
- Define the stakeholders involved in this project:
  Who is pursuing which objectives in this project? Who else does the project affect?

Questions to be asked

- May the project's objectives potentially harm third parties or society as a whole? For example:
  - Do the project objectives imply redundancies?
  - What impact does the project have on users/society?
  - Does it have a particular impact on vulnerable people, for example, children?
- Could important project stakeholders feel threatened by the project itself, or the involvement of certain specialists? For example: an employee whose collaboration is essential for the project, might feel threatened as the project puts their job at risk.
- Are the responsibilities of the different project stakeholders clear? Are there actors with an overview of the project? Is the responsibility distributed according to the different project stages or with regard to the different domains? What are the roles, rights and obligations implied by this responsibility?

Best practices for the data team

- Know the client's business objectives to better assess the impact the project may have on third parties/society.
- Based on similar experiences, discuss with the client to help them understand the potential impacts the project may have.
- Ensure a good understanding from the client of which stakeholders are impacted by the project.
- Draw up an ethical charter for the project that may accompany its realisation, including a discussion on the underlying values of the project, the company's culture and values, the commitments made and the red lines the project must not cross.

AI Fairness Guide

Step 2
DATA UNDERSTANDING

ethix

ACTIVITY DETERMINE THE DATA NEEDS

Description

- Understand the client's data analysis needs
- Precisely define the final objectives of data collection and analysis

Questions to be asked

- Does a cross-check of data sources reveal sensitive data? (e.g. Allowing the identification of an individual and their activities)
- Are the final objectives of data collection and analysis clear?
- Do these objectives of data analysis align with the business objectives detailed above? In other words, does the data and its intended use allow us to answer the relevant business questions and only those questions?
- Is there a common understanding of the objectives and their consequences between the data team and the client? In particular, does the client detect the potential impact of data mining for their company?

Best practices for the data team

- Be aware of legal constraints relating to combinations of data sources (e.g. sensitive data under GDPR, risk of deanonymisation when combining derived data)
- Establish clear and precise objectives with the client
- Plan iterative checks that the evolving project maintains the defined basic aims

AI Fairness Guide

Step 2
DATA UNDERSTANDING

ethix

ACTIVITY DATA COLLECTION

Description
- Data selection and collection

Questions to be asked

- Who is in charge of data collection (client, third party)? Is there a possible conflict of interest?
- Are legal obligations/ best practices respected (consent, Data Protection Laws, GDPR)?
- Is the data stored adequately (security, storage period)?

Best practices for the data team

- Know the data collection process: explain the potential impact of negligence in this domain to the client.
- Make the client aware of the risks when using data, especially with regard to private data.
    - For example; identify particularly sensitive data, considering the protection of personal data, and plan security measures to prohibit non-authorised persons from accessing raw data, put in place a pre-emptive policy to delete non-essential data, to anonymise or pseudonymise, or even define an “expiration” date after which the data will be deleted.

AI Fairness Guide

Step 2
DATA UNDERSTANDING

ethix

ACTIVITY DATA DESCRIPTION

Description

- Describe data sources (origin, relevance, frequency of updates, etc.)
- Definition of the data (data type, the data's limits – what it portrays, what it leaves out–, what the categories mean (classes), a priori explanation of possible lacking data, ...)

Questions to be asked

- Is the client aware of the impact on the project's quality when a data description is erroneous (e.g. bad interpretation of the results, wrong conclusions, lack of context)?
- Which assumptions/biases are used while categorising data, in particular for a categorisation that is up for discussion (e.g. gender = [male, female], or other categories)?

Best practices for the data team

- Make the client aware of the potential impact of classification/labelling
- Describe the data exhaustively, with particular care placed in explaining the categories and their assumptions/ biases

AI Fairness Guide

Step 2
DATA UNDERSTANDING

ethix

ACTIVITY DATA VERIFICATION

Description

- Quality control of collected data
- Information control: Does the data correspond to the objectives defined at the beginning of the project?

Questions to be asked

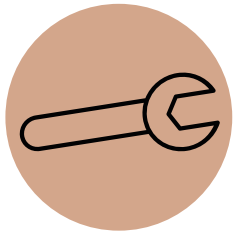
- How is the appropriate moment to stop checking, and "accept" the data determined? After which amount/ quality of testing?
- During quality control, how is it decided which data is trustworthy, and which data should be eliminated?
- What are the security measures put in place to protect "data privacy"?

Best practices for the data team

- Systematic control of data quality, even when data providers guarantee high quality. (e.g. carry out descriptive statistics for each identified variable in the dataset in order to identify any incoherence.)
- Define a "label" for data quality control, for example by defining sources that are trustworthy/reliable or a benchmark to refer to.
- Outline the normative benchmark that was used. This will help judge data quality and identify potential biases within the dataset.

AI Fairness Guide

Step 3
DATA PREPARATION



ethix

ACTIVITY DATA SELECTION – CLEANSING – CONSTRUCTION – INTEGRATION

Description

- Main data preparation process
  - Selection of important data
  - Data cleansing
  - Handling of missing data
  - Labeling – In preparation for model training and testing for supervised learning approaches.

Questions to be asked

- Is the data complete and balanced?
  - Does the generation of new data influence the final result?
- Are there biases in the data?
- Was data labelling done in a systematic way? Is the quality of labelling sufficient?

Best practices for the data team

- Complete description of categories (classes) used, and concepts present within the data (complete the descriptions created during the activity “data description”)
- Carry out synthetic data generation to complete missing data only when these are essential. Document the method of data generation. Keep in mind that the final result may be influenced consciously or unconsciously. Analyse potential modelling biases due/linked to data generation, by documenting these explicitly and transparently
- Control the effects of data correction (e.g. Compare models and results using different corrections)
- Uphold transparency towards the client and the partner by documenting meticulously the measures of data cleansing
- Explain the normative dimensions of the data preparation techniques in relation to the project objectives.
- Carefully document missing or unusable data and the techniques used to handle or correct them
- Carefully choose different methods for correcting data quality in order to minimise bias, and describe with care their potential implications (in terms of standards and further process).

AI Fairness Guide

ethix

Step 4
MODELING

ACTIVITY MODELING TECHNIQUE – TEST DESIGN – PARAMETERS – ASSESSMENT

This step is strongly linked to the next one (Evaluation). Both are largely interdependent and may be dealt with in parallel.

Description

- Create machine learning models
- Test and validate the models
- Document goals and expected effects

Questions to be asked

- What are the relevant criteria for evaluating the performance of the model?
- Based on the state of the art in this field and for the type of data examined, fix performance metrics
- Do these criteria relate to the objectives of the project? In what way? Are there contradictions between the different objectives? If so, which criteria should be favoured? On this basis, is the choice of model optimal? What are the alternatives?
- How to express discrepancies and define responsibilities when validating the model with the client?

Best practices for the data team

- Justify the choice of model on a mathematical basis and on the basis of experience (state of the art) - Explain the objective(s) the model needs to achieve (performance metrics)
- Explain the model's performance evaluation criteria in relation to the project's objectives - Document trade-offs between different legitimate objectives, which could be taken into account in the project's design and implementation
- List the implications, impacts and drawbacks of choosing this model, as well as the methods to assess the quality of the model
- Describe the limitations and vulnerabilities of the model: what is it the model (not) created for? What does it fail to detect?
- Consider adversarial machine learning techniques to identify potential model vulnerabilities.
- Perform a risk magnitude assessment that combines the probability of system failure with the seriousness of this failure.

AI Fairness Guide

Step 5
EVALUATION

ethix

ACTIVITY EVALUATE RESULTS

Description

- Validation of results
- Definition of next steps for the project

Questions to be asked

- Is the model stable?
- Is the model evaluated/ assessed on representative real world data or only on an unrepresentative training sample?
- Do results of the evaluation correspond to the client's needs?

Best practices for the data team

- Carry out a cross-validation to assess the quality of the model
- Systematically verify the evaluation data (benchmark) (see stage 2 "Data Verification"
- Precisely describe the evaluation data used to assess the results of the model (see Stage 2 ‚Data Description' and following steps).
- Ensure that the model is not biased for a particular subset of population/data subjects.
- Based on the results, specify the description of the system's scope and limitations. Organise a space for discussion with the client about the system's scope of application, the results it can (or cannot) achieve and communicate this clearly to stakeholders (project management).
- Plan an ongoing assessment with the client (agile development)

AI Fairness Guide

Step 5
DEPLOYMENT

ethix

ACTIVITY DEPLOYMENT

Description

- Going from evaluation (pilot phase) to operational phase: “model in production”
- Use of the model on real data (beyond a sample as in the previous phases)
- Integration of the data analysis tool into the processes of the company/organisation

Questions to be asked

- Do third parties give feedback to further develop the model (active learning)? If so, what precautions are taken to avoid any deviations from the model as a result of active learning, whether intentional or not?
- Is the final use of the tool in line with the initial objectives?
- Has the tool been well integrated into the processes of the company/organisation? Are the responsibilities for its end use defined?

Best practices for the data team

- Continuous monitoring of the results obtained, using metrics to ensure the stability of the model.
- Define responsibilities during deployment: e.g. a business expert identifies problematic results and a data scientist trains and verifies the model by going through the CRISP-DM process described here
- Inform stakeholders (including end-users if relevant) of the implications when using the system (limitations, etc.)
- Provide tools for end users to report results they deem problematic, and a procedure for analysing them.
- Put in place communication tools to inform users how the model works and ensure that the results obtained are understandable.

V1- 5/2021
ethix - Lab for Innovation Ethics
www.ethix.ch