TesseractAcademy

The lean AI ethics framework

AI ethics made easy

With the larger and larger role that AI is playing in our lives, it is clear that issues such as ethics, transparency and accountability will play a key role.

There are many opinions from different regulators around how the issue of AI ethics should be handled. Many of them end up in long documents which are difficult to read, understand and act upon.

The mission of the Tesseract Academy is to simplify technology so that decision makers know how to make the right moves in the competitive arena of technology.

This is a simple framework which can ensure that your organisation has nailed the fundamentals of AI ethics, irrespective of what current or future legislation will bring.

This framework assumes that ethical use of AI concerns three main aspects:

- Bias
- Privacy and security
- Accuracy

Bias refers to those situations where a model might be unfairly treating certain individuals. For example, in a famous case, an algorithm called COMPAS had been used to make automated decisions on whether someone should be released from prison. This algorithm was biased against certain ethnicities and was later scrapped. This was an algorithm used in a sensitive domain, trained with biased data, which led to a very bad outcome for many individuals.

Privacy and Security are topics which have traditionally been associated with the domain of data, but it has now started playing a larger role in machine learning models. There are certain ways that models can be manipulated by attackers in order to produce fraudulent results.

Finally, **Accuracy** is important in those domains where the results of a model can have life-changing circumstances.

The framework

The framework is built upon 3 axes:

- Data
- Algorithms
- Applications

The questions assessed in the data part have somewhat general applicability. The questions asked in the algorithms and application part assume that you are interested in developing an AI application with a particular predictive goal in mind.

Examples of such applications (where bias might exist) are:

- An algorithm that can decide who proceeds to the next interview round in the hiring process.

- An algorithm that detects whether there is a human face on a camera. There have been instances of algorithms in the past that had trouble identifying non-white people.
- An algorithm that predicts instances of recidivism of inmates. There was a big scandal where such an algorithm was biased against certain racial groups (i.e. the COMPAS software case study).

Self-assessment

This is a self-assessment framework. The objective of the self-assessment is to understand where you stand with regards to potential ethical issues with your data, models and applications. Some of those points (e.g. adversarial attacks) are the object of active research and would take a whole book to analyse them.

The goal of this framework is to make sure that:

- You are aware of all these issues.
- You have informed yourself of potential risks.

These are the first and most important steps to setting up an AI ethics strategy and engaging with the right specialists.

Data

Data itself can raise ethical issues if you have acquired it in an unethical way, or you are not complying with privacy regulations. However, data can cause an even larger issue if they contain potential sources of bias and are fed into an algorithm.

- Where is data being sourced from
    - If the data involves human subjects, have they given their consent?
    - Has this been acquired by a third party? If yes, then has this collection taken place in an ethical manner?

- Bias
    - Are there any variables in the dataset which could cause bias? Examples of such variables are gender, and ethnicity.
    - Are there any variables which can be affected by the variables in the previous question? One such example, can be salary. Different ethnicities might have different average salaries. For example, if you are developing an algorithm to suggest a salary to new hires, maybe this algorithm will suggest a lower salary to certain minorities.
- Data privacy
    - Are you GDPR compliant?
    - Are you cybersecurity compliant? (e.g. ISO27001)

Self-assessment

- Interpretability
    - Can the algorithm provide interpretable output? This usually means that it can identify the most important variables it uses to make predictions.
    - If the algorithm can't provide interpretable output, then is your data science team familiar with model-agnostic methods such as Shapley value explanations, which can help you assess feature importance?
- Effect of contentious variables
    - For the variables that you have identified as potential sources of bias, or potentially affected by bias, what results does the algorithm provide? For example, if gender and salary are part of the algorithm, then how does salary change as you change gender?
    - If you remove these variables from the dataset, and retrain, do the results of the model radically change?
- Drift
    - If the algorithm is being restrained constantly, can this lead to the algorithm developing bias or performance degradation later on?

For example, if you keep getting more data that include sensitive variables, could this affect the performance of an algorithm? Or if you are in the medical domain, could the model somehow

- If the answer to the previous question was positive, do you have a process in place to detect this phenomenon (called model drift in the machine learning literature), so you can take appropriate measures?

Applications

- Usage of the model
  - Are you aware of how the model you are developing will be used? This only applies if the model is not going to be used by the same company that develops it.
  - If you are not sure how the model will be used, have you identified whether (in the worst case), the model could be used in a way that can potentially harm someone?
- Monitoring of the model
  - If you do not keep ownership of the model, are there appropriate measures in place to monitor drift?
  - If not, then is the new owner of the model aware of this phenomenon?
- Model privacy and security
  - Is it possible for someone to use the model in a way that they can identify personal features? This is called a model inversion attack.
  - Would someone be able to “trick” the algorithm into making wrong decisions? This is called an adversarial attack.

We are on a mission to educate and help decision-makers understand and implement data science and AI……

Join Tesseract Academy