Artificial Intelligence (AI) Ethics
Workshop for Nonprofits

December 2020

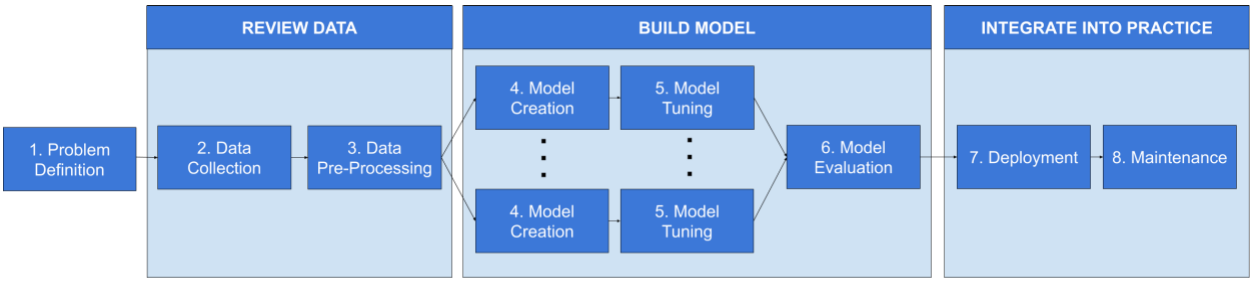
Key AI Ethics Concepts

Overview

- AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies. (Alan Turing Institute)
- Values are broad beliefs held by individuals or groups that reflect concepts of social and cultural importance and norms of appropriate behavior. Ethical values define our moral conduct and help us to determine what is right or what is wrong. An example of a value is Integrity, another one is Respect.
- Principles guide responsible innovation by providing direction for how to embed values in design and use of solutions. Example principles: Fairness, 'Do No Harm".
- Ethical technology solution is a solution that supports individual and collective well-being and enhances our ability to tackle global challenges.
- Responsible Innovation is a transparent, interactive, sustainable process by which organizations proactively evaluate how they can design and use technology in ways that are aligned with their values and missions.

Overview of the Machine Learning Process

The basic steps for the ML process are shown below - in practice, the process is iterative, and steps are revisited until the desired outcome is achieved.



1. **Problem definition** is where the team defines the objectives and the data needed to address the objectives.
2. **Data collection** is where the team consolidates different data, either collected internally or from external sources.
3. **Data pre-processing** is where the team cleans data and labels data (in supervised learning), preparing it to be used by model.
4. **Model creation** is the core technical step where the team selects and develops potential models using the preprocessed data. Data is split into a training set for building the models and test set for validating the models.
5. **Model tuning** is where appropriate threshold values and hyperparameters are set to optimize the model's performance.
6. **Model evaluation** is where models are tested against predefined criteria (such as accuracy, performance, etc) to determine which approach is best suited for the problem.
7. **Deployment** is where the model is used in real-world applications, ideally starting with a smaller beta testing phase.
8. **Maintenance** involves constantly checking the model to make sure it works as intended, revisiting earlier phases and/or retraining the models when new data comes in.

Ethical considerations with the focus on principle of Fairness

What are some of the fairness concerns?

How might data and ML model implementation cause disproportionate harm?

- **Equity** refers to the extent to which an ML model may disproportionately benefit or harm some individuals or groups more than others.
- **Representativeness** refers to whether the data used to develop AI/ML models is representative of the regions, communities, and contexts that will be affected by their use.
- **Bias** refers to systematically favoring or disfavoring different groups based on erroneous assumptions. Bias is defined in terms of attributes such as gender, economic standing, or ethnicity, among others. Consider different types of bias that may be present and how they affect the equity of ML/AI outcomes. Often bias will be embedded in data unintentionally as an artifact of the power dynamics that exist in the world.

How well do we understand how ML models work?

- **Explainability** refers to the extent to which individual predictions made by an ML model can be communicated in terms non-technical experts can understand.
- **Auditability** refers to the extent to which an AI/ML model's decision-making processes and recommendations can be queried by external actors or made transparent to a broader community of actors. Audits can sometimes help to identify concerns about equity, representativeness, and explainability.

What happens when things go wrong?

- **Accountability** refers to whether there are mechanisms in place to ensure that someone will be responsible for responding to feedback and redressing harms if necessary.

How can we address some of these concerns?

- Asking the right question
- Defining the protected attributes, making sure that data and outcomes are not correlated with them
- Identifying sources of bias (historical biases, individual biases, biases in data)
- Technical approaches to testing for bias in data
- Reviewing/strengthening data for representativeness
- Implementing fairness algorithmically
- Diversifying team of people working on AI/ML solutions
- Auditing model outcomes

What are some of the other concerns?

- What are the ethical concerns outside of fairness?
- Is AI/ML a good fit for my development problem?
- How is AI/ML better than existing approaches?
- How will the AI/ML approach align with organization structure and value?
- What other resources and capabilities may be needed to effectively implement AI/ML?