Artificial Intelligence Ethics: 3rd lesson – Identifying Bias in AI

Machine Learning (ML) has the potential to improve lives, but it can also be a source of harm. ML applications have discriminated against individuals on the basis of race, sex, religion, socioeconomic status, and other categories. In this tutorial, you’ll learn about bias, which refers to negative, unwanted consequences of ML applications, especially if the consequences disproportionately affect certain groups.

We’ll cover six different types of bias that can affect any ML application. Then you’ll put your new knowledge to work in a hands-on exercise, where you will identify bias in a real-world scenario.

Bias complexity:

Many ML practitioners are familiar with “biased data” and the concept of “*garbage in, garbage out*”. For example, if you’re training a chatbot using a dataset containing anti-Semitic online conversations (“garbage in”), the chatbot will likely make anti-Semitic remarks (“garbage out”). This example details an important type of bias (called historial bias, as you’ll see below) that should be recognized and addressed.

This is not the only way that bias can ruin ML applications. Bias in data is complex. Flawed data can also result in ***representation bias*** (covered later in this tutorial), if a group is underrepresented in the training data. For instance, when training a facial detection system, if the training data contains mostly individuals with lighter skin tones, it will fail to perform well for users with darker skin tones. A third type of bias that can arise from the training data is called measurement bias, which you’ll learn about below.

And it’s not just biased data that can lead to unfair ML applications: as you’ll learn, bias can also result from the way in which the ML model is defined, from the way the model is compared to other models, and from the way that everyday users interpret the final results of the model. Harm can come from anywhere in the ML process.

Types of bias:

Once we’re aware of the different types of bias, we are more likely to detect them in ML projects. Furthermore, with a common vocabulary, we can have fruitful conversations about how to mitigate (or reduce) the bias. We will closely follow a research paper from early 2020 that characterizes six different types of bias.

* ***Historical bias***

Historical bias occurs when the state of the world in which the data was generated is flawed.

* ***Representation bias***

Representation bias occurs when building datasets for training a model, if those datasets poorly represent the people that the model will serve.

* ***Measurement bias***

Measurement bias occurs when the accuracy of the data varies across groups. This can happen when working with proxy variables (variables that take the place of a variable that cannot be directly measured), if the quality of the proxy varies in different groups.

* ***Aggregation bias***

Aggregation bias occurs when groups are inappropriately combined, resulting in a model that does not perform well for any group or only performs well for the majority group. (this is often not an issue, but most commonly arises in medical applications).

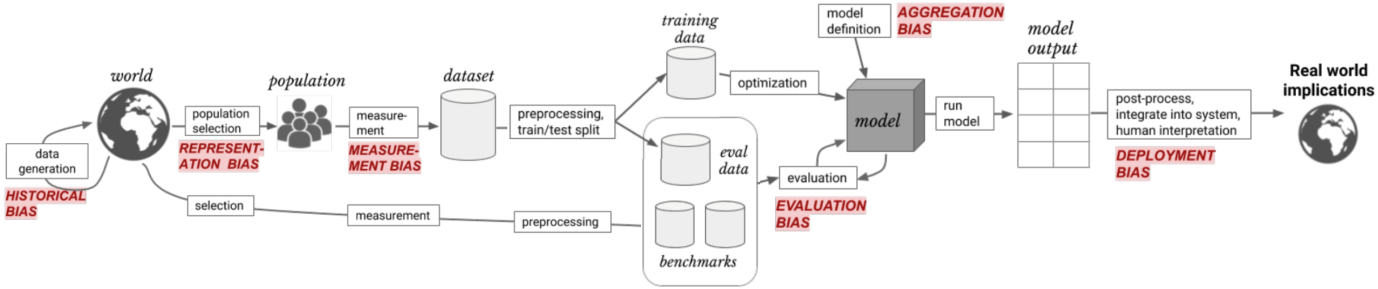
* ***Evaluation bias***

Evaluation bias occurs when evaluating a model, if the benchmark data (used to compare the model to other models that perform similar tasks) does not represent the population that the model will serve.

* ***Deployment bias***

Deployment bias occurs when the problem the model is intended to solve is different from the way it is actually used. If the end users don’t use the model in the way it is intended, there is no guarantee that the model will perform well.

We can visually represent these different types of bias, which occur at different stages in the ML workflow.



Note that these are not mutually exclusive: that is, an ML application can easily suffer from more than one type of bias. For example, as Rachel Thomas describes in a recent research talk, ML applications in wearable fitness devices can suffer from:

* Representation bias: if the dataset used to train the models exclude darker skin tones
* Measurement bias: if the measurement apparatus shows reduced performance with dark skin tones
* Evaluation bias: if the dataset used to benchmark the model exclude darker skin tones