***Bias***

* It refers to the negative, unwanted consequences of ML applications, especially if the consequences disproportionately affect certain groups.
* **Garbage in - Garbage out** - 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”).

## ***Different type of biases:***

### **Historical Bias :**

* + - Occurs when the state of the world in which the data was generated is flawed.
    - Ex: When data is collected from social media, where the basic trend is to antagonize Islamic religion and Black population, it will negatively affect the model we train using that data.

### **Representation Bias :**

* + - Occurs when building datasets for training a model, if those datasets poorly represent the people that the model will serve.
    - Ex: When data collected has under-represented data for few of the targeted groups like targeting old-aged people for mobile advertising, the model may have disastrous effects.

### **Measurement Bias** :

* + - Occurs when the accuracy of the data varies across groups.
    - Ex: Building a medical diagnosis model based on the overall health care costs may be unintentionally biased based on race color even though it’s not a factor for the model as the overall health care cost of Black people are lesser compared to White people due to them being ‘*less trustworthy*’ population while filing for insurance.

### **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.
    - Ex: Different ethnic group may have different rates of Diabetes, which means it always better to have the models sensitive to these differences or building different models for each ethinc groups.

### **Evaluation Bias** :

* + - Occurs when evaluating a model, if the benchmark data does not represent the population that the model will serve.
    - Ex: When the datasets used to train a model like one for facial analysis we use disproportionate data like having lighter-skinned subjects as the majority, it will be biased towards people of color.

### **Deployment Bias** :

* + - Occurs when the problem the model is intended to solve is different from the way it is actually used.
    - Ex: When a model built to predict the likelihood of a criminal relapsing is used to judge a criminal activity and decide appropriate punishment.
* Examples of Bias in real world:
  + A report in the periodical Propublica claimed serious bias against African Americans in a tool to score criminal defendants for recidivism risk.
  + Amazon shut down a model to score candidates for employment after they realized that it penalized women.
  + Predictive policing systems have come under close scrutiny and their use has been curtailed due to discovered biases.
  + Content personalization systems create filter bubbles and ad ranking systems have been accused of racial and gender profiling.

***Understanding Data Bias***

* *“The most important aspect of a statistical analysis is not what you do with the data, it’s what data you use” -* A quote by Andrew Gelman.
* The common definition of data bias is that the available data is not representative of the population or phenomenon of study.
* Data bias occurs due to structural characteristics of the systems that produce the data.
* In a broader sense, Bias also denotes:
  + Data does not include variables that properly capture the phenomenon we want to predict
  + Data includes content produced by humans which may contain bias against groups of people

## ***Common types of Data Bias:***

### **Response/Activity Bias:**

* + - Occurs in content generated by humans: reviews on Amazon, Twitter tweets, Facebook posts, Wikipedia entries,etc.
    - Only a small proportion of people contribute this content and their opinions and preferences are unlikely to reflect the opinions of the population as a whole.
    - Ex:
      * 7% of users produce 50% of the posts on Facebook.
      * 4% of users produce 50% of the reviews on Amazon
      * 0.04% of Wikipedia’s registered editors (about 2000 people) produced the first version of half the entries of English Wikipedia.

### **Selection Bias due to Feedback loops:**

* + - Occurs when a model itself influences the generation of data that is used to train it. This occurs in systems that rank content such as in content and ad personalization, recommender systems which present or give priority to some items over others.
    - Users’ responses to items presented are collected, responses to items not presented are unknown. User responses are also influenced by the position of the items on the page and the details of presentation such as font, media.
    - Ex:
      * Systems for online advertising, content personalization, recommendations, all have built-in feedback loops. These systems embed ML models that influence the data generated, which in turn feeds back into the system as training data for the model.
      * ***Selection bias*** occurs due to the non-random subset of items presented to users.
      * Presenting these items in a ranked list introduces ***position bias*** — since users scan items from left to right and top down
      * ***Presentation bias*** is introduced if fonts and media types vary across items.

### **Bias due to System Drift:**

* + - Drift refers to changes over time to the system generating the data. Changes include the definition of the attributes captured in the data (including the outcome) or the underlying model or algorithm that changes how users interact with the system.
    - The addition of new modes of user interaction such as like or share buttons, addition of search assist feature.
    - Ex: GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States. One of the key reasons behind the failure is due to changes Google periodically makes to its search interface

### **Omitted Variable Bias:**

* + - Occurs in data in which critical attributes that influence the outcome are missing.
    - This typically happens when data generation relies on human input or the process recording the data does not have access to key attributes.
    - Ex: A laptop manufacturer has an online chat system which its customers can use for support requests or to ask questions. The manufacturer wants to use the opportunity to cross-sell products and has developed a model to score users on how likely they are to buy additional products. The score is intended help agents working the chat system to allocate their time efficiently. When they are busy agents put in more effort (and time) trying to cross-sell to users with high-scores and less effort on users with lower scores. However, the time (and effort) expended by the agents is not recorded. Without this data it will appear that the scoring model is performing very well, whereas the time spent by agents might be much better explanation for user purchase decisions.

### **Societal Bias:**

* + - Occurs in content produced by humans, whether it be social media content or curated news articles.
    - The use of gender or race stereotypes. This type of bias can be considered a form of ***label bias****.*
    - Ex:
      * Google News articles exhibit female/male gender stereotypes. For example: females were associated with professions of nurse and nanny whereas males were associated with professions of doctor and financier.
      * Amazon tried to build an AI tool to screen candidates until management discovered that it had learned to penalize women candidates. The problem is that in most companies today, technical roles are filled by men and this bias creeps into any models that use current employee data to train models.

***Tools to Detect and Eliminate Bias***

* If the data you work with has some inherent biases, the model will not only learn those biases but will end up amplifying them.

## ***Different tools used for detecting and eliminating bias:***

### **What-if (by Google):**

* + - To check if your machine learning model is biased or not, you will need to ask many questions and test different scenarios within your data.
    - ***What-If*** is an open-source interactive tool that makes it easier for everyone — even non-programmers — to test, explore and debug machine learning models. What-If gives you the ability to manipulate data points, edit them, generate plots, and specify criteria to evaluate your model, all using a clear and simple GUI.

### **AI Fairness 360 (by IBM):**

* + - An open-source and comprehensive toolkit for both the detection and elimination of bias in machine learning models.
    - ***AI Fairness 360*** includes over 70 fairness metrics that help you detect bias in your models, such as Manhattan and Euclidean.
    - The tool also includes over 10 algorithms that help you eliminate the basis if you found one. These bias mitigating algorithms include optimizing the preprocessing stage, prejudice remover, and regular and more.

### **Crowdsourcing (by Microsoft):**

* + - Used crowdsourcing to precisely detect [bias in natural language processing](https://arxiv.org/pdf/1812.08769.pdf) applications.
    - Crowdsourcing is a term used to describe the practice of engaging people — crowd — to innovate, solve a problem, or increase efficiency. Using crowdsourcing can be used to look into different categories of the problem to identify potential causes of bias.
    - Using crowdsourcing to detect bias in machine learning applications was inspired by the Implicit Association Test (IAT).

### **Local Interpretable Model-Agnostic Explanations (LIME):**

* + - It is a tool that is used to generate explanations for the different machine learning models' behavior.
    - ***LIME*** allows you to manipulate the different components of your model so that you can gain a better understanding of it and be able to point out the source of bias if one exists.

### **FairML (Python toolbox):**

* + - It is a Python open-source toolbox that is used to audit machine learning predictive models to detect bias.
    - FairML was built and developed to answer the question, how much does a specific input affect the performance of a model? The ability to test your model's performance easily using different sets of input data can lead you to detect the existence of bias in your model.
    - FairML offers an end-to-end tool that allows you to test your model performance by quantifying specific inputs' relative significance.

# **References:**

* <https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai/tutorial>
* <https://towardsdatascience.com/survey-d4f168791e57>
* <https://towardsdatascience.com/5-tools-to-detect-and-eliminate-bias-in-your-machine-learning-models-fb6c7b28b4f1>
* <https://towardsdatascience.com/preventing-machine-learning-bias-d01adfe9f1fa>