Artificial Intelligence and Biases

Abdelwahid Benslimane
wahid.benslimane@gmail.com

I) Introduction

Today, artificial intelligence (AI)/machine learning (ML) is omnipresent, sometimes even in our daily lives without us realizing it: personal voice assistants integrated into our smartphones or devices like Alexa, personalized recommendations on streaming platforms, computer intrusion detection and software protection (antivirus, etc.), surveillance and security (facial recognition, detection of suspicious behaviours, incident detection, prevention of failures, etc.), and more. In the coming years, we can expect AI to play an even more prominent role, particularly in fields like healthcare.

These advancements rely on models which are trained using large to very large datasets to perform tasks with a high degree of confidence. However, the data used for training these models can introduce a critical element known as bias. The biases can lead models to make harmful decisions, even perpetuating discrimination or stereotypes or simply failing to provide satisfactory results. We can recall the unfortunate example of Tay, the autonomous Twitter account created by Microsoft (https://en.wikipedia.org/wiki/Tay_(chatbot)) which began to make offensive remarks.

In this document, I aim to explain the main categories of bias, according to me, based on the technical source of the problem (sampling bias, measurement bias, omitted variable bias, algorithmic bias).

Finally, I will discuss some ways to reduce, if not eliminate, these biases.

II) Biases

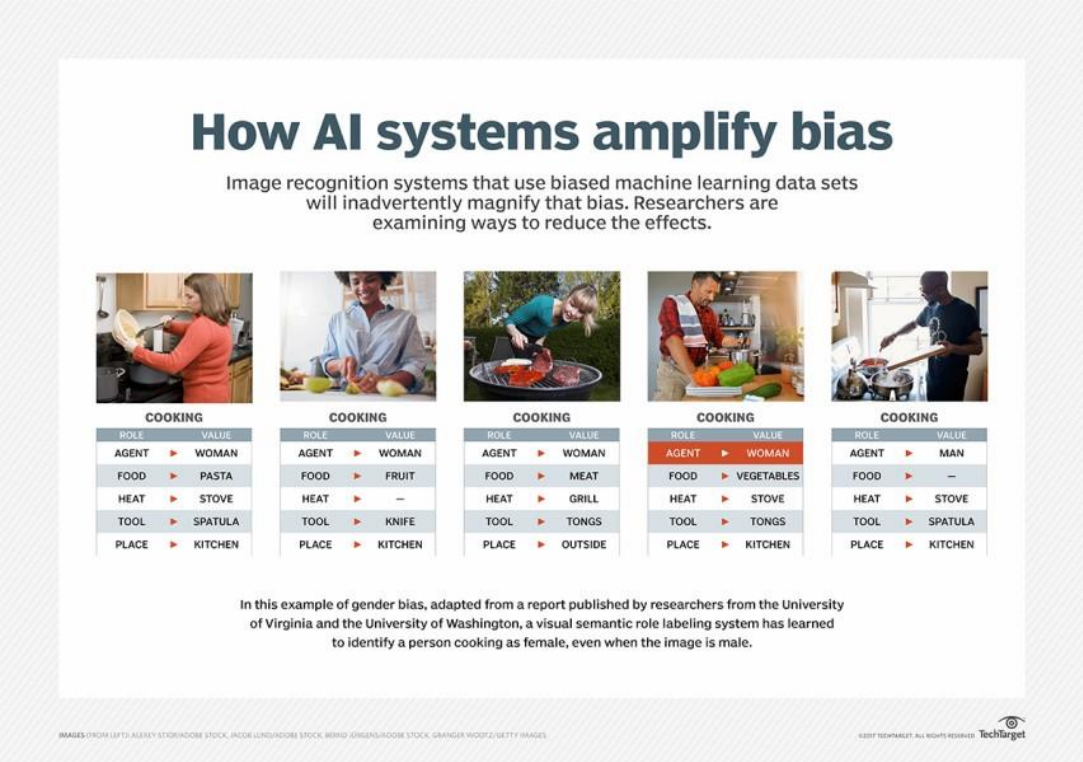
Sampling Bias: This occurs when different samples in the dataset are not representative of the domain to which we want to apply the model. In some cases, it may not be necessary for the samples to match real-world proportions, and intentionally introducing a sampling bias may be preferable to avoid replicating unfortunate situations. For example, if an American company wants to train a model to select the best candidate for a leadership position, such as a CEO or COO, with a dataset that includes a list of CEOs or COOs in major companies in the country and may contain information about the candidates' ethnic background or gender. Without a diverse dataset, the model is likely to consistently select a white male candidate, excluding non-white individuals and women.

However, sampling bias, if not corrected, can be problematic in cases like marketing studies where certain consumer groups may be underrepresented or overrepresented, making it impossible to draw conclusions about the entire population.

Image recognition models are frequently trained using public image datasets, including ImageNet. As explained in this article on the Stanford University website, “more than 45% of ImageNet data […] comes from the United States, home to only 4% of the world’s population. By contrast, China and

India together contribute just 3% of ImageNet data, even though these countries represent 36% of the world's population. This lack of geodiversity partly explains why computer vision algorithms label a photograph of a traditional US bride dressed in white as 'bride', 'dress', 'woman', 'wedding', but a photograph of a North Indian bride as 'performance art' and 'costume'."

The image below provides another illustration of a bias amplified by AI.



Measurement bias: It results from systematic errors during the collection of numerical or even qualitative data, which are linked to imperfections in the measuring instruments or inappropriate conditions during the measurement process. This occurs, for example, when the measuring instruments used, such as humidity or light sensors, are faulty, or when the approach is not consistent throughout the data collection process. For instance, in the evaluation of athletes' performance in a specific discipline, if the evaluators (if the evaluation involves a subjective component) or the evaluation scale vary. It can introduce skewed correlations between the explanatory data and the data we are trying to predict.

Omitted variable bias: The omission of a pertinent variable refers to the exclusion of a critical element from the description of statistical individuals, resulting in systematic errors and predictive inconsistencies. A straightforward example can best illustrate this bias.

Let's consider, for instance, a scenario where a company aims to predict employee churn using a ML/statistical model. It's reasonable to assume that the decision to quit is, at least partially, tied to how long an employee has been with the company. If the data collected on employees do not encompass their length of service, even when other variables, such as salary range, commute time, and family situation, also partially contribute to explaining the resignation rate, the outcome may yield false results.

Sometimes, certain variables may be intentionally omitted to precisely avoid situations of discrimination. For instance, variables indicating a person's gender, country of origin, or religion may be excluded from the dataset. However, a model can still discern these details using other data like title (Mr, Miss, Mrs, Ms, Sir), name, or surname if they are included in the dataset.

Algorithmic bias: An algorithmic bias is a more general term that specifically refers to the failure to account for certain scenarios or categories of individuals during the execution of an algorithm, which can manifest as discrimination or racial profiling, for example. It often stems from the technical biases mentioned earlier but can also result from conscious or unconscious cognitive biases of the designers of intelligent systems (or intended to be) who infuse their own worldview into the model-building process. For example, in the case of reinforcement learning in AI (to understand what reinforcement learning is, you can refer to this document on training of LLMs, although reinforcement learning applies not only to LLMs).

One well-known case of algorithmic bias is the one highlighted by Joy Buolamwini, an American-Guinean computer scientist and activist, who, during her research at MIT, was able to demonstrate that early facial recognition systems did not accurately recognize dark or black skin. The report of her research can be found here.

III) How to fight biases

The fight against biases, at least those of technical origin, necessarily involves data verification, which means implementing a data quality process. Such a process should ensure good representation of different population categories. Another solution is to assign different “weights” to each category to rebalance the dataset. An underrepresented category would then have a greater impact on the model.

Joy Buolamwini suggests making the data open and verifiable by third parties through her website at the Algorithmic Justice League in the movement towards equitable and accountable AI, where you can seek assistance if you are affected by an AI-related injustice. The site offers various services such as algorithm audits, legal assistance, and publishes testimonials from individuals who have been victims of AI-related issues.

There are also technical solutions, such as a toolkit designed for researchers and academics, which IBM contributes to, called AI Fairness 360, to enable them to create bias detection solutions.