4 Types of Machine Learning Bias

ALEGION™

INTRODUCTION

AI is far from infallible. Whether it's autonomous vehicle accidents or facial recognition mishaps, it's tempting for the public to think that AI can't be trusted.

As a company that specializes in training AI systems, we know that models in fact do precisely what they are taught to do.

AI models comprise algorithms and data, and they are only as good as their underlying mathematics and the data they are trained on.

When things go wrong with AI it's for one of two reasons: either the model of the world at the heart of the AI is flawed, or the algorithm driving the model has been insufficiently or incorrectly trained.

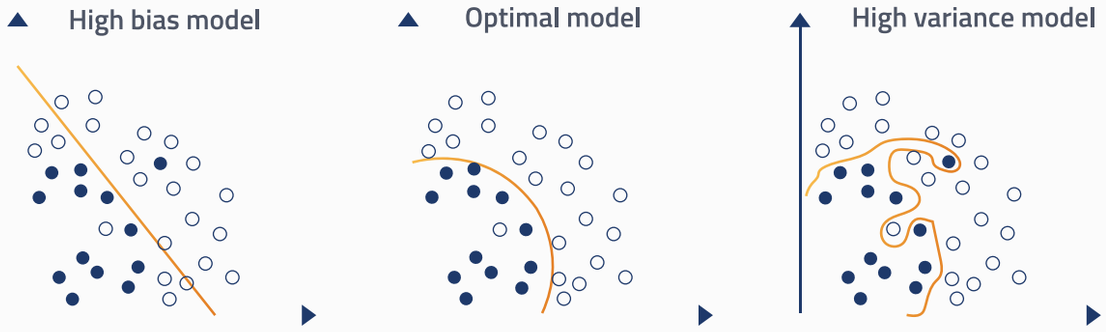
Bias in one form or another is behind many algorithm and data issues. If not mitigated, bias will cause the model to behave - or misbehave - in ways that reflect the bias.

In our experience there are four distinct kinds of bias that data scientists and AI developers need to be aware of and guard against.

From this paper AI project leads and business sponsors will better understand the four distinct types of bias that can affect machine learning, and how each can be mitigated.

1 | ALGORITHM BIAS

This first kind of bias actually has nothing to do with data. Instead, it refers to a property of the AI algorithm itself.



When used in the context of machine learning, the word **bias** has a different meaning. For data scientists, bias, along with variance, describes an algorithm property that influences prediction performance.

Bias and variance are interdependent, and data scientists typically seek a balance between the two.

As you can see here, models with high variance tend to flex to fit the training data very well. They can more easily accommodate complexity, but they are also are more sensitive to noise and may not generalize well to data outside the training data set.

Models with high bias are rigid. They are less sensitive to variations in data and may miss underlying complexities. At the same time, they are more resistant to noise.

Finding the appropriate balance between these two properties for a given model in a given environment is a critical data science skill set.

The data science discipline has developed a lot of maturity round this topic. Optimizing prediction error in machine learning across the bias-variance trade-off is well understood.

Unlike algorithmic bias, the other three sources of AI bias are all found in the system's training data.

BIAS IN THE DATA

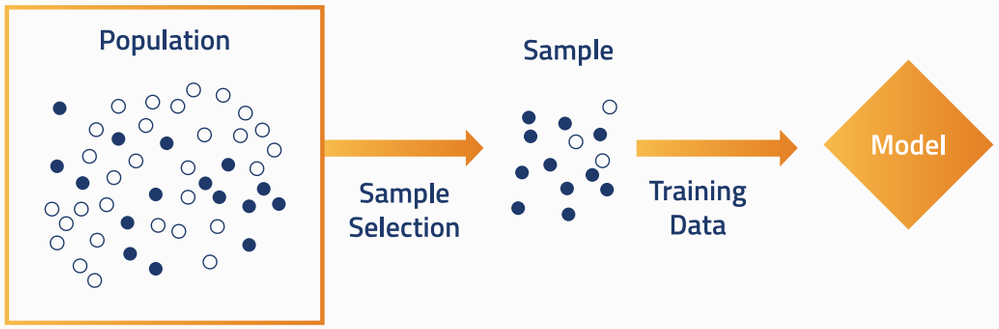
When used in the context of training data, bias returns to its popular connotation.

All three forms of data bias can be overcome with established techniques and methodologies.

Importantly, however, correcting bias in the data takes place in the data. It does not involve algorithm adjustments by data scientists.

2 | SAMPLE BIAS

Sample bias occurs when the data used to train the model does not accurately represent the problem space the model will operate in.



There are a variety of techniques for both selecting samples from populations and validating their representativeness. There are a number of techniques for identifying population characteristics that need to be captured in samples, and for analyzing a sample's fit with the population.

In other words, like algorithmic bias, mitigating sample bias is a technique that is well-understood across multiple disciplines, including psychology and social sciences. Data science teams would do well to look to these disciplines if they lack experimental sampling expertise.

To cite an obvious but illustrative example, if an autonomous vehicle is expected to operate in the daytime and at night, but is trained only on daytime data, its training data is said to reflect sample bias. The model driving the vehicle is highly unlikely to learn how to operate at night with such incomplete and unrepresentative training data. interdependent, and data scientists typically seek a balance between the two.

3 | PREJUDICAL BIAS

Prejudicial bias occurs when training data content is influenced by stereotypes or prejudice coming from the population.

This kind of bias tends to dominate the headlines around AI failures, because it touches on salient cultural and political issues outside of automation.

It becomes an issue when data scientists or the organizations that employ them do not want the model to learn, and then manifest, behaviors that echo these prejudices.

For example, an algorithm that is exposed to annotated images of people at home and at work could deduce that mothers are female. This, of course, would be true in both the sample data and the overall population.

However, if thought isn't given to the images that are introduced to the algorithm it could also deduce that nurses are female. This could happen because in reality – and in random samples of photos of people at work – nurses are statistically more often female than male.

But even if the population of nurses today is overwhelmingly female, it is not true that nurses are female in the way that mothers are female. And we may deem it inappropriate for the algorithm to produce results that incorrectly infer a causal relationship. Mitigating prejudicial bias requires insight into the ways that prejudice and stereotyping can make their way into data. It also requires forethought about the goals and acceptable behavior of a particular AI application.

Mitigating prejudicial bias requires insight into the ways that prejudice and stereotyping can make their way into data. It also requires forethought about the goals and acceptable behavior of a particular AI application.

Addressing this form of bias typically requires placing constraints on input (training) data or outputs (results).

So, for example, a model will not conclude that all nurses are female if it is exposed to images of male nurses in numbers that are disproportionate to what can be found in the workplace. A chatbot that has learned hate speech can be constrained to stop using it. And the humans who label and annotate training data can be trained to avoid introducing their own societal prejudices or stereotypes into the training data.

The Model "Learns" from the Data that Both of These Are True

Mother → **Female**

Nurse → ~~**Female**~~

The Second Is Hinted at by the Data, but is Not True

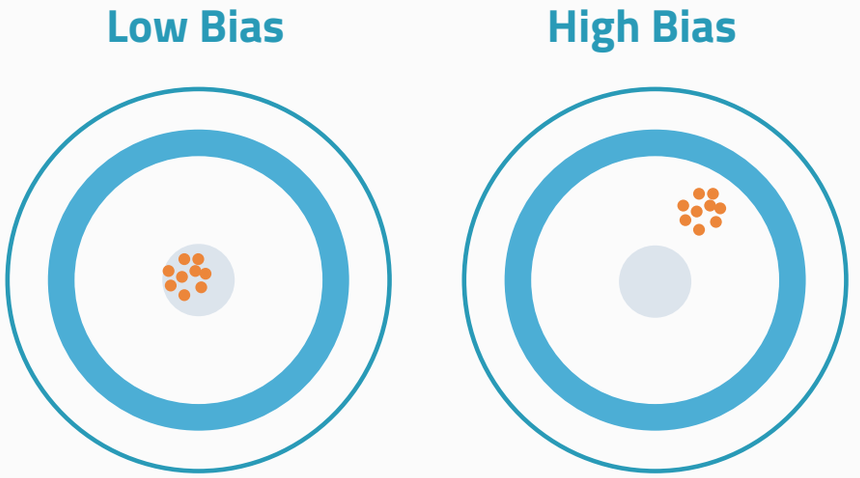
4. MEASUREMENT BIAS

This kind of bias results from faulty measurement. The outcome is a systematic distortion of all the data.

The distortion could be the fault of a device. For example, a camera with a chromatic filter will generate images with a consistent color bias. An 11-7/8 inch long "foot ruler" will always overrepresent lengths.

It could also stem from badly designed data collection. A survey with leading questions will influence responses in a consistent direction; and the output of a data labeling tool may inadvertently be influenced by workers' regional phraseology.

As with sample bias, there are established techniques for detecting and mitigating measurement bias. It's good practice to compare the outputs of different measuring devices, for example. Survey design has well-understood practices for avoiding systematic distortion. And it's essential to train labeling and annotation workers before they are put to work on real data.



"Algorithm and data-driven products will always reflect the design choices of the humans who built them, and it's irresponsible to assume otherwise."

- Fred Benenson, Kickstarter

IGNORE AI BIAS AT YOUR PERIL

AI models and algorithms are built by humans, and the data that train these algorithms are assembled, cleaned, labeled and annotated by humans. The sometimes poorly fit math built into algorithms seeks out patterns in the sometimes biased data. The results, predictably, are not always what the designers intended.

Not all data science teams have the skills in-house to avoid and mitigate training data bias. New data science teams, especially, may struggle with this. And those teams that by virtue of training or experience do have the necessary skills still find that it's a lot of work.

ALEGION'S ROLE

Alegion often is called in when a data science team is dealing with issues of bias and the team understands that offloading these issues can be beneficial.

Sadly, teams that are new to machine learning tend to bring us in late in their projects, when they have exhausted 80% of their budget and still have a model operating at unacceptably low confidence levels.

More seasoned teams tend to bring us in much earlier because experience has taught them that offloading the management of data bias is more efficient.

ABOUT ALEGION

Alegion provides ground truth training data for machine learning initiatives. Our offering operates at massive scale, combining a data and task management software platform with a global pool of trained data specialists.

We assist data science teams throughout the AI life cycle, delivering custom training datasets, providing human-scored model testing, and making available human-in-the-loop exception handling.

With our white glove level of service we completely offload these activities, freeing data professionals to focus on their areas of specialization.

We support machine learning projects broadly, with particular emphasis on Computer Vision, Natural Language Processing and Entity Resolution, in retail, financial services, defense, technology and manufacturing.