ALEGION

Responsible AI: How to Mitigate Bias in Your Training Data

Responsible AI: How to Mitigate Bias in Your Training Data

Introduction

The sheer variety and possibility of artificial intelligence applications is exciting. What if we could expedite recruitment and hiring using a candidate screening tool, analyze a large collection of records to evaluate potential criminal recidivism, or determine the patients most likely to benefit from specific healthcare programs?

These applications and many more are already at work within our organizations and communities. So what's the issue? Unless these AI tools are very carefully designed, they can perpetuate and scale all sorts of biases, with disastrous unintended consequences.

Today, bias in AI is one of the top challenges for the industry. Bias in the context of machine learning is also quite nuanced. There are hundreds of different types of bias (confirmation, association, recall, exclusion, group attribution, measurement, etc) and a quick Google search turns up a wide variety of different classifications and lists.

At Alegion, our data scientists authoritatively group all the hundreds of types of bias into four major categories. In this paper, we will explore and define the four categories of bias and explain how each of them can be mitigated.

Bias in the Data

Almost all popularly-reported examples of machine learning model bias arise from bias in the data used to train those models.

At the end of this paper, we also discuss bias-variance trade-offs that data scientists can make by modifying the algorithm. Importantly, however, correcting most bias takes place in the training data itself and does not involve algorithm adjustments by data scientists.

There are three categories of data bias that can be overcome with established techniques and methodologies.

Mitigating Sample Bias

What Is 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.

To cite an obvious but illustrative example, if an autonomous vehicle is expected to operate in the daytime and at night, but is trained primarily 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.

How to Address Sample Bias

There are a variety of techniques for both selecting samples from populations and validating their representativeness. These techniques are widely used across multiple disciplines including experimental sciences, medical studies, and social sciences.

These techniques include inspecting the data through data distribution analysis and visualizations, using probabilistic sampling procedures, collecting more data, and limiting the application of the model to the scope of the training data.

Mitigating Prejudicial Bias

What Is Prejudicial Bias?

Prejudicial bias occurs when training data content is influenced by stereotypes or prejudice coming from the population and/or from human annotators. This kind of bias tends to dominate the headlines around AI failures, because it touches on salient cultural and political issues outside of automation. Mitigating this type of bias is crucial 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 and their roles 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.

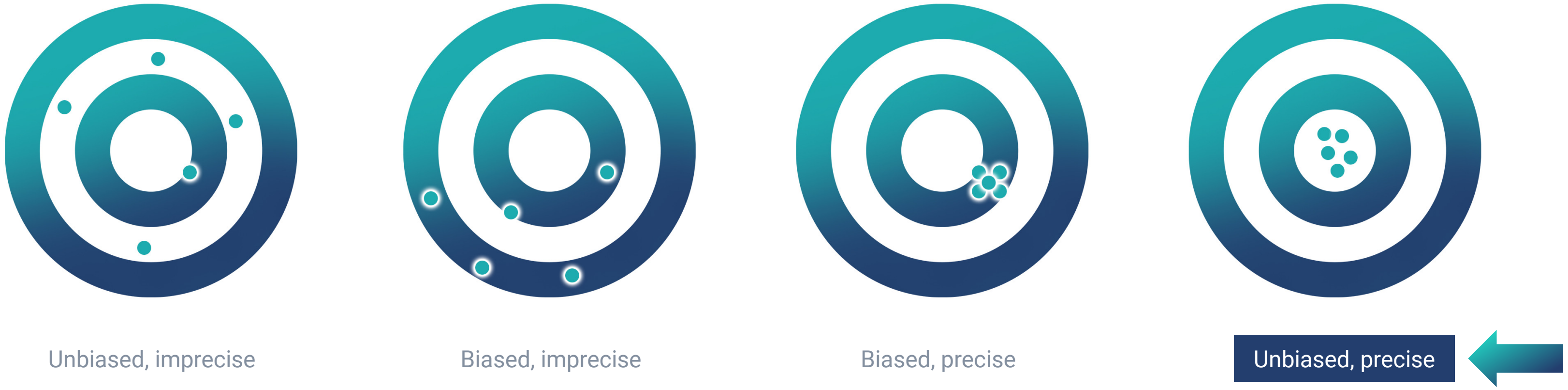
Mitigating Prejudicial Bias

How to Address Prejudicial Bias

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.

Mitigating Measurement Bias



Unbiased, imprecise

Biased, imprecise

Biased, precise

Unbiased, precise

What Is Measurement Bias?

This kind of bias results from faulty measurement. The outcome is a systematic distortion of all the data, and the distortion could be the fault of a device or a human cognitive bias.

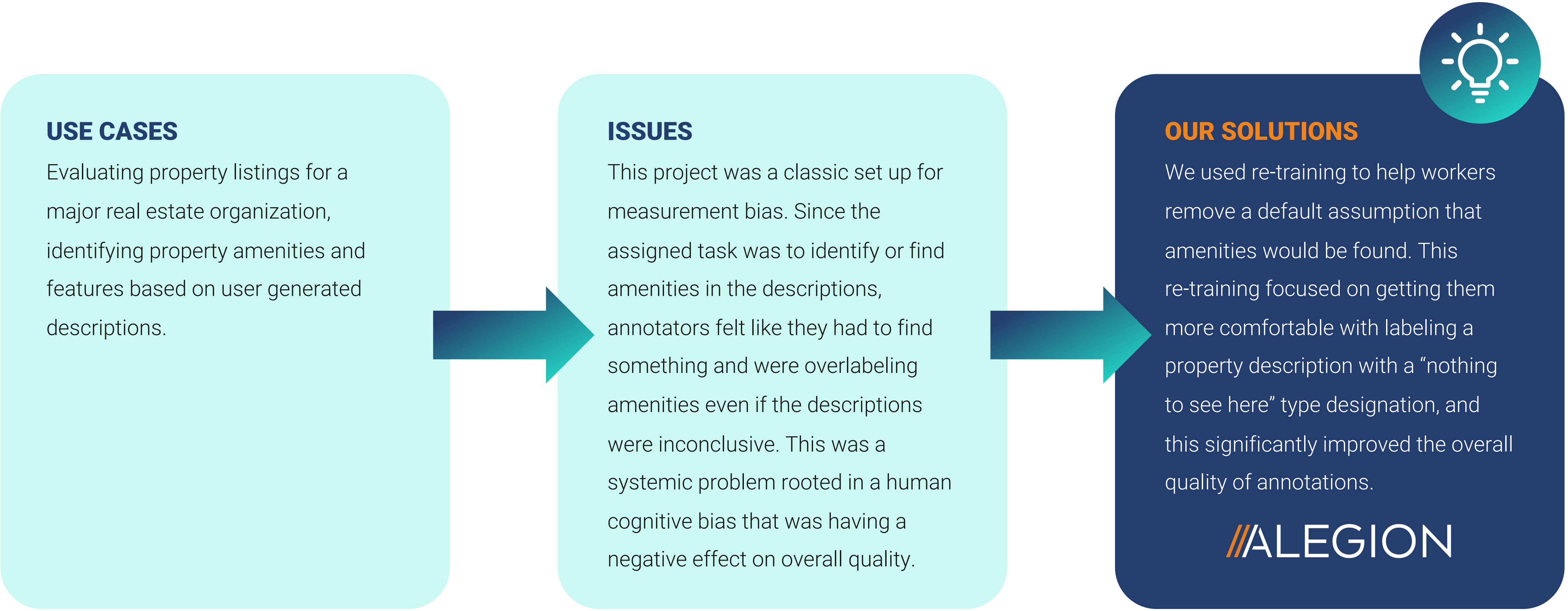
For example, a camera with a chromatic filter will generate images with a consistent color bias. An 11-7/8 inch long “foot ruler” will always overrepresent lengths. It could also stem from badly designed data collection: a survey with leading questions will influence responses in a consistent direction and the output of a data labeling tool may inadvertently be influenced by workers’ regional phraseology or societal factors.

How to Address Measurement Bias

As with sample bias, there are established techniques for detecting and mitigating measurement bias. It’s good practice to diversify methodologies and 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 to avoid systematic errors and the influence of cognitive biases.

How We Eliminated Measurement Bias for One Customer



USE CASES

Evaluating property listings for a major real estate organization, identifying property amenities and features based on user generated descriptions.

ISSUES

This project was a classic set up for measurement bias. Since the assigned task was to identify or find amenities in the descriptions, annotators felt like they had to find something and were overlabeling amenities even if the descriptions were inconclusive. This was a systemic problem rooted in a human cognitive bias that was having a negative effect on overall quality.

OUR SOLUTIONS

We used re-training to help workers remove a default assumption that amenities would be found. This re-training focused on getting them more comfortable with labeling a property description with a “nothing to see here” type designation, and this significantly improved the overall quality of annotations.

ALEGION

Bias in the Algorithm

Mitigating Algorithm Bias

What Is Algorithm Bias?

Now that we've looked at the three kinds of bias that can affect training data, it's worth mentioning a fourth kind of bias that can affect machine learning models. Importantly, this kind of bias has nothing to do with data or to popular connotations of the word "bias"; it refers to a property of the AI algorithm itself.

For data scientists, bias, along with variance, describes an algorithm property that influences prediction performance.

How to Address Algorithm Bias

As model properties, bias and variance are interdependent, and data scientists typically seek a balance between the two. Models with high variance tend to flex to fit the training data very well. They can more easily accommodate complexity, but they are also 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 context is a critical data science skill set. The data science discipline has developed a lot of maturity around this topic. Optimizing prediction error in machine learning across the bias-variance trade-off is well understood.

A data scientist can reduce the algorithmic bias in a model using well understood techniques like training more iterations or switching to a different or bigger model (i.e., more features, more nodes).

Alegion: A Trusted Partner for Responsible AI

Artificial intelligence has the capacity to transform the world as we know it. We are seeing the impact of machine learning spill out across industries and fields, from finance to retail to criminal justice. We must address issues of bias in our training data and algorithms precisely because of how powerful AI will be for growth and development in the coming decades. As the famous line from Spider-Man reminds us, "with great power comes great responsibility."

At Alegion, we are dedicated to the highest quality training data and to responsible AI. We care about how our customers will develop and deploy algorithms, and we engage deeply in order to ensure that your training data is free of bias, prejudicial or otherwise.

Outsourcing training data development to an expert partner like Alegion can help address the crucial issues of bias, saving you time, money, and negative consequences for both performance and reputation.

ALEGION

When you partner with Alegion, you get:

- ✓ Access to a highly trained global workforce of data labeling specialists
- ✓ More freedom and time back for your internal specialists to think about other aspects of model performance
- ✓ Access to our expertise in complex use cases in Computer Vision, Natural Language Processing, and Entity Resolution for industries including retail, financial services, defense, technology and manufacturing.

*Running into issues of bias in your training data? Not getting the results you need from your model? Reach out to us today at **solutions@alegion.com** to learn how we can support your unique use case.*