Algorithmic Bias

What’s the Impact of Automated Decision-Making?

Machine learning creates worlds of possibilities for prediction. Every day more companies adopt machine learning algorithms as alternatives to traditional systems. These automated systems bring considerable benefits. But what if an algorithm is biased or otherwise flawed? In this lesson, you'll learn about algorithmic bias—what it is, how to identify it, and how to avoid it.

Algorithmic Bias and Data Ethics

Data ethics researchers use the term **algorithmic bias** to refer to situations in which a computer system makes decisions that impact different people (or different groups of people) in different ways. Let's break down this term by examining what its two words mean. We’ll start with “algorithmic,” which describes an algorithm.

What’s an Algorithm?

You already know that an algorithm is a set of logical, sequentially ordered steps. An algorithm might be simple, such as a set of instructions that sorts a list of numbers from the highest to the lowest. An algorithm might also be complex, such as an algorithm that is used to categorize emails as spam or not spam.

Whether an algorithm is simple or complex, we frequently apply one to a set of data to process that data in some way. For example, we might apply a sorting algorithm to a list of words to alphabetize them. Or, we might apply a machine learning algorithm to a dataset of plant characteristics to group together the plants that have common traits. Note that **machine learning** is a type of artificial intelligence (AI) that you'll learn more about later in this boot camp.

But to understand algorithmic bias, it’s helpful to know the types of algorithms. We’ll learn about those next, and we’ll then examine what we mean by bias.

Types of Algorithms

To consider how algorithms might impact different people in different ways, researcher Nicholas Diakopoulos, in Algorithmic Accountability: On the Investigation of Black Boxes from Columbia Journalism Review, categorized them into the following four types:

- Prioritization
- Classification
- Association
- Filtering

These types are based on how they help us understand data. Let’s examine each in more detail.

Prioritization

Prioritization algorithms help you figure out where to direct your attention. These algorithms sort or rank data based on factors that the algorithm design specifies.

For example, say that your friend has a T-shirt with "Data Champion" printed on it that you love. You're not much of a T-shirt person, so you want to find a tote bag that has the same printing. You go to a crafts site and type "data champion tote bag" into the search box. The site then lists products for you in a particular order—that is, it prioritizes the available products. Ideally, this list goes in order of relevance to what you're searching for. It might also promote products that either come from more reliable sellers or earn more money for the site.

Classification

Classification algorithms help you sort information into categories. The algorithm design might include criteria that determine the categories. Or, the algorithm might automatically create the categories based on common characteristics.

For example, say that you feel disappointed with the data-themed accessory selection on the crafts site. So, you raise all your entrepreneurial energy to create your own line of products. To make money from this venture, you need a small loan to fund a bulk order of Data Champion tote bags. You apply for this loan at an AI lending company by filling out a form that includes information about your business proposal, your income, and your credit history. The company reviews this information and immediately approves your loan with a 4% interest rate. It does this by using a classification algorithm to sort loan applications into categories based on the predicted likelihood of timely repayment. That is, it classifies loan applications. Because the information that you shared in your application resembled those from people who previously repaid their loans on time, the algorithm classified you as a good candidate, so it approved your loan request. The classification algorithm also determined the interest rate, which is designed to optimize profits for the company.

Association

Association algorithms connect pieces of information that are typically either related or used at the same time. Such an algorithm is usually based on a large amount of existing information.

For example, say that you want to find out the types of "Data Champion" products that your customer base might want. You realize that when you start typing text into a search engine, you usually receive an autocomplete suggestion. You confirm that these suggestions are based on what other users typically search for. So, you decide that this provides an excellent way to gain insight into your market. You search for "data champion" in a few search engines to get their autocomplete suggestions.

Filtering

Filtering algorithms either remove or separate information that meets a certain criterion. These algorithms resemble classification algorithms. But instead of sorting many items into categories, they evaluate a binary criterion.

For example, say that your Data Champion apparel and accessory business is up and running, but you're getting feedback from customers that they're not getting the emails you've been sending to confirm that they like their purchases. So, you send test emails to some friends and family members, and you discover that their spam filters consistently separate your emails from others by placing them in the spam folder. After some online investigation, you figure out what's been happening. As a humorous nod to overly enthusiastic infomercials, you've been including the phrase, "Act now, don't hesitate!" at the end of your emails. Unfortunately, spam emails commonly use this phrase, so your emails aren't making it through.

What Do We Mean by "Bias"?

We usually define **bias** as a situation in which one group, person, or thing is treated differently than other groups, people, or things, respectively.

For example, say that you grew up in a family that always had a beloved hamster as a pet. As an adult, you decide to get a pet rodent, but because of space constraints, you can get only one. When you choose your pet, you'll likely choose a hamster over a rat because of your positive past experiences with hamsters. In this case, you're biased in favor of hamsters.

Your partner further influences your decision. He grew up in New York City and had a scary childhood experience in the subway when a rat stole his pizza. He's been scared of rats ever since. This makes him biased against rats.

People often discuss bias in terms of fairness or unfairness. Additionally, this discussion often occurs in the context of highly personal experiences of bias related to gender, race, or class. But as the pet rodent example shows, bias can exist in many contexts. And in the abstract, it's not necessarily good or bad.

Before moving on, practice reflecting on how you make decisions:

> Think through all the decisions that you made today. Examples include what you had for breakfast and how you chose to get to work. Another example is whether you took a shower, took a bath, or just washed your face when you were getting ready this morning. How long did it take you to make these decisions? Can you identify the factors that went into them?

It's not always easy to know why we make the decisions that we do. This can make identifying biased decision-making challenging. We often have to make decisions quickly, and we might have to choose between options that seem basically equal. In these situations, we tend to make quick decisions based on our past experiences, and we're not always aware of the rationale behind our decisions.

Unconscious bias occurs when we make biased decisions based on factors that we're not aware of. This isn't a big deal if, say, you're unfairly biased against oatmeal because the Quaker Oats man scared you as a child. But if you're reviewing resumes for data analysts, and you subconsciously dislike candidates who attended University of Texas because of an old football rivalry, that might have more serious consequences.

Now that we've learned what algorithmic bias means in terms of data ethics, let's learn how to identify algorithmic bias.

Identifying Algorithmic Bias

As we've learned, making quick, biased decisions (which might have significant impacts on others) is easy. And even quick decision-makers move at a slow pace compared to computers. So, when an algorithm is biased, the harms that result from that bias will quickly accumulate. Moreover, if no one specifically checks for algorithmic bias, it can be hard to notice. That's because a computer doesn't have the capacity to reflect and say, "Wow, I've rejected six qualified candidates who went to the same university. Maybe I should reconsider this."

So, how can we identify algorithmic bias? Well, algorithmic bias occurs in different ways depending on the type of algorithm and how it's being used. To explore this further, we'll examine some famous examples of algorithmic bias.

Example: Discrimination in Online Ad Delivery

In the 2013 article named Discrimination in Online Ad Delivery, Dr. Latanya Sweeney (whom you might remember for data privacy work), demonstrated that the personalized ads Google showed alongside a person's name differed greatly depending on the name.

Dr. Sweeney found that compared to searches including White-identifying names, such as Geoffrey and Emma, those including Black-identifying names, such as Darnelle and Jermaine, were much more likely to show ads suggesting that the person had been arrested. The ForeverData website shows examples of the types of ads that Dr. Sweeney studied.

Question: When do you think people are most likely to search for your name online? This typically happens when they're trying to figure out who you are. Maybe they met you at a conference and want to work with you. Or maybe you've applied for a job at their company or they're about to go on a first date with you. In all these situations, you want to appear at your best—or at least, accurately appear as you are. So, how would you feel if the search results erroneously suggested that you'd been arrested?

Answer: Answers will vary but might include angry, frustrated, or disgusted.

Example: Automated Disqualification from SNAP

The Supplemental Nutrition Assistance Program (SNAP) is a United States Department of Agriculture (USDA) program that assists low- and no-income people with purchasing food. People can use SNAP benefits as payment for certain types of food at certain stores. For grocery stores in low-income neighborhoods, SNAP purchases might provide the majority of their revenue. This was the case for three Somalian markets in Seattle that were disqualified from the SNAP program in 2002, forcing customers to shop elsewhere.

Why were these stores disqualified? The reason is that the USDA used a computer program to detect irregular purchasing patterns and flag them for fraud, and the program noticed that purchases at these grocery stores tended to be hundreds of dollars. Furthermore, the amounts tended to be numbers of dollars without cents, such as $150. By contrast, similar stores had an average transaction of $13.83.

The USDA fraud detection program didn't consider the ways that different cultures shop for food. In the Somalian immigrant community in Seattle, families would pool their SNAP resources and make bulk purchases of goods that would last them for an entire month. For example, they might request $100 worth of beef that they could then divide and freeze. This entirely legal, strategic use of resources caused these grocery stores to be effectively shut down. The store owners lost most of their business because the majority of purchases were made by using SNAP benefits. And, they faced criminal charges. Although the charges were eventually dropped and the stores were allowed to resume accepting SNAP payments, the law prevented the store owners from taking legal action against the USDA to recoup lost sales.

To read the initial reporting about the 2002 case, refer to USDA disqualifies three Somalian markets from accepting federal food stamps. To read about the eventual resolution, refer to USDA drops case against embattled Somali grocers . Although the USDA said that they made changes to address this issue, reports of a similar problem occurred in 2018.

Example: Gender Shades

Researchers at MIT completed the Gender Shades project to evaluate three facial recognition programs. These programs, which IBM, Microsoft, and Face++ created, promised to classify photos of faces by gender. These companies are three of the major ones that create facial recognition products.

Note: according to the Gender Shades website, the researchers used male and female categories for gender because the products that they were evaluating provided only binary sex labels for gender classification. They write that their research "inherits these sex labels and this reduced view of gender which is a more complex construct."

To test the programs, the researchers assembled a library of images of people. They labeled these images according to both skin shade and the binary male/female classification that the commercial programs used. They then used the three programs to label the gender, comparing the generated labels with the actual categories.

Although all three programs had a high overall accuracy, they performed better on lighter-skinned subjects than on darker-skinned subjects. And, they performed much worse on darker-skinned females.

The Gender Shades project found that high overall accuracy scores can be deceptive. If a product is trusted as accurate—but is accurate only in limited situations, such as identifying the gender of lighter-skinned men—any harmful decisions that result from errors in the product will rest disproportionately on specific groups, such as darker-skinned women. For statistics and more details about the Gender Shades project, refer to Gender Shades.

Now that we've observed some results of algorithmic bias in practice, we'll learn about some causes of algorithmic bias.

Causes of Algorithmic Bias

It can be difficult to determine the causes of algorithmic bias. That's because teams of company employees create most software programs. The relevant decisions and responsibilities are distributed across a wide range of people and often made over lengthy spans of time. Furthermore, algorithms can be complex and act in ways that their developers don't understand. This is particularly true for certain types of AI models, such as neural networks, that make connections between data points in ways that people find challenging to understand.

Still, in cases of algorithmic bias, two factors come up again and again:

- The people who developed the algorithm come from different backgrounds than those impacted by the algorithm.

- The algorithm was trained with biased training data.

We can notice the first factor in the earlier example of the Somalian markets. Somalian immigrants make up a minority in the US, so they were unlikely to be well represented on the development team that created the SNAP fraud detection algorithm. Because these immigrants used the system differently than the majority of other users, and differently than the developers anticipated, the algorithm was biased against them.

This factor also likely applied to the bias that the Gender Shades researchers found. Tech companies are overwhelmingly staffed by lighter-skinned men. And as we learned earlier, we all make decisions that reflect our backgrounds, whether or not we realize it. So, unconscious bias might cause the developers of an algorithm to make decisions that reflect their backgrounds. Hence, that's probably why the facial recognition algorithms worked better on lighter-skinned men than on anyone else.

Let's now consider the second factor: biased training data. **Training data** refers to a labeled dataset that's used to teach an algorithm how to make decisions. The algorithm bases its decisions on the patterns that it identifies regarding how the data points match their labels.

Recall that the Gender Shades project tested three facial recognition algorithms. And each of these algorithms was trained with a dataset—specifically, with one that associated faces with their genders. Say that these training datasets disproportionately included lighter-skinned faces that were labeled male. In that case, the algorithms would have had more information about the facial patterns of lighter-skinned males than about the other types of people. This would then make the error rate worse for the other types of people by comparison.

Note that, because the Gender Shades project evaluated commercial products, we don't have transparency about how those products labeled the genders. They might have used a purchased database of images with their associated genders, for example. Or, they might have scraped online data from labeled sources. In any case, the issue doesn't concern the accuracy of the training data labels. Instead, it concerns the distribution of the types of people that the training dataset includes, and how that impacts prediction accuracy.

Now that we've learned about the causes of algorithmic bias, you might be wondering how we can address this bias. Let's learn about that next.

Addressing Algorithmic Bias

Algorithmic bias is an enormous problem that's caused harm to many people. But, it's also receiving lots of scholarly attention. Researchers have identified many ways that we can identify algorithmic bias, prevent it, and mitigate any harms that it might cause.

One way is to use an **auditing system**—that is, a proactive review of an algorithm to seek out bias. Two types of auditing systems exist: external audits and internal audits.

The Gender Shades project provides an example of an **external audit**. That is, the researchers had neither a relationship to the companies that created the algorithms nor any insight into how they worked. Instead, they used their own data to seek out bias.

A company can also develop a system to complete an **internal audit**. That is, the system will regularly test an algorithm for different types of bias. This testing occurs when the algorithm is both in development and in use.

Nicholas Diakopoulos, who categorized algorithms into the types that we discussed earlier, recommends making algorithmic systems more transparent (Accountability in Algorithmic Decision Making). Specifically, they should be transparent about the following:

- The source of the data that it's processing.
- The specific information that the algorithmic uses and how it prioritizes that information.
- Standards for what to expect as a result from the algorithm.

This transparency and documentation can help both external and internal audits. But transparency has limits. In many cases, companies resist transparency to protect their intellectual property. Regarding the release of data, privacy is an issue, and with certain complex types of models, the amount of transparency that's possible also has limits.

Finally, a movement exists to make algorithmic systems **contestable**. Doing so means supplying ways to contest, or disagree with, the findings of an algorithmic system.

In some cases, developers can build contestability into the system for users. For example, Netflix allows you to give a thumbs up or a thumbs down to movies and TV shows that it recommends. A thumbs down is a way to contest a result of the recommendation algorithm, which predicted that you'd like that movie or show. Other systems might need a more-formal and human-supported process for contesting results.

Question: Imagine that a computer system automatically rejects you for a loan without providing any insight into how or why it made this decision. What might you do to contest the decision?

Answer: Answers will vary. But, you might first try to find a person to ask about the decision and then try to get the decision reversed if you felt that it was made unfairly.

Now that we've learned about ways to address algorithmic bias, wouldn't it be nice to have a checklist for avoiding algorithmic bias? That's what we'll explore next.

Checklists for Avoiding Algorithmic Bias

To avoid algorithmic bias in an existing system, you might want to use the following checklist:

- Investigate how the system works. Was it trained by using historical data? If so, think about biases that the historical data might contain. Also, seek out evidence that the developers have controlled for biased training data.

- Research similar systems to find out if they include examples of bias.

- Try auditing the system by testing it to find out if the accuracy of its results vary according to your input.

- Check that a process for contesting the results of the system exists.

To avoid algorithmic bias in a system that's in development, you might want to use the following checklist:

- Consider the types of people that your system will serve or impact. How do these people compare to those who are developing the system? Make sure that your team is aware of any differences, and teach them to seek out potentially biased decisions.

- Carefully review your training and testing datasets to find out if they underrepresent or overrepresent any groups of people.

- Consistently document your work in plain language so that others in your organization can both review and understand it. As much as possible, choose modeling techniques that people can understand.

- Create a process for internally auditing your system, and identify someone outside your development team who can test that process.

The topic of algorithmic bias is growing increasingly important as more systems replace human decision-making with algorithmic decision-making. Now that you know what it is, be on the lookout for systems that could be biased.