NO PREP

What Does "Biased AI" Actually Mean?

MORE LESSONS LESS PLANNING · MLLP · FREE TODAY · FREE TOMORROW · FREE FOREVER

Includes: Student reading, case study, differentiated questions, teacher key

Name: Period: Date:

What Does "Biased AI" Actually Mean?

Not Just a Glitch — A Pattern

When people hear that an AI system is "biased," they sometimes picture a bug in the code — a line of programming that someone accidentally wrote wrong. But AI bias is almost never that simple. Bias in artificial intelligence refers to a systematic and repeatable pattern in which an AI produces outputs that are unfair or inaccurate for certain groups of people. It is not usually a single mistake; it is a built-in

Name: Period: Date:

What Does "Biased AI" Actually Mean?

Not Just a Glitch: A Pattern

When people hear that an AI system is "biased," they sometimes picture a bug in the code, a line of programming that someone accidentally wrote wrong. But AI bias is almost never that simple. Bias in artificial intelligence refers to a systematic and repeatable pattern in which an AI produces outputs that are unfair or inaccurate for certain groups of people. It is not usually a single mistake; it is a built-in tendency that shows up consistently, often without anyone intending it. Understanding what causes this tendency requires looking at where AI learns in the first place: data. AI systems, particularly those based on machine learning, learn by analyzing enormous datasets: sometimes billions of examples. A language model might be trained on hundreds of millions of web pages. An image classifier might study millions of photographs. The AI detects patterns in all that information and uses those patterns to make predictions or decisions. The problem is that the data it learns from was created by human beings, who live in societies with long histories of discrimination, inequality, and prejudice. When those patterns exist in the training data, the AI absorbs them and reflects them back as if they were natural rules about the world.

Three Types of Bias You Should Know

Researchers who study AI fairness have identified several distinct types of bias. The first is historical bias. This occurs when the data accurately reflects the real world, but the real world itself contains discrimination. For example, if an AI is trained on historical hiring records from a company where women were rarely promoted to senior roles, the AI may learn to deprioritize women's applications, not because anyone programmed it to discriminate, but because discrimination was embedded in the history it studied. The second type is representation bias. This

happens when certain groups are underrepresented or missing from the training data entirely. If a facial recognition system is trained mostly on photos of light-skinned individuals, it will become much less accurate when it encounters darker-skinned faces. The AI was not told to perform worse on those faces; it simply had far fewer examples to learn from. The third type is measurement bias, which occurs when the data itself is collected in a flawed way. If a predictive policing algorithm uses arrest rates as a proxy for crime rates, it is measuring police activity, not actual crime; police activity has historically been distributed unequally across communities.

The Feedback Loop Problem

One of the most troubling features of biased AI systems is that they can make bias worse over time. Imagine a hiring algorithm that learns from past hiring data. That data reflects years of discrimination against certain groups. The algorithm begins rejecting candidates from those groups at higher rates. Fewer candidates from those groups get hired. Future data now shows even lower representation. The algorithm trains again on this new data and becomes even more biased. This cycle is called a feedback loop, and it is one reason why bias in AI systems can compound and accelerate far faster than bias in purely human decision-making. Social media recommendation algorithms illustrate a similar problem. If users from certain communities are shown certain types of content more often, they click on it more. The algorithm interprets these clicks as evidence that those users want more of that content. More is shown. The cycle continues. The AI is optimizing for engagement, which sounds neutral, but the actual result can be that entire communities are systematically pushed toward certain categories of information: a form of bias that shapes what people believe and how they see the world.

Why Neutral-Sounding Systems Are Not Always Neutral

One of the most common misconceptions about AI is that because computers run on math and logic, they must be objective. This belief is dangerous. A credit-scoring algorithm might not mention race at all — but if it uses ZIP codes, and certain ZIP codes are predominantly populated by one racial group due to decades of discriminatory housing policy, the algorithm is effectively using race as a factor even without naming it. This is called proxy discrimination. The variable looks neutral on the surface, but it carries demographic information in disguise. Researchers at MIT and other institutions have demonstrated that many commercially deployed AI

systems perform significantly worse for women and darker-skinned individuals even when those characteristics are not directly included in the model. In a 2018 study called "Gender Shades" by researcher Joy Buolamwini and Timnit Gebru, commercial facial recognition systems from major technology companies showed error rates as high as 34.7 percent for darker-skinned women, compared to less than 1 percent for lighter-skinned men. The systems had not been programmed to discriminate — they had been trained on unrepresentative data and deployed without adequate testing.

Who Is Responsible, and What Can Be Done?

Fixing AI bias is not a simple technical patch. It requires decisions about values: What does fairness mean? Does it mean equal accuracy across all groups? Equal false-positive rates? Equal outcomes? Different answers lead to different technical solutions, and some of those solutions are mathematically incompatible with each other. This means that reducing bias in AI is as much an ethical and political question as a technical one. Researchers, governments, and advocacy groups have proposed several approaches. Diverse training data is a starting point — systems that are trained on data representing a wider range of people, geographies, and circumstances tend to perform more equitably. Bias audits, conducted by independent third parties before AI systems are deployed in high-stakes settings like hiring, lending, or criminal justice, are another tool. Transparency requirements — forcing companies to explain how their systems make decisions — help advocates and regulators identify problems. Some countries, including members of the European Union, have begun requiring human oversight of AI decisions that significantly affect individuals. But none of these solutions is perfect, and none removes the fundamental challenge: AI systems learn from human history, and human history contains a great deal of injustice. Looking ahead, some researchers argue that the solution to AI bias requires not just technical fixes but structural changes in how AI systems are developed. This means diversifying the teams that build AI — bringing in people who have themselves experienced discrimination and who can recognize problems that homogeneous teams might miss. It also means establishing external accountability structures: independent review boards with authority to examine AI systems before and after deployment, funded separately from the companies that profit from those systems. The history of anti-discrimination law in the United States suggests that voluntary compliance with

fairness standards has rarely been sufficient; enforcement mechanisms with real consequences have usually been necessary to produce lasting change.

Key Information: AI Bias Basics

Detail	Information
Definition of AI bias	A systematic pattern in which AI produces unfair or inaccurate outputs for certain groups
Primary cause	Training data created by humans in societies with histories of discrimination
Historical bias	AI learns from data that accurately reflects a discriminatory past
Representation bias	Certain groups are underrepresented or missing from training data
Measurement bias	Data is collected in ways that introduce existing inequalities
Feedback loop	Biased outputs become future training data, amplifying the original bias over time
Proxy discrimination	Neutral-seeming variables (like ZIP code) carry hidden demographic information
"Gender Shades" study (2018)	Found error rates up to 34.7% for darker-skinned women in facial recognition systems

Case Study: Maya's Job Application

Maya is 24 years old and recently finished a two-year certification program in medical billing. She spent the past three years working part-time at a community health clinic while raising her younger siblings after her mother got sick. Her resume is strong, but she applies through an online portal at a large hospital network. The portal uses an AI screening tool to rank applicants before any human sees the resumes. Maya is rejected before a recruiter reads a single word she wrote. She does not know why. She calls the HR department. The representative tells her the system flagged her application based on "resume patterns" but cannot explain what that means. A friend who studies computer science tells Maya that screening algorithms often downrank employment gaps and non-traditional work histories. "The system wasn't built for people like us," her friend says. "It was built on data from people who got hired the easy way." Maya is angry but also confused. She did everything right. She got the credential. She did the work. She just did not fit the pattern the machine expected.

1. Based on the reading, what type of AI bias most likely affected Maya's application, and what does that type of bias involve?

2. The reading says that fixing AI bias is "as much an ethical and political question as a technical one." Using Maya's situation as an example, explain what this means.

Multiple Choice

1. According to the reading, what is the PRIMARY source of bias in AI systems?

a) Deliberate programming errors by developers

b) Lack of internet access in rural communities across the country

c) Training data that reflects existing human inequalities

d) Hardware malfunctions in server systems

2. What is a "feedback loop" in the context of AI bias?

a) A quality-control process that catches errors and problems before the system is deployed

b) A method for collecting diverse data to reduce discrimination

c) A legal requirement for companies to explain their algorithms

d) A cycle in which biased outputs become future training data, increasing bias over time

3. What did the 2018 "Gender Shades" study find about commercial facial recognition systems?

a) They had error rates as high as 34.7 percent for darker-skinned women

b) They often refused to process images of individuals who were over age 50

c) They performed worse on men than on women in most cases

d) They were equally accurate for all demographic groups

Matching

Write the letter of the correct definition on the line next to each term.

Term	Definition
1. ___ Historical bias	A) Bias that occurs when certain groups are underrepresented in training data
2. ___ Representation bias	B) When a neutral-seeming variable carries hidden demographic information
3. ___ Proxy discrimination	C) Bias that occurs when training data accurately reflects a discriminatory past
4. ___ Feedback loop	D) Bias introduced when data is collected in a flawed or unequal way
5. ___ Measurement bias	E) A cycle in which biased outputs become future training data, amplifying bias

Fill-in-the-Blank

1. AI bias is a _____ and repeatable pattern in which an AI produces outputs that are unfair for certain groups.

2. When a neutral variable like a ZIP code carries hidden demographic information, researchers call this _____ discrimination.

3. The 2018 study called "Gender _____" revealed high error rates in commercial facial recognition for darker-skinned women.

Short Answer

Answer each question in 2-4 complete sentences.

1. Explain why a statement like "computers are objective because they run on math" is misleading when it comes to AI bias.

2. Describe what representation bias is and give a specific example from the reading that illustrates it.

3. What are two approaches the reading mentions for reducing AI bias, and what are the limitations of those approaches?