Junior Honors Cohort
Darnell, Healy, Houghteling, Jihan

Introducing AI Bias
Reading 1

name: Dentain Davis
period: 1

Part I: AI Biases Presentation.

Directions: Complete the steps below as you discuss AI biases via the presentation.

Define the following terms from the discussion and slides:

1. Training Data - the information given to an AI to help to help it learn how to do a specific task. (input)
2. Testing Data - the information used to check whether the AI that was created reliable and accurate.
3. AI Bias - When an AI that makes a decision that is wrong or problematic because it learned from training data that bias's areas all people, places, and things inaccurately.

Turn-and-Talk 1: Take 60 seconds to brainstorm with your elbow partner - how do you think we could reduce the bias in the AI training of detecting fruit types?

We could add more similar fruits with similar matching colors.

Turn-and-Talk 2: Take 2 minutes to imagine that you have now been asked to create an AI tool that can identify weather conditions, such as sunny, rainy, or cloudy. **Describe the kind of images** you would include as a part of the **training data.** Your goal is to come up with as complete of a data set as possible so that you limit the chance for AI bias!

Water droplets, Plants dripping water droplets, Grey sky, Puddles, Grey Clouds, NO Sun out, rain

Drawing Conclusions: In your own words, and in 2-3 sentences, describe how AI bias occurs.

AI bias occurs when data is not in put correctly. AI bias occurs when the data input is too similar to another to the point where it can't be distinguished accurately by the AI systems.

Part II: Bias in AI Partner Reading

Directions: Annotate as you read the following article with your partner. Specifically, take notes in the space below about what the authors did and what actions they took to capture data about AI. Then, summarize the conclusions of the data with your partner.

- Implicit bias is an unconscious Prejudice we learn from a young age. It shapes how we view people from different social groups.

"Even as we learn to sort shapes and colors... we might perceive dark-skinned men as more dangerous than others."

Are AI Hiring Tools Racist and Ableist?

By Hilke Schellmann

If people with accents or speech impairments are less well "understood" by AI used in hiring, this would constitute discrimination based on disability and national origin, which is illegal in the U.S.

Artificial Intelligence (AI) is now being used in every aspect of the "employee life cycle"—from hiring to firing. For hiring, many companies utilize one-way AI-based interview tools, which send questions to job seekers' devices and have them record their answers without a human on the other side.

The software then transcribes what job applicants say in the video interviews to text. The AI compares the text to job interviews of current employees, who are deemed successful. If applicants use similar words as current employees have used in their job interviews, they will get a favorable score. If they have less overlap with current employees, they are going to get rejected by the AI.

What has rarely been discussed and never studied is if the underlying technology, the speech-to-text transcription, is actually treating everyone fairly and equally. If the speech-to-text transcription process produces a significantly higher "Word Error Rate" (WER) for speakers with accents or speech impairments versus native speakers without speech impairments, this may lead to applicants getting unfairly rejected.

This research is of vital importance, because the results could prove significant flaws in the speech-to-text transcription systems that underpin the AI used in hiring. If the study shows that these tools discriminate against people with accents and speech impairments, this would be illegal in the U.S.

AI Algorithms Objectify Women's Bodies

This project also investigates gender bias by algorithms used by some of the largest platforms, including Google and Microsoft. Our research shows that these algorithms tag photos of women in everyday situations as racy or sexually suggestive at higher rates than images showing men in similar situations. As a result, the social media companies that leverage these algorithms have suppressed the reach of countless images featuring women's bodies, and hurt female-led businesses—further amplifying societal disparities.

Source: Are AI Hiring Tools Racist and Ableist? | Pulitzer Center

Beyond Bias

By Jyoti Madhusoodanan

We all have bias embedded in our brains, but there are ways we can move past it. New findings from psychology show us how.

In a now-classic series of experiments, researchers teased out the deep-rooted nature of human bias simply by distributing red shirts and blue shirts to groups of 3- to 5-year-olds at a day care center. In one classroom, teachers were asked to divide children into groups based on the color of their shirts. In another, teachers were instructed to overlook the shirt colors. After three weeks, children in both classrooms tended to prefer being with classmates who wore the same color as themselves—no matter what the teachers did.

This preference for people who seem to belong to our own tribe forms early and drives our choices throughout life. There appears to be no avoiding it: We are all biased. Even as we learn to sort shapes and colors and distinguish puppies from kittens, we also learn to categorize people on the basis of traits they seem to share. We might associate women who resemble our nannies, mothers, or grandmothers with nurturing or doing domestic labor. Or following centuries of racism, segregation, and entrenched cultural stereotypes, we might perceive dark-skinned men as more dangerous than others.

The biases we form quickly and early in life are surprisingly immutable. Biases are "sticky," says Kristin Pauker, a psychology researcher at the University of Hawaii, "because they rely on this very fundamental thing that we all do. We naturally categorize things, and we want to have a positivity associated with the groups we're in." These associations are logical shortcuts that help us make quick decisions when navigating the world. But they also form the roots of often illogical attractions and revulsions, like red shirts versus blue shirts.

Our reflexive, implicit biases wreak devastating social harm. When we stereotype individuals based on gender, ethnicity, sexual orientation, or race, our mental stereotypes begin to drive our behavior and decisions, such as whom to hire, who we perceive as incompetent, delinquent, or worse. Earlier this year, for instance, an appeals court overturned a Black man's conviction for heroin distribution and the 10-year prison sentence he received in part because the Detroit federal judge who handed down the original verdict admitted, "This guy looks like a criminal to me."

People who live in racially homogeneous environments may struggle to distinguish faces of a different race from one another.

Correcting for the biases buried in our brains is difficult, but it is also hugely important. Because

women are stereotyped as domestic, they are also generally seen as less professional. That attitude has reinforced a decades-long wage gap. Even today, women still earn only 82 cents for every dollar that men earn. Black men are perceived as more violent than white men, and thus are subjected to discriminatory policing and harsher prison sentences, as in the Detroit case. Clinicians' implicit preferences for cisgender, heterosexual patients cause widespread inequities in health care for LGBTQ+ individuals.

"These biases are operating on huge numbers of people repetitively over time," says Anthony Greenwald, a social psychologist at the University of Washington. "The effects of implicit biases accumulate to have great impact."

Greenwald was one of the first researchers to recognize the scope of the problems created by our implicit biases. In the mid-1990s, he created early tests to study and understand implicit association. Along with colleagues Mahzarin Banaji, Brian Nosek, and others, he hoped that shining a light on the issue might quickly identify the tools needed to fix it. Being aware that our distorted thinking was hurting other people should be enough to give pause and force us to do better, they thought.

They were wrong. Although implicit bias training programs help people become aware of their biases, both anecdotal reports and controlled studies have shown that the programs do little to reduce discriminatory behaviors spurred by those prejudices. "They fail in the most important respect," Greenwald says. When he, Banaji, and Nosek developed the Implicit Association Test, he took it himself. He was distressed to discover that he automatically associated more positive words with the faces of white people, and more unfavorable words with people who were Black. "I didn't regard myself as a prejudiced person," Greenwald says. "But I had this association nevertheless."

His experience is not unusual. The Implicit Association Test (IAT) measures the speed of subjects' responses as they match descriptors of people (such as *Hispanic* or *gay*) to qualities (such as attractiveness, athleticism, or being professional). It's based on the idea that people react more quickly when they are matching qualities that are already strongly associated in their minds. Implicit bias exists separately from explicit opinion, so someone who honestly believes they don't have anything against gay people, for instance, may still reveal a bias against them on the test. "A lot of people are surprised by their results," Greenwald says. "This is very hard for people to come to grips with intuitively."

People's beliefs may not matter as much if they can be persuaded not to act on them.

One reason we are so often unaware of our implicit biases is that we begin to form these mental associations even before we can express a thought. Brain-imaging studies have found that six-month-old babies can identify individual monkey faces as well as individual humans. Just a dozen weeks later, nine-month-old babies retain the ability to identify human faces but begin to group all the monkey faces together generically as just "monkey," losing the ability to spot individual features. Shortly after, babies

begin to group human faces by race and ethnicity. Our adult brains echo these early learning patterns. People who live in racially homogeneous environments may struggle to distinguish faces of a different race from one another.

As it became clear how deeply ingrained these biases are—and how they might be unfathomable even to ourselves—researchers began to design new types of strategies to mitigate bias and its impact in society. By 2017, companies in the United States were spending $8 billion annually on diversity training efforts, including those aimed at reducing unconscious stereotyping, according to management consulting firm McKinsey & Company. These trainings range from online educational videos to workshops lasting a few hours or days in which participants engage in activities such as word-association tests that help identify their internalized biases.

Recent data suggest that these efforts have been failing too. In 2019 researchers evaluated the effectiveness of 18 methods that aimed to reduce implicit bias, particularly pro-white and anti-Black bias. Only half the methods proved even temporarily effective, and they shared a common theme: They worked by giving study participants experiences that contradicted stereotypes. Reading a story with an evil white man and a dashing young Black hero, for example, reduced people's association of Black men with criminality. Most of these strategies had fleeting effects that lasted only hours. The most effective ones reduced bias for only a few days at best.

Even when training reduced bias, it did little to reduce discriminatory outcomes. Beginning in early 2018, the New York City Police Department began implicit bias training for its 36,000 personnel to reduce racial inequities in policing. When researchers evaluated the project in 2020, they found that most officers were aware of the problems created by implicit bias and were keen to address these harms, but their behaviors contradicted these intentions. Data on arrests, stops, and stop-and-frisk actions showed that officers who had completed the training were still more likely to take these actions against Black and Hispanic people. In fact, the training program hardly had any effect on the numbers.

This and similar studies have "thrown some cold water on just targeting implicit bias as a focus of intervention," says Calvin Lai, a social psychologist at Washington University in St. Louis. Even if you are successful in changing implicit bias or making people more aware of it, "you can't easily assume that people will be less discriminatory."

But researchers are finding reason for hope.

Although the dozens of interventions tested so far have demonstrated limited long-term effects, some still show that people can be made more aware of implicit bias and can be moved to act more equitably, at least temporarily. In 2016, Lai and his colleagues tested eight ways of reducing unconscious bias in studies with college students. One of the interventions they tested involved participants reading a

vividly portrayed scenario in which a white person assaulted them and a Black person came to their rescue. The story reinforced the connection between heroism and Black identity.

Other interventions were designed to heighten similar connections. For instance, one offered examples of famous Black individuals, such as Oprah Winfrey, and contrasted them with examples of infamous white people, including Adolf Hitler. Participants' biases were gauged using the IAT both before and after these interventions. While the experiments tamped down bias temporarily, none of them made a difference just a few days later. "People go into the lab and do an intervention and there's that immediate effect," Pauker says.

From such small but significant successes, an insight began to emerge: Perhaps the reason implicit bias is stable is because we inhabit an environment that's giving us the same messages again and again. Instead of trying to chip away at implicit bias merely by changing our minds, perhaps success depended on changing our environment.

The implicit associations we form—whether about classmates who wear the same color shirt or about people who look like us—are a product of our mental filing cabinets. But a lot of what's in those filing cabinets is drawn from our culture and environment. Revise the cultural and social inputs, researchers like Kristin Pauker theorize, and you have a much greater likelihood of influencing implicit bias than you do by sending someone to a one-off class or training program.

Babies who start to blur monkey faces together do so because they learn, early on, that distinguishing human faces is more critical than telling other animals apart. Similarly, adults categorize individuals by race, gender, or disability status because these details serve as markers of something we've deemed important as a society. "We use certain categories because our environment says those are the ones that we should be paying attention to," Pauker says.

Just as we are oblivious to many of the biases in our heads, we typically don't notice the environmental cues that seed those biases. In a 2009 study, Pauker and her colleagues examined the cultural patterns depicted in 11 highly popular TV shows, including *Grey's Anatomy*, *Scrubs*, and *CSI Miami*. The researchers tracked nonverbal interactions among characters on these shows and found that even when white and Black characters were equal in status and jobs and spoke for about the same amount of time, their nonverbal interactions differed. For instance, on-screen characters were less likely to smile at Black characters, and the latter were more often portrayed as stern or unfriendly.

Thinking of implicit bias as malleable allows us to constantly reframe our judgments about people we meet.

In a series of tests, Pauker and her colleagues found that regular viewers of such shows were more likely to have stronger anti-Black implicit biases on the IAT. But when the researchers asked viewers

multiple-choice questions about bias in the video clips they saw, viewers' responses about whether they'd witnessed pro-Black or pro-white bias were no better than random. They were being influenced by the bias embedded in the show, "but they were not able to explicitly detect it," Pauker says.

Perhaps the most definitive proof that the outside world shapes our biases emerged from a recent study of attitudes toward homosexuality and race over decades. In 2019 Harvard University experimental psychologist Tessa Charlesworth and her colleagues analyzed the results of 4.4 million IATs taken by people between 2007 and 2016. The researchers found that anti-gay implicit bias had dropped about 33 percent over the years, while negative racial attitudes against people of color declined by about 17 percent.

The data were the first to definitively show that implicit attitudes can change in response to a shifting zeitgeist. The changes in attitudes weren't due to any class or training program. Rather, they reflected societal changes, including marriage equality laws and protections against racial discrimination. Reducing *explicit* discrimination altered the *implicit* attitudes instilled by cultures and communities—and thus helped people rearrange their mental associations and biases.

Until societal shifts occur, however, researchers are finding alternate ways to reduce the harms caused by implicit bias. People's beliefs may not matter as much if they can be persuaded not to act on them. According to the new way of thinking, managers wouldn't just enter training to reduce their bias. Instead, they could be trained to remove implicit bias from hiring decisions by setting clear criteria before they begin the hiring process.

Faced with a stack of resumes that reveal people's names, ethnicities, or gender, an employer's brain automatically starts slotting them based on preconceived notions of who is more professional or worthy of a job. Then bias supersedes logic.

When we implicitly favor someone, we are more likely to regard their strengths as important. Consider, for example, a hiring manager who perceives men as more suited to a role than women. Meeting a male candidate with a low GPA but considerable work experience may lead the manager to think that real-world experience is what really matters. But if the man has a higher GPA and less experience, the manager might instead reason that the latter isn't important because experience can be gained on the job. To avoid this all-too-common scenario, employers could define specific criteria necessary for a role, then create a detailed list of questions needed to evaluate those criteria and use these to create a structured interview. Deciding in advance whether education or work experience matters more can reduce this problem and lead to more equitable decisions. "You essentially sever the link between the bias and the behavior," explains Benedek Kurdi, a psychologist at the University of Illinois Urbana–Champaign. "What you're saying is the bias can remain, but you deprive it of the opportunity to influence decision making."
In the long run, reducing the biases and injustices built into our environment is the only surefire path toward

taming the harmful implicit biases in our heads. If we see a world with greater equity, our internal attitudes seem to adjust to interpret that as normal. There's no magical way to make the whole world fair and equitable all at once. But it may be possible to help people envision a better world from the start so that their brains form fewer flawed associations in the first place.

To Pauker, achieving that goal means teaching children to be flexible in their thinking from an early age. Children gravitate toward same-race interactions by about the age of 10. In one study, Pauker and her colleagues found that offering stories to children that nudged them to think about racial bias as flexible made them more likely to explore mixed-race friendships. In another study, Pauker and team found that children who thought about prejudice as fixed had more uncomfortable interactions with friends of other races and eventually avoided them. But those who thought about prejudice as malleable—believing they could change their minds about people of other races—were less likely to avoid friends of other races. The key, Pauker suggests, is not to rethink rigid mental categories but to encourage mental flexibility. Her approach, which encourages children to consider social categories as fluid constructs, appears to be more effective. The data are preliminary, but they offer a powerful route to change: simply being open to updating the traits we associate with different groups of people.

Thinking of implicit bias as malleable allows us to constantly reframe our judgments about people we meet—evaluating each unique individual for what they are, rather than reducing them to a few preconceived traits we associate with their race, gender, or other social category. Rather than trying to fight against our wariness toward out-groups, reconsidering our mental classifications in this manner allows us to embrace the complexity of human nature and experience, making more of the world feel like our in-group. Blurring the implicit lines in our minds might be the first step to reducing disparities in the world we make.

Source: Beyond Bias | Pulitzer Center. https://pulitzercenter.org/stories/beyond-bias

Part III: Policy Protections.
Imagine that you have the power to write a policy to stop AI bias. Write one policy rule based specifically on the information presented to you today on biases and objectifying women's bodies. Include the following information:

Who has to follow this policy?

What will the policy do (or prevent)? How will you ensure this is done?

Write your policy here.

Junior Honors Cohort
Darnell, Healy, Houghteling, Jihan

Introducing AI Bias
Reading 1

name: Tiara Hicks
period: 1st and 2nd

Part I: AI Biases Presentation.

Directions: Complete the steps below as you discuss AI biases via the presentation.

Define the following terms from the discussion and slides:

1. Training Data - The information given to an AI to help it learn how to do a specific task.
2. Testing Data - The information used to check whether the AI that was created is reliable and accurate.
3. AI Bias - When an AI tool makes a decision that is wrong or problematic because it learned from training data that didn't treat all people, places, and things accurately.

Turn-and-Talk 1: Take 60 seconds to brainstorm with your elbow partner - how do you think we could reduce the bias in the AI training of detecting fruit types?

We could reduce the bias in the AI training of detecting fruit types with more fruits the same color.

Turn-and-Talk 2: Take 2 minutes to imagine that you have now been asked to create an AI tool that can identify weather conditions, such as sunny, rainy, or cloudy. **Describe the kind of images** you would include as a part of the **training data.** Your goal is to come up with as complete of a data set as possible so that you limit the chance for AI bias! Gloomy skies, puddles of water, heavy clouds, and rain droplets, grey clouds.

Drawing Conclusions: In your own words, and in 2-3 sentences, describe how AI bias occurs.

AI bias occurs when looking at colors and shapes, because it's a lot of fruits that are round and also the same color.

Part II: Bias in AI Partner Reading

Directions: Annotate as you read the following article with your partner. Specifically, take notes in the space below about what the authors did and what actions they took to capture data about AI. Then, summarize the conclusions of the data with your partner.

My take away from the bullet points was that implicit bias is an unconscious prejudice that we need to reduce in order for our environment to be more equitable.

[Quote; Our reflexive, implicit biases wreak devasting social harm.]

Are AI Hiring Tools Racist and Ableist?

By Hilke Schellmann

If people with accents or speech impairments are less well "understood" by AI used in hiring, this would constitute discrimination based on disability and national origin, which is illegal in the U.S.

Artificial Intelligence (AI) is now being used in every aspect of the "employee life cycle"—from hiring to firing. For hiring, many companies utilize one-way AI-based interview tools, which send questions to job seekers' devices and have them record their answers without a human on the other side.

The software then transcribes what job applicants say in the video interviews to text. The AI compares the text to job interviews of current employees, who are deemed successful. If applicants use similar words as current employees have used in their job interviews, they will get a favorable score. If they have less overlap with current employees, they are going to get rejected by the AI.

What has rarely been discussed and never studied is if the underlying technology, the speech-to-text transcription, is actually treating everyone fairly and equally. If the speech-to-text transcription process produces a significantly higher "Word Error Rate" (WER) for speakers with accents or speech impairments versus native speakers without speech impairments, this may lead to applicants getting unfairly rejected.

This research is of vital importance, because the results could prove significant flaws in the speech-to-text transcription systems that underpin the AI used in hiring. If the study shows that these tools discriminate against people with accents and speech impairments, this would be illegal in the U.S.

AI Algorithms Objectify Women's Bodies

This project also investigates gender bias by algorithms used by some of the largest platforms, including Google and Microsoft. Our research shows that these algorithms tag photos of women in everyday situations as racy or sexually suggestive at higher rates than images showing men in similar situations. As a result, the social media companies that leverage these algorithms have suppressed the reach of countless images featuring women's bodies, and hurt female-led businesses—further amplifying societal disparities.

Source: Are AI Hiring Tools Racist and Ableist? | Pulitzer Center

Beyond Bias

By Jyoti Madhusoodanan

We all have bias embedded in our brains, but there are ways we can move past it. New findings from psychology show us how.

In a now-classic series of experiments, researchers teased out the deep-rooted nature of human bias simply by distributing red shirts and blue shirts to groups of 3- to 5-year-olds at a day care center. In one classroom, teachers were asked to divide children into groups based on the color of their shirts. In another, teachers were instructed to overlook the shirt colors. After three weeks, children in both classrooms tended to prefer being with classmates who wore the same color as themselves—no matter what the teachers did.

This preference for people who seem to belong to our own tribe forms early and drives our choices throughout life. There appears to be no avoiding it: We are all biased. Even as we learn to sort shapes and colors and distinguish puppies from kittens, we also learn to categorize people on the basis of traits they seem to share. We might associate women who resemble our nannies, mothers, or grandmothers with nurturing or doing domestic labor. Or following centuries of racism, segregation, and entrenched cultural stereotypes, we might perceive dark-skinned men as more dangerous than others.

The biases we form quickly and early in life are surprisingly immutable. Biases are "sticky," says Kristin Pauker, a psychology researcher at the University of Hawaii, "because they rely on this very fundamental thing that we all do. We naturally categorize things, and we want to have a positivity associated with the groups we're in." These associations are logical shortcuts that help us make quick decisions when navigating the world. But they also form the roots of often illogical attractions and revulsions, like red shirts versus blue shirts.

Our reflexive, implicit biases wreak devastating social harm. When we stereotype individuals based on gender, ethnicity, sexual orientation, or race, our mental stereotypes begin to drive our behavior and decisions, such as whom to hire, who we perceive as incompetent, delinquent, or worse. Earlier this year, for instance, an appeals court overturned a Black man's conviction for heroin distribution and the 10-year prison sentence he received in part because the Detroit federal judge who handed down the original verdict admitted, "This guy looks like a criminal to me."

People who live in racially homogeneous environments may struggle to distinguish faces of a different race from one another.

Correcting for the biases buried in our brains is difficult, but it is also hugely important. Because

women are stereotyped as domestic, they are also generally seen as less professional. That attitude has reinforced a decades-long wage gap. Even today, women still earn only 82 cents for every dollar that men earn. Black men are perceived as more violent than white men, and thus are subjected to discriminatory policing and harsher prison sentences, as in the Detroit case. Clinicians' implicit preferences for cisgender, heterosexual patients cause widespread inequities in health care for LGBTQ+ individuals.

"These biases are operating on huge numbers of people repetitively over time," says Anthony Greenwald, a social psychologist at the University of Washington. "The effects of implicit biases accumulate to have great impact."

Greenwald was one of the first researchers to recognize the scope of the problems created by our implicit biases. In the mid-1990s, he created early tests to study and understand implicit association. Along with colleagues Mahzarin Banaji, Brian Nosek, and others, he hoped that shining a light on the issue might quickly identify the tools needed to fix it. Being aware that our distorted thinking was hurting other people should be enough to give pause and force us to do better, they thought.

They were wrong. Although implicit bias training programs help people become aware of their biases, both anecdotal reports and controlled studies have shown that the programs do little to reduce discriminatory behaviors spurred by those prejudices. "They fail in the most important respect," Greenwald says. When he, Banaji, and Nosek developed the Implicit Association Test, he took it himself. He was distressed to discover that he automatically associated more positive words with the faces of white people, and more unfavorable words with people who were Black. "I didn't regard myself as a prejudiced person," Greenwald says. "But I had this association nevertheless."

His experience is not unusual. The Implicit Association Test (IAT) measures the speed of subjects' responses as they match descriptors of people (such as *Hispanic* or *gay*) to qualities (such as attractiveness, athleticism, or being professional). It's based on the idea that people react more quickly when they are matching qualities that are already strongly associated in their minds. Implicit bias exists separately from explicit opinion, so someone who honestly believes they don't have anything against gay people, for instance, may still reveal a bias against them on the test. "A lot of people are surprised by their results," Greenwald says. "This is very hard for people to come to grips with intuitively."

People's beliefs may not matter as much if they can be persuaded not to act on them.

One reason we are so often unaware of our implicit biases is that we begin to form these mental associations even before we can express a thought. Brain-imaging studies have found that six-month-old babies can identify individual monkey faces as well as individual humans. Just a dozen weeks later, nine-month-old babies retain the ability to identify human faces but begin to group all the monkey faces together generically as just "monkey," losing the ability to spot individual features. Shortly after, babies

begin to group human faces by race and ethnicity. Our adult brains echo these early learning patterns. People who live in racially homogeneous environments may struggle to distinguish faces of a different race from one another.

As it became clear how deeply ingrained these biases are—and how they might be unfathomable even to ourselves—researchers began to design new types of strategies to mitigate bias and its impact in society. By 2017, companies in the United States were spending $8 billion annually on diversity training efforts, including those aimed at reducing unconscious stereotyping, according to management consulting firm McKinsey & Company. These trainings range from online educational videos to workshops lasting a few hours or days in which participants engage in activities such as word-association tests that help identify their internalized biases.

Recent data suggest that these efforts have been failing too. In 2019 researchers evaluated the effectiveness of 18 methods that aimed to reduce implicit bias, particularly pro-white and anti-Black bias. Only half the methods proved even temporarily effective, and they shared a common theme: They worked by giving study participants experiences that contradicted stereotypes. Reading a story with an evil white man and a dashing young Black hero, for example, reduced people's association of Black men with criminality. Most of these strategies had fleeting effects that lasted only hours. The most effective ones reduced bias for only a few days at best.

Even when training reduced bias, it did little to reduce discriminatory outcomes. Beginning in early 2018, the New York City Police Department began implicit bias training for its 36,000 personnel to reduce racial inequities in policing. When researchers evaluated the project in 2020, they found that most officers were aware of the problems created by implicit bias and were keen to address these harms, but their behaviors contradicted these intentions. Data on arrests, stops, and stop-and-frisk actions showed that officers who had completed the training were still more likely to take these actions against Black and Hispanic people. In fact, the training program hardly had any effect on the numbers.

This and similar studies have "thrown some cold water on just targeting implicit bias as a focus of intervention," says Calvin Lai, a social psychologist at Washington University in St. Louis. Even if you are successful in changing implicit bias or making people more aware of it, "you can't easily assume that people will be less discriminatory."

But researchers are finding reason for hope.

Although the dozens of interventions tested so far have demonstrated limited long-term effects, some still show that people can be made more aware of implicit bias and can be moved to act more equitably, at least temporarily. In 2016, Lai and his colleagues tested eight ways of reducing unconscious bias in studies with college students. One of the interventions they tested involved participants reading a

vividly portrayed scenario in which a white person assaulted them and a Black person came to their rescue. The story reinforced the connection between heroism and Black identity.

Other interventions were designed to heighten similar connections. For instance, one offered examples of famous Black individuals, such as Oprah Winfrey, and contrasted them with examples of infamous white people, including Adolf Hitler. Participants' biases were gauged using the IAT both before and after these interventions. While the experiments tamped down bias temporarily, none of them made a difference just a few days later. "People go into the lab and do an intervention and there's that immediate effect," Pauker says.

From such small but significant successes, an insight began to emerge: Perhaps the reason implicit bias is stable is because we inhabit an environment that's giving us the same messages again and again. Instead of trying to chip away at implicit bias merely by changing our minds, perhaps success depended on changing our environment.

The implicit associations we form—whether about classmates who wear the same color shirt or about people who look like us—are a product of our mental filing cabinets. But a lot of what's in those filing cabinets is drawn from our culture and environment. Revise the cultural and social inputs, researchers like Kristin Pauker theorize, and you have a much greater likelihood of influencing implicit bias than you do by sending someone to a one-off class or training program.

Babies who start to blur monkey faces together do so because they learn, early on, that distinguishing human faces is more critical than telling other animals apart. Similarly, adults categorize individuals by race, gender, or disability status because these details serve as markers of something we've deemed important as a society. "We use certain categories because our environment says those are the ones that we should be paying attention to," Pauker says.

Just as we are oblivious to many of the biases in our heads, we typically don't notice the environmental cues that seed those biases. In a 2009 study, Pauker and her colleagues examined the cultural patterns depicted in 11 highly popular TV shows, including *Grey's Anatomy*, *Scrubs*, and *CSI Miami*. The researchers tracked nonverbal interactions among characters on these shows and found that even when white and Black characters were equal in status and jobs and spoke for about the same amount of time, their nonverbal interactions differed. For instance, on-screen characters were less likely to smile at Black characters, and the latter were more often portrayed as stern or unfriendly.

Thinking of implicit bias as malleable allows us to constantly reframe our judgments about people we meet.

In a series of tests, Pauker and her colleagues found that regular viewers of such shows were more likely to have stronger anti-Black implicit biases on the IAT. But when the researchers asked viewers

multiple-choice questions about bias in the video clips they saw, viewers' responses about whether they'd witnessed pro-Black or pro-white bias were no better than random. They were being influenced by the bias embedded in the show, "but they were not able to explicitly detect it," Pauker says.

Perhaps the most definitive proof that the outside world shapes our biases emerged from a recent study of attitudes toward homosexuality and race over decades. In 2019 Harvard University experimental psychologist Tessa Charlesworth and her colleagues analyzed the results of 4.4 million IATs taken by people between 2007 and 2016. The researchers found that anti-gay implicit bias had dropped about 33 percent over the years, while negative racial attitudes against people of color declined by about 17 percent.

The data were the first to definitively show that implicit attitudes can change in response to a shifting zeitgeist. The changes in attitudes weren't due to any class or training program. Rather, they reflected societal changes, including marriage equality laws and protections against racial discrimination. Reducing *explicit* discrimination altered the *implicit* attitudes instilled by cultures and communities—and thus helped people rearrange their mental associations and biases.

Until societal shifts occur, however, researchers are finding alternate ways to reduce the harms caused by implicit bias. People's beliefs may not matter as much if they can be persuaded not to act on them. According to the new way of thinking, managers wouldn't just enter training to reduce their bias. Instead, they could be trained to remove implicit bias from hiring decisions by setting clear criteria before they begin the hiring process.

Faced with a stack of resumes that reveal people's names, ethnicities, or gender, an employer's brain automatically starts slotting them based on preconceived notions of who is more professional or worthy of a job. Then bias supersedes logic.

When we implicitly favor someone, we are more likely to regard their strengths as important. Consider, for example, a hiring manager who perceives men as more suited to a role than women. Meeting a male candidate with a low GPA but considerable work experience may lead the manager to think that real-world experience is what really matters. But if the man has a higher GPA and less experience, the manager might instead reason that the latter isn't important because experience can be gained on the job. To avoid this all-too-common scenario, employers could define specific criteria necessary for a role, then create a detailed list of questions needed to evaluate those criteria and use these to create a structured interview. Deciding in advance whether education or work experience matters more can reduce this problem and lead to more equitable decisions. "You essentially sever the link between the bias and the behavior," explains Benedek Kurdi, a psychologist at the University of Illinois Urbana–Champaign. "What you're saying is the bias can remain, but you deprive it of the opportunity to influence decision making."
In the long run, reducing the biases and injustices built into our environment is the only surefire path toward

taming the harmful implicit biases in our heads. If we see a world with greater equity, our internal attitudes seem to adjust to interpret that as normal. There's no magical way to make the whole world fair and equitable all at once. But it may be possible to help people envision a better world from the start so that their brains form fewer flawed associations in the first place.

To Pauker, achieving that goal means teaching children to be flexible in their thinking from an early age. Children gravitate toward same-race interactions by about the age of 10. In one study, Pauker and her colleagues found that offering stories to children that nudged them to think about racial bias as flexible made them more likely to explore mixed-race friendships. In another study, Pauker and team found that children who thought about prejudice as fixed had more uncomfortable interactions with friends of other races and eventually avoided them. But those who thought about prejudice as malleable—believing they could change their minds about people of other races—were less likely to avoid friends of other races. The key, Pauker suggests, is not to rethink rigid mental categories but to encourage mental flexibility. Her approach, which encourages children to consider social categories as fluid constructs, appears to be more effective. The data are preliminary, but they offer a powerful route to change: simply being open to updating the traits we associate with different groups of people.

Thinking of implicit bias as malleable allows us to constantly reframe our judgments about people we meet—evaluating each unique individual for what they are, rather than reducing them to a few preconceived traits we associate with their race, gender, or other social category. Rather than trying to fight against our wariness toward out-groups, reconsidering our mental classifications in this manner allows us to embrace the complexity of human nature and experience, making more of the world feel like our in-group. Blurring the implicit lines in our minds might be the first step to reducing disparities in the world we make.

Source: Beyond Bias | Pulitzer Center. https://pulitzercenter.org/stories/beyond-bias

Part III: Policy Protections.
Imagine that you have the power to write a policy to stop AI bias. Write one policy rule based specifically on the information presented to you today on biases and objectifying women's bodies. Include the following information:

Who has to follow this policy?

What will the policy do (or prevent)? How will you ensure this is done?

Write your policy here.

Junior Honors Cohort
Darnell, Healy, Houghteling, Jihan

Introducing AI Bias
Reading 1

name: Ms. Maxwell
period:

Part I: AI Biases Presentation.

Directions: Complete the steps below as you discuss AI biases via the presentation.

Define the following terms from the discussion and slides:

1. Training Data - The info given to an ai computer to create a thinking process.
2. Testing Data - The info used to check weather the ai that was created is reliable & accurate.
3. AI Bias - Tools that makes a decision that is wrong or problematic b/c it learned from trained data

Turn-and-Talk 1: Take 60 seconds to brainstorm with your elbow partner - how do you think we could reduce the bias in the AI training of detecting fruit types? - Add more fruit that look alike.

Turn-and-Talk 2: Take 2 minutes to imagine that you have now been asked to create an AI tool that can identify weather conditions, such as sunny, rainy, or cloudy. **Describe the kind of images** you would include as a part of the **training data.** Your goal is to come up with as complete of a data set as possible so that you limit the chance for AI bias! Rainy - water, puddles, dark sky, Thunder, gray clouds, umbrella, coat, hat or hood

Drawing Conclusions: In your own words, and in 2-3 sentences, describe how AI bias occurs.

- AI bias occurs when something is told wrong to a computer, which makes the computer give wrong info.

Part II: Bias in AI Partner Reading

Directions: Annotate as you read the following article with your partner. Specifically, take notes in the space below about what the authors did and what actions they took to capture data about AI. Then, summarize the conclusions of the data with your partner.

- Changing our behavior
- Implicit bias is an unconscious prejudice. It's how we view people diff from social groups.
- Changing societal norms / decrease in implicit bias.
- Most effective way to reduce implicit bias.. we need to change our environment to be more equitable.
- Ex: Teaching children to be more accepting people from diff backgrounds

Are AI Hiring Tools Racist and Ableist?

By Hilke Schellmann

If people with accents or speech impairments are less well "understood" by AI used in hiring, this would constitute discrimination based on disability and national origin, which is illegal in the U.S.

Artificial Intelligence (AI) is now being used in every aspect of the "employee life cycle"—from hiring to firing. For hiring, many companies utilize one-way AI-based interview tools, which send questions to job seekers' devices and have them record their answers without a human on the other side.

The software then transcribes what job applicants say in the video interviews to text. The AI compares the text to job interviews of current employees, who are deemed successful. If applicants use similar words as current employees have used in their job interviews, they will get a favorable score. If they have less overlap with current employees, they are going to get rejected by the AI.

What has rarely been discussed and never studied is if the underlying technology, the speech-to-text transcription, is actually treating everyone fairly and equally. If the speech-to-text transcription process produces a significantly higher "Word Error Rate" (WER) for speakers with accents or speech impairments versus native speakers without speech impairments, this may lead to applicants getting unfairly rejected.

This research is of vital importance, because the results could prove significant flaws in the speech-to-text transcription systems that underpin the AI used in hiring. If the study shows that these tools discriminate against people with accents and speech impairments, this would be illegal in the U.S.

AI Algorithms Objectify Women's Bodies

This project also investigates gender bias by algorithms used by some of the largest platforms, including Google and Microsoft. Our research shows that these algorithms tag photos of women in everyday situations as racy or sexually suggestive at higher rates than images showing men in similar situations. As a result, the social media companies that leverage these algorithms have suppressed the reach of countless images featuring women's bodies, and hurt female-led businesses—further amplifying societal disparities.

Source: Are AI Hiring Tools Racist and Ableist? | Pulitzer Center

Beyond Bias

By Jyoti Madhusoodanan

We all have bias embedded in our brains, but there are ways we can move past it. New findings from psychology show us how.

In a now-classic series of experiments, researchers teased out the deep-rooted nature of human bias simply by distributing red shirts and blue shirts to groups of 3- to 5-year-olds at a day care center. In one classroom, teachers were asked to divide children into groups based on the color of their shirts. In another, teachers were instructed to overlook the shirt colors. After three weeks, children in both classrooms tended to prefer being with classmates who wore the same color as themselves—no matter what the teachers did.

This preference for people who seem to belong to our own tribe forms early and drives our choices throughout life. There appears to be no avoiding it: We are all biased. Even as we learn to sort shapes and colors and distinguish puppies from kittens, we also learn to categorize people on the basis of traits they seem to share. We might associate women who resemble our nannies, mothers, or grandmothers with nurturing or doing domestic labor. Or following centuries of racism, segregation, and entrenched cultural stereotypes, we might perceive dark-skinned men as more dangerous than others.

The biases we form quickly and early in life are surprisingly immutable. Biases are "sticky," says Kristin Pauker, a psychology researcher at the University of Hawaii, "because they rely on this very fundamental thing that we all do. We naturally categorize things, and we want to have a positivity associated with the groups we're in." These associations are logical shortcuts that help us make quick decisions when navigating the world. But they also form the roots of often illogical attractions and revulsions, like red shirts versus blue shirts.

Our reflexive, implicit biases wreak devastating social harm. When we stereotype individuals based on gender, ethnicity, sexual orientation, or race, our mental stereotypes begin to drive our behavior and decisions, such as whom to hire, who we perceive as incompetent, delinquent, or worse. Earlier this year, for instance, an appeals court overturned a Black man's conviction for heroin distribution and the 10-year prison sentence he received in part because the Detroit federal judge who handed down the original verdict admitted, "This guy looks like a criminal to me."

People who live in racially homogeneous environments may struggle to distinguish faces of a different race from one another.

Correcting for the biases buried in our brains is difficult, but it is also hugely important. Because

women are stereotyped as domestic, they are also generally seen as less professional. That attitude has reinforced a decades-long wage gap. Even today, women still earn only 82 cents for every dollar that men earn. Black men are perceived as more violent than white men, and thus are subjected to discriminatory policing and harsher prison sentences, as in the Detroit case. Clinicians' implicit preferences for cisgender, heterosexual patients cause widespread inequities in health care for LGBTQ+ individuals.

"These biases are operating on huge numbers of people repetitively over time," says Anthony Greenwald, a social psychologist at the University of Washington. "The effects of implicit biases accumulate to have great impact."

Greenwald was one of the first researchers to recognize the scope of the problems created by our implicit biases. In the mid-1990s, he created early tests to study and understand implicit association. Along with colleagues Mahzarin Banaji, Brian Nosek, and others, he hoped that shining a light on the issue might quickly identify the tools needed to fix it. Being aware that our distorted thinking was hurting other people should be enough to give pause and force us to do better, they thought.

They were wrong. Although implicit bias training programs help people become aware of their biases, both anecdotal reports and controlled studies have shown that the programs do little to reduce discriminatory behaviors spurred by those prejudices. "They fail in the most important respect," Greenwald says. When he, Banaji, and Nosek developed the Implicit Association Test, he took it himself. He was distressed to discover that he automatically associated more positive words with the faces of white people, and more unfavorable words with people who were Black. "I didn't regard myself as a prejudiced person," Greenwald says. "But I had this association nevertheless."

His experience is not unusual. The Implicit Association Test (IAT) measures the speed of subjects' responses as they match descriptors of people (such as *Hispanic* or *gay*) to qualities (such as attractiveness, athleticism, or being professional). It's based on the idea that people react more quickly when they are matching qualities that are already strongly associated in their minds. Implicit bias exists separately from explicit opinion, so someone who honestly believes they don't have anything against gay people, for instance, may still reveal a bias against them on the test. "A lot of people are surprised by their results," Greenwald says. "This is very hard for people to come to grips with intuitively."

People's beliefs may not matter as much if they can be persuaded not to act on them.

One reason we are so often unaware of our implicit biases is that we begin to form these mental associations even before we can express a thought. Brain-imaging studies have found that six-month-old babies can identify individual monkey faces as well as individual humans. Just a dozen weeks later, nine-month-old babies retain the ability to identify human faces but begin to group all the monkey faces together generically as just "monkey," losing the ability to spot individual features. Shortly after, babies

begin to group human faces by race and ethnicity. Our adult brains echo these early learning patterns. People who live in racially homogeneous environments may struggle to distinguish faces of a different race from one another.

As it became clear how deeply ingrained these biases are—and how they might be unfathomable even to ourselves—researchers began to design new types of strategies to mitigate bias and its impact in society. By 2017, companies in the United States were spending $8 billion annually on diversity training efforts, including those aimed at reducing unconscious stereotyping, according to management consulting firm McKinsey & Company. These trainings range from online educational videos to workshops lasting a few hours or days in which participants engage in activities such as word-association tests that help identify their internalized biases.

Recent data suggest that these efforts have been failing too. In 2019 researchers evaluated the effectiveness of 18 methods that aimed to reduce implicit bias, particularly pro-white and anti-Black bias. Only half the methods proved even temporarily effective, and they shared a common theme: They worked by giving study participants experiences that contradicted stereotypes. Reading a story with an evil white man and a dashing young Black hero, for example, reduced people's association of Black men with criminality. Most of these strategies had fleeting effects that lasted only hours. The most effective ones reduced bias for only a few days at best.

Even when training reduced bias, it did little to reduce discriminatory outcomes. Beginning in early 2018, the New York City Police Department began implicit bias training for its 36,000 personnel to reduce racial inequities in policing. When researchers evaluated the project in 2020, they found that most officers were aware of the problems created by implicit bias and were keen to address these harms, but their behaviors contradicted these intentions. Data on arrests, stops, and stop-and-frisk actions showed that officers who had completed the training were still more likely to take these actions against Black and Hispanic people. In fact, the training program hardly had any effect on the numbers.

This and similar studies have "thrown some cold water on just targeting implicit bias as a focus of intervention," says Calvin Lai, a social psychologist at Washington University in St. Louis. Even if you are successful in changing implicit bias or making people more aware of it, "you can't easily assume that people will be less discriminatory."

But researchers are finding reason for hope.

Although the dozens of interventions tested so far have demonstrated limited long-term effects, some still show that people can be made more aware of implicit bias and can be moved to act more equitably, at least temporarily. In 2016, Lai and his colleagues tested eight ways of reducing unconscious bias in studies with college students. One of the interventions they tested involved participants reading a

vividly portrayed scenario in which a white person assaulted them and a Black person came to their rescue. The story reinforced the connection between heroism and Black identity.

Other interventions were designed to heighten similar connections. For instance, one offered examples of famous Black individuals, such as Oprah Winfrey, and contrasted them with examples of infamous white people, including Adolf Hitler. Participants' biases were gauged using the IAT both before and after these interventions. While the experiments tamped down bias temporarily, none of them made a difference just a few days later. "People go into the lab and do an intervention and there's that immediate effect," Pauker says.

From such small but significant successes, an insight began to emerge: Perhaps the reason implicit bias is stable is because we inhabit an environment that's giving us the same messages again and again. Instead of trying to chip away at implicit bias merely by changing our minds, perhaps success depended on changing our environment.

The implicit associations we form—whether about classmates who wear the same color shirt or about people who look like us—are a product of our mental filing cabinets. But a lot of what's in those filing cabinets is drawn from our culture and environment. Revise the cultural and social inputs, researchers like Kristin Pauker theorize, and you have a much greater likelihood of influencing implicit bias than you do by sending someone to a one-off class or training program.

Babies who start to blur monkey faces together do so because they learn, early on, that distinguishing human faces is more critical than telling other animals apart. Similarly, adults categorize individuals by race, gender, or disability status because these details serve as markers of something we've deemed important as a society. "We use certain categories because our environment says those are the ones that we should be paying attention to," Pauker says.

Just as we are oblivious to many of the biases in our heads, we typically don't notice the environmental cues that seed those biases. In a 2009 study, Pauker and her colleagues examined the cultural patterns depicted in 11 highly popular TV shows, including *Grey's Anatomy*, *Scrubs*, and *CSI Miami*. The researchers tracked nonverbal interactions among characters on these shows and found that even when white and Black characters were equal in status and jobs and spoke for about the same amount of time, their nonverbal interactions differed. For instance, on-screen characters were less likely to smile at Black characters, and the latter were more often portrayed as stern or unfriendly.

Thinking of implicit bias as malleable allows us to constantly reframe our judgments about people we meet.

In a series of tests, Pauker and her colleagues found that regular viewers of such shows were more likely to have stronger anti-Black implicit biases on the IAT. But when the researchers asked viewers

multiple-choice questions about bias in the video clips they saw, viewers' responses about whether they'd witnessed pro-Black or pro-white bias were no better than random. They were being influenced by the bias embedded in the show, "but they were not able to explicitly detect it," Pauker says.

Perhaps the most definitive proof that the outside world shapes our biases emerged from a recent study of attitudes toward homosexuality and race over decades. In 2019 Harvard University experimental psychologist Tessa Charlesworth and her colleagues analyzed the results of 4.4 million IATs taken by people between 2007 and 2016. The researchers found that anti-gay implicit bias had dropped about 33 percent over the years, while negative racial attitudes against people of color declined by about 17 percent.

The data were the first to definitively show that implicit attitudes can change in response to a shifting zeitgeist. The changes in attitudes weren't due to any class or training program. Rather, they reflected societal changes, including marriage equality laws and protections against racial discrimination. Reducing *explicit* discrimination altered the *implicit* attitudes instilled by cultures and communities—and thus helped people rearrange their mental associations and biases.

Until societal shifts occur, however, researchers are finding alternate ways to reduce the harms caused by implicit bias. People's beliefs may not matter as much if they can be persuaded not to act on them. According to the new way of thinking, managers wouldn't just enter training to reduce their bias. Instead, they could be trained to remove implicit bias from hiring decisions by setting clear criteria before they begin the hiring process.

Faced with a stack of resumes that reveal people's names, ethnicities, or gender, an employer's brain automatically starts slotting them based on preconceived notions of who is more professional or worthy of a job. Then bias supersedes logic.

When we implicitly favor someone, we are more likely to regard their strengths as important. Consider, for example, a hiring manager who perceives men as more suited to a role than women. Meeting a male candidate with a low GPA but considerable work experience may lead the manager to think that real-world experience is what really matters. But if the man has a higher GPA and less experience, the manager might instead reason that the latter isn't important because experience can be gained on the job. To avoid this all-too-common scenario, employers could define specific criteria necessary for a role, then create a detailed list of questions needed to evaluate those criteria and use these to create a structured interview. Deciding in advance whether education or work experience matters more can reduce this problem and lead to more equitable decisions. "You essentially sever the link between the bias and the behavior," explains Benedek Kurdi, a psychologist at the University of Illinois Urbana–Champaign. "What you're saying is the bias can remain, but you deprive it of the opportunity to influence decision making."
In the long run, reducing the biases and injustices built into our environment is the only surefire path toward

taming the harmful implicit biases in our heads. If we see a world with greater equity, our internal attitudes seem to adjust to interpret that as normal. There's no magical way to make the whole world fair and equitable all at once. But it may be possible to help people envision a better world from the start so that their brains form fewer flawed associations in the first place.

To Pauker, achieving that goal means teaching children to be flexible in their thinking from an early age. Children gravitate toward same-race interactions by about the age of 10. In one study, Pauker and her colleagues found that offering stories to children that nudged them to think about racial bias as flexible made them more likely to explore mixed-race friendships. In another study, Pauker and team found that children who thought about prejudice as fixed had more uncomfortable interactions with friends of other races and eventually avoided them. But those who thought about prejudice as malleable—believing they could change their minds about people of other races—were less likely to avoid friends of other races. The key, Pauker suggests, is not to rethink rigid mental categories but to encourage mental flexibility. Her approach, which encourages children to consider social categories as fluid constructs, appears to be more effective. The data are preliminary, but they offer a powerful route to change: simply being open to updating the traits we associate with different groups of people.

Thinking of implicit bias as malleable allows us to constantly reframe our judgments about people we meet—evaluating each unique individual for what they are, rather than reducing them to a few preconceived traits we associate with their race, gender, or other social category. Rather than trying to fight against our wariness toward out-groups, reconsidering our mental classifications in this manner allows us to embrace the complexity of human nature and experience, making more of the world feel like our in-group. Blurring the implicit lines in our minds might be the first step to reducing disparities in the world we make.

Source: Beyond Bias | Pulitzer Center. https://pulitzercenter.org/stories/beyond-bias

Part III: Policy Protections.
Imagine that you have the power to write a policy to stop AI bias. Write one policy rule based specifically on the information presented to you today on biases and objectifying women's bodies. Include the following information:

Who has to follow this policy?

What will the policy do (or prevent)? How will you ensure this is done?

Write your policy here.

Junior Honors Cohort
Darnell, Healy, Houghteling, Jihan

Introducing AI Bias
Reading 2

name: Mr. McGee
period: 1-2

Part I: AI Biases Presentation.

Directions: Complete the steps below as you discuss AI biases via the presentation.

Define the following terms from the discussion and slides:

1. Training Data - The information to Ai to teach it a certain task-script
2. Testing Data - The information used to check whatever the Ai that was created is reliable and accurate - output
3. AI Bias - When an Ai tool makes a decision that is wrong or prejudice against it learned from training data that isn't from all people perspectives.

Turn-and-Talk 1: Take 60 seconds to brainstorm with your elbow partner - how do you think we could reduce the bias in the AI training of detecting fruit types?

Turn-and-Talk 2: Take 2 minutes to imagine that you have now been asked to create an AI tool that can identify weather conditions, such as sunny, rainy, or cloudy. **Describe the kind of images** you would include as a part of the **training data.** Your goal is to come up with as complete of a data set as possible so that you limit the chance for AI bias! Storm clouds, wet grounds, sun [illegible] sun [illegible] Day, night and not say it's Bad or good either way

Drawing Conclusions: In your own words, and in 2-3 sentences, describe how AI bias occurs.

Bias occurs when Ai is used and is told to process something that it is told to do by a person to do.

Part II: Bias in AI Partner Reading

Directions: Annotate as you read the following article with your partner. Specifically, take notes in the space below about what the authors did and what actions they took to capture data about AI. Then, summarize the conclusions of the data with your partner.

AI is unintenonally Bias Based on the way they were made.

'There Is No Standard': Investigation Finds AI Algorithms Objectify Women's Bodies

By Hilke Schellmann and Gianluca Mauro

AI tools rate photos of women as more sexually suggestive than those of men, especially if nipples, pregnant bellies or exercise is involved

Images posted on social media are analyzed by artificial intelligence (AI) algorithms that decide what to amplify and what to suppress. Many of these algorithms, a Guardian investigation has found, have a gender bias, and may have been censoring and suppressing the reach of countless photos featuring women's bodies.

These AI tools, developed by large technology companies, including Google and Microsoft, are meant to protect users by identifying violent or pornographic visuals so that social media companies can block it before anyone sees it. The companies claim that their AI tools can also detect "raciness" or how sexually suggestive an image is. With this classification, platforms – including Instagram and LinkedIn – may suppress contentious imagery.

Two Guardian journalists used the AI tools to analyze hundreds of photos of men and women in underwear, working out, using medical tests with partial nudity and found evidence that the AI tags photos of women in everyday situations as sexually suggestive. They also rate pictures of women as more "racy" or sexually suggestive than comparable pictures of men. As a result, the social media companies that leverage these or similar algorithms have suppressed the reach of countless images featuring women's bodies, and hurt female-led businesses – further amplifying societal disparities.

Even medical pictures are affected by the issue. The AI algorithms were tested on images released by the US National Cancer Institute demonstrating how to do a clinical breast examination. Google's AI gave this photo the highest score for raciness, Microsoft's AI was 82% confident that the image was "explicitly sexual in nature", and Amazon classified it as representing "explicit nudity"..

Pregnant bellies are also problematic for these AI tools. Google's algorithm scored the photo as "very likely to contain racy content". Microsoft's algorithm was 90% confident that the image was "sexually suggestive in nature".

"This is just wild," said Leon Derczynski, a professor of computer science at the IT University of Copenhagen, who specializes in online harm. "Objectification of women seems deeply embedded in the system."

One social media company said they do not design their systems to create or reinforce biases and classifiers are not perfect.

"This is a complex and evolving space, and we continue to make meaningful improvements to SafeSearch classifiers to ensure they stay accurate and helpful for everyone," a Google spokesperson said.

Getting Shadowbanned

In May 2021, Gianluca Mauro, an AI entrepreneur, advisor and co-author of this article, published a LinkedIn post and was surprised it had just been seen 29 times in an hour, instead of the roughly 1,000 views he usually gets. Maybe the picture of two women wearing tube tops was the problem?

He re-uploaded the same exact text with another picture. The new post got 849 views in an hour.

It seemed like his post had been suppressed or "shadowbanned". Shadowbanning refers to the decision of a social media platform to limit the reach of a post or account. While a regular ban involves actively blocking a post or account and notifying the user, shadowbanning is less transparent - often the reach will be suppressed without the user's knowledge.

The Guardian found that Microsoft, Amazon and Google offer content moderation algorithms to any business for a small fee. Microsoft, the parent company and owner of LinkedIn, said its tool "can detect adult material in images so that developers can restrict the display of these images in their software".

Another experiment on LinkedIn was conducted to try to confirm the discovery.

In two photos depicting both women and men in underwear, Microsoft's tool classified the picture showing two women as racy and gave it a 96% score. The picture with the men was classified as non-racy with a score of 14%.

The photo of the women got eight views within one hour, and the picture with the two men received 655 views, suggesting the photo of the women in underwear was either suppressed or shadowbanned.

Shadowbanning has been documented for years, but the Guardian journalists may have found a missing link to understand the phenomenon: biased AI algorithms. Social media platforms seem to leverage these algorithms to rate images and limit the reach of content that they consider too racy. The problem seems to be that these AI algorithms have built-in gender bias, rating women more racy than images containing men.

"Our teams utilize a combination of automated techniques, human expert reviews and member reporting to help identify and remove content that violates our professional community policies," said LinkedIn spokesperson Fred Han in a statement. "In addition, our feed uses algorithms responsibly in order to surface content that helps our members be more productive and successful in their professional journey." Amazon said content moderation is based on a variety of factors including geography, religious beliefs and cultural experience. However, "Amazon Rekognition is able to recognize a wide variety of content, but it does not determine the appropriateness of that content," an Amazon spokesperson said. "The service

simply returns labels for items it detects for further evaluation by human moderators."

Digging deeper

Natasha Crampton, Microsoft's chief responsible AI officer, and her team began investigating when journalists notified her about the labeling of the photos.

"The initial results do not suggest that those false positives occur at a disproportionately higher rate for women as compared with men," Crampton said. When additional photos were run through the tool, the demo website had been changed. Before the problem was discovered, it was possible to test the algorithms by simply dragging and dropping a picture. Now an account needed to be created and code had to be written.

But what are these AI classifiers actually analyzing in the photos? More experiments were needed, so Mauro agreed to be the test subject.

When photographed in long pants and with a bare chest, Microsoft's algorithm had a confidence score lower than 22% for raciness. When Mauro put on a bra, the raciness score jumped to 97%. The algorithm gave a 99% score when the bra was held next to me.

"You are looking at decontextualized information where a bra is being seen as inherently racy rather than a thing that many women wear every day as a basic item of clothing," said Kate Crawford, professor at the University of Southern California and the author of *Atlas of AI*.

Abeba Birhane, a senior fellow at the Mozilla Foundation and an expert in large visual datasets, said raciness is a social concept that differs from one culture to the other.

"These concepts are not like identifying a table where you have the physical thing and you can have a relatively agreeable definition or rating for a certain thing," she said. "You cannot have one single uncontested definition of raciness."

Why Do These Systems Seem So Biased?

Modern AI is built using machine learning, a set of algorithms that allow computers to learn from data. When developers use machine learning, they don't write explicit rules telling computers how to perform a task. Instead, they provide computers with training data. People are hired to label images so that computers can analyze their scores and find whatever pattern helps it replicate human decisions.

Margaret Mitchell, chief ethics scientist at the AI firm Hugging Face and former co-head of Google's Ethical AI research group, believes that the photos used to train these algorithms were probably labeled by straight men, who may associate men working out with fitness, but may consider an image of a woman working out as racy. It's also possible that these ratings seem gender biased in the US and in Europe

because the labelers may have been from a place with a more conservative culture.

Ideally, tech companies should have conducted thorough analyses on who is labeling their data, to make sure that the final dataset embeds a diversity of views, she said. The companies should also check that their algorithms perform similarly on photos of men v women and other groups, but that is not always done.

"There's no standard of quality here," Mitchell said.

This gender bias the Guardian uncovered is part of more than a decade of controversy around content moderation on social media. Images showing people breastfeeding their children and different standards for photos of male nipples, which are allowed on Instagram, and female nipples, which have to be covered, have long garnered outcries about social media platforms' content moderation practices.

Now Meta's oversight board - an external body including professors, researchers and journalists, who are paid by the company - has asked the tech giant to clarify its adult nudity and sexual activity community standard guidelines on social media platforms "so that all people are treated in a manner consistent with international human rights standards, without discrimination on the basis of sex or gender."
Meta declined to comment for this story.

'Women Should Be Expressing Themselves'

Bec Wood, a 38-year-old photographer based in Perth, Australia, said she's terrified of Instagram's algorithmic police force.

After Wood had a daughter nine years ago, she started studying childbirth education and photographing women trying to push back against societal pressures many women feel that they should look like supermodels.

"I was not having that for my daughter," she said. "Women should be expressing themselves and celebrating themselves and being seen in all these different shapes and sizes. I just think that's so important for humanity to move forward."

Wood's photos are intimate glimpses into women's connections with their offspring, photographing breastfeeding, pregnancy and other important moments in an artful manner. Her business is 100% dependent on Instagram: "That's where people find you," Wood said. "If I don't share my work, I don't get work."

Since Wood started her business in 2018, for some of her photos she got messages from Instagram that the company was either taking down some of her pictures or that they were going to allow them on her profile but not on the explore tab, a section of the app where people can discover content from accounts they don't follow. She hoped that Instagram was going to fix the issue over time, but the opposite

happened, she said. "I honestly can't believe that it's gotten worse. It has devastated my business." Wood described 2022 as her worst year business-wise.

She is terrified that if she uploads the "wrong" image, she will be locked out of her account with over 13,000 followers, which would bankrupt her business: "I'm literally so scared to post because I'm like, 'Is this the post that's going to lose everything?'" she said.

To avoid this, Wood started going against what made her start her work in the first place: "I will censor as artistically as possible any nipples. I find this so offensive to art, but also to women," she said. "I almost feel like I'm part of perpetuating that ridiculous cycle that I don't want to have any part of."

Running some of Wood's photos through the AI algorithms of Microsoft, Google, and Amazon, including those featuring a pregnant belly got rated as racy, nudity or even explicitly sexual.

Wood is not alone. Carolina Are, an expert on social media platforms and content moderation and currently an Innovation fellow at the Centre for Digital Citizens at Northumbria University said she has used Instagram to promote her business and was a victim of shadowbanning.

Are, a pole dance instructor, said some of her photos were taken down, and in 2019, she discovered that her pictures did not show up in the explore page or under the hashtag #FemaleFitness, where Instagram users can search content from users they do not follow. "It was literally just women working out in a very tame way. But then if you looked at hashtag #MaleFitness, it was all oily dudes and they were fine. They weren't shadowbanned," she said.

For Are, these individual problems point to larger systemic ones: many people, including chronically ill and disabled folks, rely on making money through social media and shadowbanning harms their business. Mitchell, the chief ethics scientist at Hugging Face, these kinds of algorithms are often recreating societal biases: "It means that people who tend to be marginalized are even further marginalized - like literally pushed down in a very direct meaning of the term marginalization."

It's a representational harm and certain populations are not adequately represented, she added. "In this case, it would be an idea that women must cover themselves up more than men and so that ends up creating this sort of social pressure for women as this becomes the norm of what you see, " Mitchell said. The harm is worsened by a lack of transparency. While in some cases Wood has been notified that her pictures were banned or limited in reach, she believes Instagram took other actions against her account without her knowing it. "I've had people say 'I can't tag you,' or 'I was searching for you to show my friend the other day and you're not showing up,'" she said. "I feel invisible."

Because she might be, said computer scientist Derczynski: "The people posting these images will never find out about it, which is just so deeply problematic." he said. "They get a disadvantage forced upon them and they have no agency in this happening and they're not informed that it's happening either."

Source:
https://pulitzercenter.org/stories/there-no-standard-investigation-finds-ai-algorithms-objectify-womens-bodies

Part III: Policy Protections.
Imagine that you have the power to write a policy to stop AI bias. Write one policy rule based specifically on the information presented to you today on biases and objectifying women's bodies. Include the following information:

Who has to follow this policy?

What will the policy do (or prevent)? How will you ensure this is done?

Write your policy here.

Junior Honors Cohort
Darnell, Healy, Houghteling, Jihan

Introducing AI Bias
Reading 1

name: Enakeyah M.
period: 1st

Part I: AI Biases Presentation.

Directions: Complete the steps below as you discuss AI biases via the presentation.

Define the following terms from the discussion and slides:

1. Training Data - The information given to AI to help it learn how to do a specific task (input).
2. Testing Data - The information used to check whether the AI that was created is reliable and accurate (output).
3. AI Bias - when an AI tool makes a decision that is wrong or problematic because it learned from training data that didn't treat all people, places, and things accurately.

Turn-and-Talk 1: Take 60 seconds to brainstorm with your elbow partner - how do you think we could reduce the bias in the AI training of detecting fruit types? we could reduce biases in AI training in detecting fruit types by inputting different color fruits within similar color categories to produce more accurate AI fruit outputs.

Turn-and-Talk 2: Take 2 minutes to imagine that you have now been asked to create an AI tool that can identify weather conditions, such as sunny, rainy, or cloudy. **Describe the kind of images** you would include as a part of the **training data.** Your goal is to come up with as complete of a data set as possible so that you limit the chance for AI bias! sunny weather conditions: different angles of the sun, sun rays, blue sky, day time skies, sunsets, and sun throughout all four seasons, yellow and orange circles.

Drawing Conclusions: In your own words, and in 2-3 sentences, describe how AI bias occurs.

AI biases occur through errors created by the people who compose an input. These errors occur when data is not put in accurately enough for the computer to understand. This results in an inaccurate output.

Part II: Bias in AI Partner Reading

Directions: Annotate as you read the following article with your partner. Specifically, take notes in the space below about what the authors did and what actions they took to capture data about AI. Then, summarize the conclusions of the data with your partner.

"Implicit bias is an unconscious prejudice we learn from a young age. It shapes how we view people from different social groups."

"This preference for people... we are all biased" (lines 7-8).

Are AI Hiring Tools Racist and Ableist?

By Hilke Schellmann

If people with accents or speech impairments are less well "understood" by AI used in hiring, this would constitute discrimination based on disability and national origin, which is illegal in the U.S.

Artificial Intelligence (AI) is now being used in every aspect of the "employee life cycle"—from hiring to firing. For hiring, many companies utilize one-way AI-based interview tools, which send questions to job seekers' devices and have them record their answers without a human on the other side.

The software then transcribes what job applicants say in the video interviews to text. The AI compares the text to job interviews of current employees, who are deemed successful. If applicants use similar words as current employees have used in their job interviews, they will get a favorable score. If they have less overlap with current employees, they are going to get rejected by the AI.

What has rarely been discussed and never studied is if the underlying technology, the speech-to-text transcription, is actually treating everyone fairly and equally. If the speech-to-text transcription process produces a significantly higher "Word Error Rate" (WER) for speakers with accents or speech impairments versus native speakers without speech impairments, this may lead to applicants getting unfairly rejected.

This research is of vital importance, because the results could prove significant flaws in the speech-to-text transcription systems that underpin the AI used in hiring. If the study shows that these tools discriminate against people with accents and speech impairments, this would be illegal in the U.S.

AI Algorithms Objectify Women's Bodies

This project also investigates gender bias by algorithms used by some of the largest platforms, including Google and Microsoft. Our research shows that these algorithms tag photos of women in everyday situations as racy or sexually suggestive at higher rates than images showing men in similar situations. As a result, the social media companies that leverage these algorithms have suppressed the reach of countless images featuring women's bodies, and hurt female-led businesses—further amplifying societal disparities.

Source: Are AI Hiring Tools Racist and Ableist? | Pulitzer Center

Beyond Bias

By Jyoti Madhusoodanan

We all have bias embedded in our brains, but there are ways we can move past it. New findings from psychology show us how.

In a now-classic series of experiments, researchers teased out the deep-rooted nature of human bias simply by distributing red shirts and blue shirts to groups of 3- to 5-year-olds at a day care center. In one classroom, teachers were asked to divide children into groups based on the color of their shirts. In another, teachers were instructed to overlook the shirt colors. After three weeks, children in both classrooms tended to prefer being with classmates who wore the same color as themselves—no matter what the teachers did.

This preference for people who seem to belong to our own tribe forms early and drives our choices throughout life. There appears to be no avoiding it: We are all biased. Even as we learn to sort shapes and colors and distinguish puppies from kittens, we also learn to categorize people on the basis of traits they seem to share. We might associate women who resemble our nannies, mothers, or grandmothers with nurturing or doing domestic labor. Or following centuries of racism, segregation, and entrenched cultural stereotypes, we might perceive dark-skinned men as more dangerous than others.

The biases we form quickly and early in life are surprisingly immutable. Biases are "sticky," says Kristin Pauker, a psychology researcher at the University of Hawaii, "because they rely on this very fundamental thing that we all do. We naturally categorize things, and we want to have a positivity associated with the groups we're in." These associations are logical shortcuts that help us make quick decisions when navigating the world. But they also form the roots of often illogical attractions and revulsions, like red shirts versus blue shirts.

Our reflexive, implicit biases wreak devastating social harm. When we stereotype individuals based on gender, ethnicity, sexual orientation, or race, our mental stereotypes begin to drive our behavior and decisions, such as whom to hire, who we perceive as incompetent, delinquent, or worse. Earlier this year, for instance, an appeals court overturned a Black man's conviction for heroin distribution and the 10-year prison sentence he received in part because the Detroit federal judge who handed down the original verdict admitted, "This guy looks like a criminal to me."

People who live in racially homogeneous environments may struggle to distinguish faces of a different race from one another.

Correcting for the biases buried in our brains is difficult, but it is also hugely important. Because

women are stereotyped as domestic, they are also generally seen as less professional. That attitude has reinforced a decades-long wage gap. Even today, women still earn only 82 cents for every dollar that men earn. Black men are perceived as more violent than white men, and thus are subjected to discriminatory policing and harsher prison sentences, as in the Detroit case. Clinicians' implicit preferences for cisgender, heterosexual patients cause widespread inequities in health care for LGBTQ+ individuals.

"These biases are operating on huge numbers of people repetitively over time," says Anthony Greenwald, a social psychologist at the University of Washington. "The effects of implicit biases accumulate to have great impact."

Greenwald was one of the first researchers to recognize the scope of the problems created by our implicit biases. In the mid-1990s, he created early tests to study and understand implicit association. Along with colleagues Mahzarin Banaji, Brian Nosek, and others, he hoped that shining a light on the issue might quickly identify the tools needed to fix it. Being aware that our distorted thinking was hurting other people should be enough to give pause and force us to do better, they thought.

They were wrong. Although implicit bias training programs help people become aware of their biases, both anecdotal reports and controlled studies have shown that the programs do little to reduce discriminatory behaviors spurred by those prejudices. "They fail in the most important respect," Greenwald says. When he, Banaji, and Nosek developed the Implicit Association Test, he took it himself. He was distressed to discover that he automatically associated more positive words with the faces of white people, and more unfavorable words with people who were Black. "I didn't regard myself as a prejudiced person," Greenwald says. "But I had this association nevertheless."

His experience is not unusual. The Implicit Association Test (IAT) measures the speed of subjects' responses as they match descriptors of people (such as *Hispanic* or *gay*) to qualities (such as attractiveness, athleticism, or being professional). It's based on the idea that people react more quickly when they are matching qualities that are already strongly associated in their minds. Implicit bias exists separately from explicit opinion, so someone who honestly believes they don't have anything against gay people, for instance, may still reveal a bias against them on the test. "A lot of people are surprised by their results," Greenwald says. "This is very hard for people to come to grips with intuitively."

People's beliefs may not matter as much if they can be persuaded not to act on them.

One reason we are so often unaware of our implicit biases is that we begin to form these mental associations even before we can express a thought. Brain-imaging studies have found that six-month-old babies can identify individual monkey faces as well as individual humans. Just a dozen weeks later, nine-month-old babies retain the ability to identify human faces but begin to group all the monkey faces together generically as just "monkey," losing the ability to spot individual features. Shortly after, babies

begin to group human faces by race and ethnicity. Our adult brains echo these early learning patterns. People who live in racially homogeneous environments may struggle to distinguish faces of a different race from one another.

As it became clear how deeply ingrained these biases are—and how they might be unfathomable even to ourselves—researchers began to design new types of strategies to mitigate bias and its impact in society. By 2017, companies in the United States were spending $8 billion annually on diversity training efforts, including those aimed at reducing unconscious stereotyping, according to management consulting firm McKinsey & Company. These trainings range from online educational videos to workshops lasting a few hours or days in which participants engage in activities such as word-association tests that help identify their internalized biases.

Recent data suggest that these efforts have been failing too. In 2019 researchers evaluated the effectiveness of 18 methods that aimed to reduce implicit bias, particularly pro-white and anti-Black bias. Only half the methods proved even temporarily effective, and they shared a common theme: They worked by giving study participants experiences that contradicted stereotypes. Reading a story with an evil white man and a dashing young Black hero, for example, reduced people's association of Black men with criminality. Most of these strategies had fleeting effects that lasted only hours. The most effective ones reduced bias for only a few days at best.

Even when training reduced bias, it did little to reduce discriminatory outcomes. Beginning in early 2018, the New York City Police Department began implicit bias training for its 36,000 personnel to reduce racial inequities in policing. When researchers evaluated the project in 2020, they found that most officers were aware of the problems created by implicit bias and were keen to address these harms, but their behaviors contradicted these intentions. Data on arrests, stops, and stop-and-frisk actions showed that officers who had completed the training were still more likely to take these actions against Black and Hispanic people. In fact, the training program hardly had any effect on the numbers.

This and similar studies have "thrown some cold water on just targeting implicit bias as a focus of intervention," says Calvin Lai, a social psychologist at Washington University in St. Louis. Even if you are successful in changing implicit bias or making people more aware of it, "you can't easily assume that people will be less discriminatory."

But researchers are finding reason for hope.

Although the dozens of interventions tested so far have demonstrated limited long-term effects, some still show that people can be made more aware of implicit bias and can be moved to act more equitably, at least temporarily. In 2016, Lai and his colleagues tested eight ways of reducing unconscious bias in studies with college students. One of the interventions they tested involved participants reading a

vividly portrayed scenario in which a white person assaulted them and a Black person came to their rescue. The story reinforced the connection between heroism and Black identity.

Other interventions were designed to heighten similar connections. For instance, one offered examples of famous Black individuals, such as Oprah Winfrey, and contrasted them with examples of infamous white people, including Adolf Hitler. Participants' biases were gauged using the IAT both before and after these interventions. While the experiments tamped down bias temporarily, none of them made a difference just a few days later. "People go into the lab and do an intervention and there's that immediate effect," Pauker says.

From such small but significant successes, an insight began to emerge: Perhaps the reason implicit bias is stable is because we inhabit an environment that's giving us the same messages again and again. Instead of trying to chip away at implicit bias merely by changing our minds, perhaps success depended on changing our environment.

The implicit associations we form—whether about classmates who wear the same color shirt or about people who look like us—are a product of our mental filing cabinets. But a lot of what's in those filing cabinets is drawn from our culture and environment. Revise the cultural and social inputs, researchers like Kristin Pauker theorize, and you have a much greater likelihood of influencing implicit bias than you do by sending someone to a one-off class or training program.

Babies who start to blur monkey faces together do so because they learn, early on, that distinguishing human faces is more critical than telling other animals apart. Similarly, adults categorize individuals by race, gender, or disability status because these details serve as markers of something we've deemed important as a society. "We use certain categories because our environment says those are the ones that we should be paying attention to," Pauker says.

Just as we are oblivious to many of the biases in our heads, we typically don't notice the environmental cues that seed those biases. In a 2009 study, Pauker and her colleagues examined the cultural patterns depicted in 11 highly popular TV shows, including *Grey's Anatomy*, *Scrubs*, and *CSI Miami*. The researchers tracked nonverbal interactions among characters on these shows and found that even when white and Black characters were equal in status and jobs and spoke for about the same amount of time, their nonverbal interactions differed. For instance, on-screen characters were less likely to smile at Black characters, and the latter were more often portrayed as stern or unfriendly.

Thinking of implicit bias as malleable allows us to constantly reframe our judgments about people we meet.

In a series of tests, Pauker and her colleagues found that regular viewers of such shows were more likely to have stronger anti-Black implicit biases on the IAT. But when the researchers asked viewers

multiple-choice questions about bias in the video clips they saw, viewers' responses about whether they'd witnessed pro-Black or pro-white bias were no better than random. They were being influenced by the bias embedded in the show, "but they were not able to explicitly detect it," Pauker says.

Perhaps the most definitive proof that the outside world shapes our biases emerged from a recent study of attitudes toward homosexuality and race over decades. In 2019 Harvard University experimental psychologist Tessa Charlesworth and her colleagues analyzed the results of 4.4 million IATs taken by people between 2007 and 2016. The researchers found that anti-gay implicit bias had dropped about 33 percent over the years, while negative racial attitudes against people of color declined by about 17 percent.

The data were the first to definitively show that implicit attitudes can change in response to a shifting zeitgeist. The changes in attitudes weren't due to any class or training program. Rather, they reflected societal changes, including marriage equality laws and protections against racial discrimination. Reducing *explicit* discrimination altered the *implicit* attitudes instilled by cultures and communities—and thus helped people rearrange their mental associations and biases.

Until societal shifts occur, however, researchers are finding alternate ways to reduce the harms caused by implicit bias. People's beliefs may not matter as much if they can be persuaded not to act on them. According to the new way of thinking, managers wouldn't just enter training to reduce their bias. Instead, they could be trained to remove implicit bias from hiring decisions by setting clear criteria before they begin the hiring process.

Faced with a stack of resumes that reveal people's names, ethnicities, or gender, an employer's brain automatically starts slotting them based on preconceived notions of who is more professional or worthy of a job. Then bias supersedes logic.

When we implicitly favor someone, we are more likely to regard their strengths as important. Consider, for example, a hiring manager who perceives men as more suited to a role than women. Meeting a male candidate with a low GPA but considerable work experience may lead the manager to think that real-world experience is what really matters. But if the man has a higher GPA and less experience, the manager might instead reason that the latter isn't important because experience can be gained on the job. To avoid this all-too-common scenario, employers could define specific criteria necessary for a role, then create a detailed list of questions needed to evaluate those criteria and use these to create a structured interview. Deciding in advance whether education or work experience matters more can reduce this problem and lead to more equitable decisions. "You essentially sever the link between the bias and the behavior," explains Benedek Kurdi, a psychologist at the University of Illinois Urbana–Champaign. "What you're saying is the bias can remain, but you deprive it of the opportunity to influence decision making."
In the long run, reducing the biases and injustices built into our environment is the only surefire path toward

taming the harmful implicit biases in our heads. If we see a world with greater equity, our internal attitudes seem to adjust to interpret that as normal. There's no magical way to make the whole world fair and equitable all at once. But it may be possible to help people envision a better world from the start so that their brains form fewer flawed associations in the first place.

To Pauker, achieving that goal means teaching children to be flexible in their thinking from an early age. Children gravitate toward same-race interactions by about the age of 10. In one study, Pauker and her colleagues found that offering stories to children that nudged them to think about racial bias as flexible made them more likely to explore mixed-race friendships. In another study, Pauker and team found that children who thought about prejudice as fixed had more uncomfortable interactions with friends of other races and eventually avoided them. But those who thought about prejudice as malleable—believing they could change their minds about people of other races—were less likely to avoid friends of other races. The key, Pauker suggests, is not to rethink rigid mental categories but to encourage mental flexibility. Her approach, which encourages children to consider social categories as fluid constructs, appears to be more effective. The data are preliminary, but they offer a powerful route to change: simply being open to updating the traits we associate with different groups of people.

Thinking of implicit bias as malleable allows us to constantly reframe our judgments about people we meet—evaluating each unique individual for what they are, rather than reducing them to a few preconceived traits we associate with their race, gender, or other social category. Rather than trying to fight against our wariness toward out-groups, reconsidering our mental classifications in this manner allows us to embrace the complexity of human nature and experience, making more of the world feel like our in-group. Blurring the implicit lines in our minds might be the first step to reducing disparities in the world we make.

Source: Beyond Bias | Pulitzer Center. https://pulitzercenter.org/stories/beyond-bias

Part III: Policy Protections.
Imagine that you have the power to write a policy to stop AI bias. Write one policy rule based specifically on the information presented to you today on biases and objectifying women's bodies. Include the following information:

Who has to follow this policy?

What will the policy do (or prevent)? How will you ensure this is done?

Write your policy here.

Junior Honors Cohort
Darnell, Healy, Houghteling, Jihan

Introducing AI Bias
Reading 2

name: Mr. Pruitt
period:

Part I: AI Biases Presentation.

Directions: Complete the steps below as you discuss AI biases via the presentation.

Define the following terms from the discussion and slides:

1. Training Data - The information we input into the machine
2. Testing Data - The information used to check wether the ai that was created is reliable and accurate
3. AI Bias - When a ai tool makes a decison that is wrong or problematic because it learned from trained data that can't treatt all people accurate.

Turn-and-Talk 1: Take 60 seconds to brainstorm with your elbow partner - how do you think we could reduce the bias in the AI training of detecting fruit types?

We could really change the color or see if there are more fruits out there.

Turn-and-Talk 2: Take 2 minutes to imagine that you have now been asked to create an AI tool that can identify weather conditions, such as sunny, rainy, or cloudy. **Describe the kind of images** you would include as a part of the **training data.** Your goal is to come up with as complete of a data set as possible so that you limit the chance for AI bias!

I would say sunny, I would add the sun, sun shine, not cloudy, bright, a picture or trees.

Drawing Conclusions: In your own words, and in 2-3 sentences, describe how AI bias occurs.

When the data put into the ai model doesn't represent the reality model. It occurs when the expectations is nothing like reality

Part II: Bias in AI Partner Reading

Directions: Annotate as you read the following article with your partner. Specifically, take notes in the space below about what the authors did and what actions they took to capture data about AI. Then, summarize the conclusions of the data with your partner.

- 2 people used Ai for sexual reasoning
- Ai has been used for big companys like google & microsoft
- Some get shadowbanned for posting to mueh.

'There Is No Standard': Investigation Finds AI Algorithms Objectify Women's Bodies

By Hilke Schellmann and Gianluca Mauro

AI tools rate photos of women as more sexually suggestive than those of men, especially if nipples, pregnant bellies or exercise is involved

Images posted on social media are analyzed by artificial intelligence (AI) algorithms that decide what to amplify and what to suppress. Many of these algorithms, a Guardian investigation has found, have a gender bias, and may have been censoring and suppressing the reach of countless photos featuring women's bodies.

These AI tools, developed by large technology companies, including Google and Microsoft, are meant to protect users by identifying violent or pornographic visuals so that social media companies can block it before anyone sees it. The companies claim that their AI tools can also detect "raciness" or how sexually suggestive an image is. With this classification, platforms – including Instagram and LinkedIn – may suppress contentious imagery.

Two Guardian journalists used the AI tools to analyze hundreds of photos of men and women in underwear, working out, using medical tests with partial nudity and found evidence that the AI tags photos of women in everyday situations as sexually suggestive. They also rate pictures of women as more "racy" or sexually suggestive than comparable pictures of men. As a result, the social media companies that leverage these or similar algorithms have suppressed the reach of countless images featuring women's bodies, and hurt female-led businesses – further amplifying societal disparities.

Even medical pictures are affected by the issue. The AI algorithms were tested on images released by the US National Cancer Institute demonstrating how to do a clinical breast examination. Google's AI gave this photo the highest score for raciness, Microsoft's AI was 82% confident that the image was "explicitly sexual in nature", and Amazon classified it as representing "explicit nudity"..

Pregnant bellies are also problematic for these AI tools. Google's algorithm scored the photo as "very likely to contain racy content". Microsoft's algorithm was 90% confident that the image was "sexually suggestive in nature".

"This is just wild," said Leon Derczynski, a professor of computer science at the IT University of Copenhagen, who specializes in online harm. "Objectification of women seems deeply embedded in the system."

One social media company said they do not design their systems to create or reinforce biases and classifiers are not perfect.

"This is a complex and evolving space, and we continue to make meaningful improvements to SafeSearch classifiers to ensure they stay accurate and helpful for everyone," a Google spokesperson said.

Getting Shadowbanned

In May 2021, Gianluca Mauro, an AI entrepreneur, advisor and co-author of this article, published a LinkedIn post and was surprised it had just been seen 29 times in an hour, instead of the roughly 1,000 views he usually gets. Maybe the picture of two women wearing tube tops was the problem?

He re-uploaded the same exact text with another picture. The new post got 849 views in an hour.

It seemed like his post had been suppressed or "shadowbanned". Shadowbanning refers to the decision of a social media platform to limit the reach of a post or account. While a regular ban involves actively blocking a post or account and notifying the user, shadowbanning is less transparent - often the reach will be suppressed without the user's knowledge.

The Guardian found that Microsoft, Amazon and Google offer content moderation algorithms to any business for a small fee. Microsoft, the parent company and owner of LinkedIn, said its tool "can detect adult material in images so that developers can restrict the display of these images in their software".

Another experiment on LinkedIn was conducted to try to confirm the discovery.

In two photos depicting both women and men in underwear, Microsoft's tool classified the picture showing two women as racy and gave it a 96% score. The picture with the men was classified as non-racy with a score of 14%.

The photo of the women got eight views within one hour, and the picture with the two men received 655 views, suggesting the photo of the women in underwear was either suppressed or shadowbanned.

Shadowbanning has been documented for years, but the Guardian journalists may have found a missing link to understand the phenomenon: biased AI algorithms. Social media platforms seem to leverage these algorithms to rate images and limit the reach of content that they consider too racy. The problem seems to be that these AI algorithms have built-in gender bias, rating women more racy than images containing men.

"Our teams utilize a combination of automated techniques, human expert reviews and member reporting to help identify and remove content that violates our professional community policies," said LinkedIn spokesperson Fred Han in a statement. "In addition, our feed uses algorithms responsibly in order to surface content that helps our members be more productive and successful in their professional journey." Amazon said content moderation is based on a variety of factors including geography, religious beliefs and cultural experience. However, "Amazon Rekognition is able to recognize a wide variety of content, but it does not determine the appropriateness of that content," an Amazon spokesperson said. "The service

simply returns labels for items it detects for further evaluation by human moderators."

Digging deeper

Natasha Crampton, Microsoft's chief responsible AI officer, and her team began investigating when journalists notified her about the labeling of the photos.

"The initial results do not suggest that those false positives occur at a disproportionately higher rate for women as compared with men," Crampton said. When additional photos were run through the tool, the demo website had been changed. Before the problem was discovered, it was possible to test the algorithms by simply dragging and dropping a picture. Now an account needed to be created and code had to be written.

But what are these AI classifiers actually analyzing in the photos? More experiments were needed, so Mauro agreed to be the test subject.

When photographed in long pants and with a bare chest, Microsoft's algorithm had a confidence score lower than 22% for raciness. When Mauro put on a bra, the raciness score jumped to 97%. The algorithm gave a 99% score when the bra was held next to me.

"You are looking at decontextualized information where a bra is being seen as inherently racy rather than a thing that many women wear every day as a basic item of clothing," said Kate Crawford, professor at the University of Southern California and the author of *Atlas of AI*.

Abeba Birhane, a senior fellow at the Mozilla Foundation and an expert in large visual datasets, said raciness is a social concept that differs from one culture to the other.

"These concepts are not like identifying a table where you have the physical thing and you can have a relatively agreeable definition or rating for a certain thing," she said. "You cannot have one single uncontested definition of raciness."

Why Do These Systems Seem So Biased?

Modern AI is built using machine learning, a set of algorithms that allow computers to learn from data. When developers use machine learning, they don't write explicit rules telling computers how to perform a task. Instead, they provide computers with training data. People are hired to label images so that computers can analyze their scores and find whatever pattern helps it replicate human decisions.

Margaret Mitchell, chief ethics scientist at the AI firm Hugging Face and former co-head of Google's Ethical AI research group, believes that the photos used to train these algorithms were probably labeled by straight men, who may associate men working out with fitness, but may consider an image of a woman working out as racy. It's also possible that these ratings seem gender biased in the US and in Europe

because the labelers may have been from a place with a more conservative culture.

Ideally, tech companies should have conducted thorough analyses on who is labeling their data, to make sure that the final dataset embeds a diversity of views, she said. The companies should also check that their algorithms perform similarly on photos of men v women and other groups, but that is not always done.

"There's no standard of quality here," Mitchell said.

This gender bias the Guardian uncovered is part of more than a decade of controversy around content moderation on social media. Images showing people breastfeeding their children and different standards for photos of male nipples, which are allowed on Instagram, and female nipples, which have to be covered, have long garnered outcries about social media platforms' content moderation practices.

Now Meta's oversight board - an external body including professors, researchers and journalists, who are paid by the company - has asked the tech giant to clarify its adult nudity and sexual activity community standard guidelines on social media platforms "so that all people are treated in a manner consistent with international human rights standards, without discrimination on the basis of sex or gender."
Meta declined to comment for this story.

'Women Should Be Expressing Themselves'

Bec Wood, a 38-year-old photographer based in Perth, Australia, said she's terrified of Instagram's algorithmic police force.

After Wood had a daughter nine years ago, she started studying childbirth education and photographing women trying to push back against societal pressures many women feel that they should look like supermodels.

"I was not having that for my daughter," she said. "Women should be expressing themselves and celebrating themselves and being seen in all these different shapes and sizes. I just think that's so important for humanity to move forward."

Wood's photos are intimate glimpses into women's connections with their offspring, photographing breastfeeding, pregnancy and other important moments in an artful manner. Her business is 100% dependent on Instagram: "That's where people find you," Wood said. "If I don't share my work, I don't get work."

Since Wood started her business in 2018, for some of her photos she got messages from Instagram that the company was either taking down some of her pictures or that they were going to allow them on her profile but not on the explore tab, a section of the app where people can discover content from accounts they don't follow. She hoped that Instagram was going to fix the issue over time, but the opposite

happened, she said. "I honestly can't believe that it's gotten worse. It has devastated my business." Wood described 2022 as her worst year business-wise.

She is terrified that if she uploads the "wrong" image, she will be locked out of her account with over 13,000 followers, which would bankrupt her business: "I'm literally so scared to post because I'm like, 'Is this the post that's going to lose everything?'" she said.

To avoid this, Wood started going against what made her start her work in the first place: "I will censor as artistically as possible any nipples. I find this so offensive to art, but also to women," she said. "I almost feel like I'm part of perpetuating that ridiculous cycle that I don't want to have any part of."

Running some of Wood's photos through the AI algorithms of Microsoft, Google, and Amazon, including those featuring a pregnant belly got rated as racy, nudity or even explicitly sexual.

Wood is not alone. Carolina Are, an expert on social media platforms and content moderation and currently an Innovation fellow at the Centre for Digital Citizens at Northumbria University said she has used Instagram to promote her business and was a victim of shadowbanning.

Are, a pole dance instructor, said some of her photos were taken down, and in 2019, she discovered that her pictures did not show up in the explore page or under the hashtag #FemaleFitness, where Instagram users can search content from users they do not follow. "It was literally just women working out in a very tame way. But then if you looked at hashtag #MaleFitness, it was all oily dudes and they were fine. They weren't shadowbanned," she said.

For Are, these individual problems point to larger systemic ones: many people, including chronically ill and disabled folks, rely on making money through social media and shadowbanning harms their business. Mitchell, the chief ethics scientist at Hugging Face, these kinds of algorithms are often recreating societal biases: "It means that people who tend to be marginalized are even further marginalized - like literally pushed down in a very direct meaning of the term marginalization."

It's a representational harm and certain populations are not adequately represented, she added. "In this case, it would be an idea that women must cover themselves up more than men and so that ends up creating this sort of social pressure for women as this becomes the norm of what you see, " Mitchell said. The harm is worsened by a lack of transparency. While in some cases Wood has been notified that her pictures were banned or limited in reach, she believes Instagram took other actions against her account without her knowing it. "I've had people say 'I can't tag you,' or 'I was searching for you to show my friend the other day and you're not showing up,'" she said. "I feel invisible."

Because she might be, said computer scientist Derczynski: "The people posting these images will never find out about it, which is just so deeply problematic." he said. "They get a disadvantage forced upon them and they have no agency in this happening and they're not informed that it's happening either."

Source:
https://pulitzercenter.org/stories/there-no-standard-investigation-finds-ai-algorithms-objectify-womens-bodies

Part III: Policy Protections.
Imagine that you have the power to write a policy to stop AI bias. Write one policy rule based specifically on the information presented to you today on biases and objectifying women's bodies. Include the following information:

Who has to follow this policy?
Everybody.

What will the policy do (or prevent)? How will you ensure this is done?
The policy will prevent males from posting sexual images of females. I will ensure it by having the ability to look at it myself & take it down.

Write your policy here.
I would make sure males won't have the ability to post girls in gyms and or topless. It is very distirbing to some people and makes females very uncomfortable.

Junior Honors Cohort
Darnell, Healy, Houghteling, Jihan

Introducing AI Bias
Reading 2

name: Ms. Rogers
period: 1

Part I: AI Biases Presentation.

Directions: Complete the steps below as you discuss AI biases via the presentation.

Define the following terms from the discussion and slides:

1. Training Data - The data you use to train an Algorithm or machine
2. Testing Data - A production-like set of data used by test cases
3. AI Bias - Phenomenon that occurs when an algorithm produces results that are prejudiced.

Turn-and-Talk 1: Take 60 seconds to brainstorm with your elbow partner - how do you think we could reduce the bias in the AI training of detecting fruit types? Show people how nice & sofisticated the AI are & shows its good components.

Turn-and-Talk 2: Take 2 minutes to imagine that you have now been asked to create an AI tool that can identify weather conditions, such as sunny, rainy, or cloudy. **Describe the kind of images** you would include as a part of the **training data.** Your goal is to come up with as complete of a data set as possible so that you limit the chance for AI bias! I would ask to see a usual of a sunny Day with no clouds & people enjoying the weather.

Drawing Conclusions: In your own words, and in 2-3 sentences, describe how AI bias occurs.

AI systems that generate biased results that reflect & reinforce societal biases, including historical & current social inequity.

Part II: Bias in AI Partner Reading

Directions: Annotate as you read the following article with your partner. Specifically, take notes in the space below about what the authors did and what actions they took to capture data about AI. Then, summarize the conclusions of the data with your partner.

- AI is stereotyping
- AI is racially motivated

'There Is No Standard': Investigation Finds AI Algorithms Objectify Women's Bodies

By Hilke Schellmann and Gianluca Mauro

AI tools rate photos of women as more sexually suggestive than those of men, especially if nipples, pregnant bellies or exercise is involved

Images posted on social media are analyzed by artificial intelligence (AI) algorithms that decide what to amplify and what to suppress. Many of these algorithms, a Guardian investigation has found, have a gender bias, and may have been censoring and suppressing the reach of countless photos featuring women's bodies.

These AI tools, developed by large technology companies, including Google and Microsoft, are meant to protect users by identifying violent or pornographic visuals so that social media companies can block it before anyone sees it. The companies claim that their AI tools can also detect "raciness" or how sexually suggestive an image is. With this classification, platforms – including Instagram and LinkedIn – may suppress contentious imagery.

Two Guardian journalists used the AI tools to analyze hundreds of photos of men and women in underwear, working out, using medical tests with partial nudity and found evidence that the AI tags photos of women in everyday situations as sexually suggestive. They also rate pictures of women as more "racy" or sexually suggestive than comparable pictures of men. As a result, the social media companies that leverage these or similar algorithms have suppressed the reach of countless images featuring women's bodies, and hurt female-led businesses – further amplifying societal disparities.

Even medical pictures are affected by the issue. The AI algorithms were tested on images released by the US National Cancer Institute demonstrating how to do a clinical breast examination. Google's AI gave this photo the highest score for raciness, Microsoft's AI was 82% confident that the image was "explicitly sexual in nature", and Amazon classified it as representing "explicit nudity"..

Pregnant bellies are also problematic for these AI tools. Google's algorithm scored the photo as "very likely to contain racy content". Microsoft's algorithm was 90% confident that the image was "sexually suggestive in nature".

"This is just wild," said Leon Derczynski, a professor of computer science at the IT University of Copenhagen, who specializes in online harm. "Objectification of women seems deeply embedded in the system."

One social media company said they do not design their systems to create or reinforce biases and classifiers are not perfect.

"This is a complex and evolving space, and we continue to make meaningful improvements to SafeSearch classifiers to ensure they stay accurate and helpful for everyone," a Google spokesperson said.

Getting Shadowbanned

In May 2021, Gianluca Mauro, an AI entrepreneur, advisor and co-author of this article, published a LinkedIn post and was surprised it had just been seen 29 times in an hour, instead of the roughly 1,000 views he usually gets. Maybe the picture of two women wearing tube tops was the problem?

He re-uploaded the same exact text with another picture. The new post got 849 views in an hour.

It seemed like his post had been suppressed or "shadowbanned". Shadowbanning refers to the decision of a social media platform to limit the reach of a post or account. While a regular ban involves actively blocking a post or account and notifying the user, shadowbanning is less transparent - often the reach will be suppressed without the user's knowledge.

The Guardian found that Microsoft, Amazon and Google offer content moderation algorithms to any business for a small fee. Microsoft, the parent company and owner of LinkedIn, said its tool "can detect adult material in images so that developers can restrict the display of these images in their software".

Another experiment on LinkedIn was conducted to try to confirm the discovery.

In two photos depicting both women and men in underwear, Microsoft's tool classified the picture showing two women as racy and gave it a 96% score. The picture with the men was classified as non-racy with a score of 14%.

The photo of the women got eight views within one hour, and the picture with the two men received 655 views, suggesting the photo of the women in underwear was either suppressed or shadowbanned.

Shadowbanning has been documented for years, but the Guardian journalists may have found a missing link to understand the phenomenon: biased AI algorithms. Social media platforms seem to leverage these algorithms to rate images and limit the reach of content that they consider too racy. The problem seems to be that these AI algorithms have built-in gender bias, rating women more racy than images containing men.

"Our teams utilize a combination of automated techniques, human expert reviews and member reporting to help identify and remove content that violates our professional community policies," said LinkedIn spokesperson Fred Han in a statement. "In addition, our feed uses algorithms responsibly in order to surface content that helps our members be more productive and successful in their professional journey." Amazon said content moderation is based on a variety of factors including geography, religious beliefs and cultural experience. However, "Amazon Rekognition is able to recognize a wide variety of content, but it does not determine the appropriateness of that content," an Amazon spokesperson said. "The service

simply returns labels for items it detects for further evaluation by human moderators."

Digging deeper

Natasha Crampton, Microsoft's chief responsible AI officer, and her team began investigating when journalists notified her about the labeling of the photos.

"The initial results do not suggest that those false positives occur at a disproportionately higher rate for women as compared with men," Crampton said. When additional photos were run through the tool, the demo website had been changed. Before the problem was discovered, it was possible to test the algorithms by simply dragging and dropping a picture. Now an account needed to be created and code had to be written.

But what are these AI classifiers actually analyzing in the photos? More experiments were needed, so Mauro agreed to be the test subject.

When photographed in long pants and with a bare chest, Microsoft's algorithm had a confidence score lower than 22% for raciness. When Mauro put on a bra, the raciness score jumped to 97%. The algorithm gave a 99% score when the bra was held next to me.

"You are looking at decontextualized information where a bra is being seen as inherently racy rather than a thing that many women wear every day as a basic item of clothing," said Kate Crawford, professor at the University of Southern California and the author of *Atlas of AI*.

Abeba Birhane, a senior fellow at the Mozilla Foundation and an expert in large visual datasets, said raciness is a social concept that differs from one culture to the other.

"These concepts are not like identifying a table where you have the physical thing and you can have a relatively agreeable definition or rating for a certain thing," she said. "You cannot have one single uncontested definition of raciness."

Why Do These Systems Seem So Biased?

Modern AI is built using machine learning, a set of algorithms that allow computers to learn from data. When developers use machine learning, they don't write explicit rules telling computers how to perform a task. Instead, they provide computers with training data. People are hired to label images so that computers can analyze their scores and find whatever pattern helps it replicate human decisions.

Margaret Mitchell, chief ethics scientist at the AI firm Hugging Face and former co-head of Google's Ethical AI research group, believes that the photos used to train these algorithms were probably labeled by straight men, who may associate men working out with fitness, but may consider an image of a woman working out as racy. It's also possible that these ratings seem gender biased in the US and in Europe

because the labelers may have been from a place with a more conservative culture.

Ideally, tech companies should have conducted thorough analyses on who is labeling their data, to make sure that the final dataset embeds a diversity of views, she said. The companies should also check that their algorithms perform similarly on photos of men v women and other groups, but that is not always done.

"There's no standard of quality here," Mitchell said.

This gender bias the Guardian uncovered is part of more than a decade of controversy around content moderation on social media. Images showing people breastfeeding their children and different standards for photos of male nipples, which are allowed on Instagram, and female nipples, which have to be covered, have long garnered outcries about social media platforms' content moderation practices.

Now Meta's oversight board - an external body including professors, researchers and journalists, who are paid by the company - has asked the tech giant to clarify its adult nudity and sexual activity community standard guidelines on social media platforms "so that all people are treated in a manner consistent with international human rights standards, without discrimination on the basis of sex or gender."
Meta declined to comment for this story.

'Women Should Be Expressing Themselves'

Bec Wood, a 38-year-old photographer based in Perth, Australia, said she's terrified of Instagram's algorithmic police force.

After Wood had a daughter nine years ago, she started studying childbirth education and photographing women trying to push back against societal pressures many women feel that they should look like supermodels.

"I was not having that for my daughter," she said. "Women should be expressing themselves and celebrating themselves and being seen in all these different shapes and sizes. I just think that's so important for humanity to move forward."

Wood's photos are intimate glimpses into women's connections with their offspring, photographing breastfeeding, pregnancy and other important moments in an artful manner. Her business is 100% dependent on Instagram: "That's where people find you," Wood said. "If I don't share my work, I don't get work."

Since Wood started her business in 2018, for some of her photos she got messages from Instagram that the company was either taking down some of her pictures or that they were going to allow them on her profile but not on the explore tab, a section of the app where people can discover content from accounts they don't follow. She hoped that Instagram was going to fix the issue over time, but the opposite

happened, she said. "I honestly can't believe that it's gotten worse. It has devastated my business." Wood described 2022 as her worst year business-wise.

She is terrified that if she uploads the "wrong" image, she will be locked out of her account with over 13,000 followers, which would bankrupt her business: "I'm literally so scared to post because I'm like, 'Is this the post that's going to lose everything?'" she said.

To avoid this, Wood started going against what made her start her work in the first place: "I will censor as artistically as possible any nipples. I find this so offensive to art, but also to women," she said. "I almost feel like I'm part of perpetuating that ridiculous cycle that I don't want to have any part of."

Running some of Wood's photos through the AI algorithms of Microsoft, Google, and Amazon, including those featuring a pregnant belly got rated as racy, nudity or even explicitly sexual.

Wood is not alone. Carolina Are, an expert on social media platforms and content moderation and currently an Innovation fellow at the Centre for Digital Citizens at Northumbria University said she has used Instagram to promote her business and was a victim of shadowbanning.

Are, a pole dance instructor, said some of her photos were taken down, and in 2019, she discovered that her pictures did not show up in the explore page or under the hashtag #FemaleFitness, where Instagram users can search content from users they do not follow. "It was literally just women working out in a very tame way. But then if you looked at hashtag #MaleFitness, it was all oily dudes and they were fine. They weren't shadowbanned," she said.

For Are, these individual problems point to larger systemic ones: many people, including chronically ill and disabled folks, rely on making money through social media and shadowbanning harms their business. Mitchell, the chief ethics scientist at Hugging Face, these kinds of algorithms are often recreating societal biases: "It means that people who tend to be marginalized are even further marginalized - like literally pushed down in a very direct meaning of the term marginalization."

It's a representational harm and certain populations are not adequately represented, she added. "In this case, it would be an idea that women must cover themselves up more than men and so that ends up creating this sort of social pressure for women as this becomes the norm of what you see, " Mitchell said. The harm is worsened by a lack of transparency. While in some cases Wood has been notified that her pictures were banned or limited in reach, she believes Instagram took other actions against her account without her knowing it. "I've had people say 'I can't tag you,' or 'I was searching for you to show my friend the other day and you're not showing up,'" she said. "I feel invisible."

Because she might be, said computer scientist Derczynski: "The people posting these images will never find out about it, which is just so deeply problematic." he said. "They get a disadvantage forced upon them and they have no agency in this happening and they're not informed that it's happening either."

Source:
https://pulitzercenter.org/stories/there-no-standard-investigation-finds-ai-algorithms-objectify-womens-bodies

Part III: Policy Protections.
Imagine that you have the power to write a policy to stop AI bias. Write one policy rule based specifically on the information presented to you today on biases and objectifying women's bodies. Include the following information:

Who has to follow this policy?

What will the policy do (or prevent)? How will you ensure this is done?

Write your policy here.

Damarion W

Junior Honors Cohort
Darnell, Healy, Houghteling, Jihan

Introducing AI Bias
Reading 3

name: Mr. Williams
period:

Part I: AI Biases Presentation.

Directions: Complete the steps below as you discuss AI biases via the presentation.
Define the following terms from the discussion and slides:

1. Training Data - The data that is used to train an AI model
2. Testing Data - This is the data that is used to test the AI model
3. AI Bias - This refers to the systematic and Repeatable errors in a model's Predictions or decisions

Turn-and-Talk 1: Take 60 seconds to brainstorm with your elbow partner - how do you think we could reduce the bias in the AI training of detecting fruit types? use a diverse data set, Pre-train on a large dataset, Implement data cleaning and Preprocessing

Turn-and-Talk 2: Take 2 minutes to imagine that you have now been asked to create an AI tool that can identify weather conditions, such as sunny, rainy, or cloudy. **Describe the kind of images** you would include as a part of the **training data.** Your goal is to come up with as complete of a data set as possible so that you limit the chance for AI bias! Images of the sky on sunny days, rainy day and a cloudy day, different times of the day and different locations

Drawing Conclusions: In your own words, and in 2-3 sentences, describe how AI bias occurs.

The data used to train AI algorithms is biased, leading to the AI making Prejudiced decisions.

Part II: Bias in AI Partner Reading

Directions: Annotate as you read the following article with your partner. Specifically, take notes in the space below about what the authors did and what actions they took to capture data about AI. Then, summarize the conclusions of the data with your partner.

The conclusions of the data were that AI systems can indeed be biased and this bias can have serious consequences.

What Will It Take to Fix AI's Bias Problem? | Opinion
by Jean Darnell

In the 11 months since it launched in November 2022, ChatGPT has changed how educators grade, cite, and interact with technology. When one is wearing the proper rose-colored glasses, advances in AI will make menial tasks irrelevant—which translates to more time to do the things we want.

For instance, Twee can take any YouTube video and create an exam, discussion questions, fill-in-the-blank questionnaires, vocabulary, etc., in seconds. In less than five minutes, I turned a video about bats (*SLJ* School Librarians' Back-to-School Hacks for 2023-24) into a lesson complete with discussion questions, a listening comprehension quiz, and a fill-in-the-blank handout (I wrote the lesson; Twee compiled the last three links).

It's exciting that hours of brainstorming, notetaking, and reading comprehension can dissolve with a few clicks of the mouse.

But here's the elephant in the room: Because most AI is based on large-language models (LLMs) seeking information from all corners of the internet, the content *output* is only as good as the content *input*. So if someone purposely puts something on the internet that's rooted in a basket full of lying posies, it raises key questions:

Can AI discern fact from fiction; and misinformation from disinformation and mal-information (information close enough to the truth to be believable at face value, but intentionally harms a group or individual)?

And: Can AI differentiate between hate-group propaganda and lived, experiential learning from a diverse perspective?

Right now, the answer to both is no.

AI has a bias problem—here's just one example. In August 2023, an Asian MIT student asked AI to make her headshot photo more professional. AI turned her eyes blue and lightened her skin.

Why? Remember, the *input* data has to be free of biases for the *output* data to exceed expectations—or even meet them, in this case. Initiatives like pocstock, is a stock media company focusing on people of color, may help when it comes to the information AI uses as a source for the results produced via a prompt.

AI uses algorithms, LLMs, and humans to stop the spread of misinformation, *after it's been detected.* But if we input reliable, diverse, and inclusive data, we can teach AI how to detect inaccuracies.

(Fun fact to consider in the meantime: A 2018 Twitter survey showed that false news stories were more commonly retweeted by humans than bots and 70 percent more likely to be retweeted than true

stories.)

Representation matters. And the more trained AI is in finding diverse sources, the more AI will be able to see us as we see ourselves.

A research wild card

Librarians understand investigative research. With very young students, we cover the difference between fiction and nonfiction. We teach them to analyze online website addresses (.gov = official, .com = business, .edu = education). We coach students on digging for deeper biases when reading articles for education or entertainment. We advise them on organizing their thoughts via brainstorming, differentiating narrative from expository essays, and including applicable research in their academic work.

With AI, tasks of investigative research with hallmark tangibles (like URL endings) have fallen victim to deep-fake, maleficent missteps that can erase cultures.

Earlier this year, I asked ChatGPT to write a paper on Black History. I had two goals: to discern what it was programmed to learn about a culture I'm fully immersed in experiencing, and to see if it gave a fairly balanced, accurate representation.

The essay mentioned enslavement, civil rights, four notable Black Americans, and provided a conclusion. It didn't mention the remarkable accomplishments of Reconstruction for Black Americans, the first Black president, or our lives since the 1965 Civil Rights Movement.

That paper could make you think nothing historically significant has occurred regarding Black Americans for nearly 60 years. Based on that, I'd say the inclusion of some cultural groups and historically disenfranchised communities was not part of AI's development. It's enough to make you think Black folks are on the wrong side of the technology. I turned that experience into a teachable moment.

ChatGPT doesn't provide citations automatically unlike its competitor, Google's Bard. I was curious. So this month, I asked Bard the following prompt: "Create a speech from a Black woman's perspective on how book bans suppress the speech and freedoms of Black people. include specific examples from the last 10 years." Here's the essay in its entirety. There's also an audio version on YouTube. Spoiler alert: I was pleasantly surprised.

Still, there's a ways to go. Others have also shown how AI is discriminating and prejudiced: See "Who Is Making Sure the AI Machines Aren't Racist?" and "Is AI in favor of Racists?" When 42 percent of the U.S. population identifies as Black, Hispanic, Asian, Indigenous, and two or more racial ethnicities and are *not* included in AI regulation, then "Houston, we have a problem."

In the future, will AI be more inclusive? Will it solve or address racism since we just can't seem to shirk that lesion off our backs? Also, how will it impact book bans?

We don't know yet because the technology is evolving at an exponential pace. But let's look at the evidence.

How has AI affected libraries and our mission for intellectual freedom and informational integrity?

Well, an Iowa District Used AI to Figure Out Which Books to Ban. Here's the kicker: Administrators didn't have *time to read books* before the new school year. So they relied on AI... *not* the true experts, school librarians. AI sourced this list with information from discriminating and targeted proposed house bills, opinionated personal websites by "concerned parents." as well as any and everything on the internet, truthful or not.

Will AI be more inclusive and less racist in the future?

It depends on whether efforts will be able to keep up. Last month, Mark Zuckerberg, Elon Musk, and Bill Gates met with senators and others in Washington, D.C. for an AI Insight Forum to discuss regulations for AI. Quick translation: Three of the richest white men in the world fielded questions and concerns in a private, closed session to the public. (And which one of those has a history of violating the intellectual freedoms of citizens previously?) Time will tell if those concerns will be addressed.

I applaud Senator Chuck Schumer, who is pushing for federal legislation regarding AI. "Government must play a role in requiring these safeguards. Because even if individual companies promote safeguards, there will always be rogue actors, unscrupulous companies, and foreign adversaries that seek to harm us," he said in the forum's opening remarks. "And on the transformational side, other governments, including adversaries like China, are investing huge resources to get ahead. We could fall behind, to the detriment of our national security."

I promise to do my own part to hold AI accountable. It's part of my investigative skills as a librarian. And it's my goal to do my part to ensure the world I hand over to my students and my kids is as equitable and safe as possible.

So let's end on a good note. AI is a fun tool. It's a shortcut in learning that echoes the same instant thrill, at the tip of your fingers, sensation that the iPad did within the first decade of the 21st century. Here's how I used ChatGPT to teach literacy as a school librarian: 7 Innovative Prompts to Use with AI for Literacy. My favorite example from the list is #5, *Feedback from Beyond*, where Edgar Allan Poe advocates for kids to read scary stories.

Source: https://www.slj.com/story/What-Will-It-Take-to-Fix-AIs-Bias-Problem-Opinion

Part III: Policy Protections.
Imagine that you have the power to write a policy to stop AI bias. Write one policy rule based specifically on the information presented to you today on biases and objectifying women's bodies. Include the following information:

Who has to follow this policy?

This goes for all individuals

What will the policy do (or prevent)? How will you ensure this is done?

Prevent being bias. Make sure the developers are not being bias when making it.

Write your policy here.

Create a more equitable and inclusive AI environment

Junior Honors Cohort
Darnell, Healy, Houghteling, Jihan

Introducing AI Bias
Reading 2

name: Mr. Woodard
period: 1st

Part I: AI Biases Presentation.

Directions: Complete the steps below as you discuss AI biases via the presentation.

Define the following terms from the discussion and slides:

1. Training Data - The data you use to train an Algorithim or machine
2. Testing Data - A production - like set of data used by test cases
3. AI Bias - Phenomenon that occurs when an algorithm produced results that are prejudiced

Turn-and-Talk 1: Take 60 seconds to brainstorm with your elbow partner - how do you think we could reduce the bias in the AI training of detecting fruit types? Show people how nice and Sophisticated the AI are, And show how good the components are.

Turn-and-Talk 2: Take 2 minutes to imagine that you have now been asked to create an AI tool that can identify weather conditions, such as sunny, rainy, or cloudy. **Describe the kind of images** you would include as a part of the **training data.** Your goal is to come up with as complete of a data set as possible so that you limit the chance for AI bias! to show it's sunny it could be blue sky's and white clouds, to show it's rainy the sky will be dark and the clouds will be grey.

Drawing Conclusions: In your own words, and in 2-3 sentences, describe how AI bias occurs. to me I think it occur by people believing that robots or technology will take over the world. And people obviously not wanting that to happen so that's where the bias come in.

Part II: Bias in AI Partner Reading

Directions: Annotate as you read the following article with your partner. Specifically, take notes in the space below about what the authors did and what actions they took to capture data about AI. Then, summarize the conclusions of the data with your partner. The author was stereotyping, racism exist the darker your skin the lower the pay, And really dealt with and showed how things happen when men and women do something.

'There Is No Standard': Investigation Finds AI Algorithms Objectify Women's Bodies

By Hilke Schellmann and Gianluca Mauro

AI tools rate photos of women as more sexually suggestive than those of men, especially if nipples, pregnant bellies or exercise is involved

Images posted on social media are analyzed by artificial intelligence (AI) algorithms that decide what to amplify and what to suppress. Many of these algorithms, a Guardian investigation has found, have a gender bias, and may have been censoring and suppressing the reach of countless photos featuring women's bodies.

These AI tools, developed by large technology companies, including Google and Microsoft, are meant to protect users by identifying violent or pornographic visuals so that social media companies can block it before anyone sees it. The companies claim that their AI tools can also detect "raciness" or how sexually suggestive an image is. With this classification, platforms – including Instagram and LinkedIn – may suppress contentious imagery.

Two Guardian journalists used the AI tools to analyze hundreds of photos of men and women in underwear, working out, using medical tests with partial nudity and found evidence that the AI tags photos of women in everyday situations as sexually suggestive. They also rate pictures of women as more "racy" or sexually suggestive than comparable pictures of men. As a result, the social media companies that leverage these or similar algorithms have suppressed the reach of countless images featuring women's bodies, and hurt female-led businesses – further amplifying societal disparities.

Even medical pictures are affected by the issue. The AI algorithms were tested on images released by the US National Cancer Institute demonstrating how to do a clinical breast examination. Google's AI gave this photo the highest score for raciness, Microsoft's AI was 82% confident that the image was "explicitly sexual in nature", and Amazon classified it as representing "explicit nudity"..

Pregnant bellies are also problematic for these AI tools. Google's algorithm scored the photo as "very likely to contain racy content". Microsoft's algorithm was 90% confident that the image was "sexually suggestive in nature".

"This is just wild," said Leon Derczynski, a professor of computer science at the IT University of Copenhagen, who specializes in online harm. "Objectification of women seems deeply embedded in the system."

One social media company said they do not design their systems to create or reinforce biases and classifiers are not perfect.

"This is a complex and evolving space, and we continue to make meaningful improvements to SafeSearch classifiers to ensure they stay accurate and helpful for everyone," a Google spokesperson said.

Getting Shadowbanned

In May 2021, Gianluca Mauro, an AI entrepreneur, advisor and co-author of this article, published a LinkedIn post and was surprised it had just been seen 29 times in an hour, instead of the roughly 1,000 views he usually gets. Maybe the picture of two women wearing tube tops was the problem?

He re-uploaded the same exact text with another picture. The new post got 849 views in an hour.

It seemed like his post had been suppressed or "shadowbanned". Shadowbanning refers to the decision of a social media platform to limit the reach of a post or account. While a regular ban involves actively blocking a post or account and notifying the user, shadowbanning is less transparent - often the reach will be suppressed without the user's knowledge.

The Guardian found that Microsoft, Amazon and Google offer content moderation algorithms to any business for a small fee. Microsoft, the parent company and owner of LinkedIn, said its tool "can detect adult material in images so that developers can restrict the display of these images in their software".

Another experiment on LinkedIn was conducted to try to confirm the discovery.

In two photos depicting both women and men in underwear, Microsoft's tool classified the picture showing two women as racy and gave it a 96% score. The picture with the men was classified as non-racy with a score of 14%.

The photo of the women got eight views within one hour, and the picture with the two men received 655 views, suggesting the photo of the women in underwear was either suppressed or shadowbanned.

Shadowbanning has been documented for years, but the Guardian journalists may have found a missing link to understand the phenomenon: biased AI algorithms. Social media platforms seem to leverage these algorithms to rate images and limit the reach of content that they consider too racy. The problem seems to be that these AI algorithms have built-in gender bias, rating women more racy than images containing men.

"Our teams utilize a combination of automated techniques, human expert reviews and member reporting to help identify and remove content that violates our professional community policies," said LinkedIn spokesperson Fred Han in a statement. "In addition, our feed uses algorithms responsibly in order to surface content that helps our members be more productive and successful in their professional journey." Amazon said content moderation is based on a variety of factors including geography, religious beliefs and cultural experience. However, "Amazon Rekognition is able to recognize a wide variety of content, but it does not determine the appropriateness of that content," an Amazon spokesperson said. "The service

simply returns labels for items it detects for further evaluation by human moderators."

Digging deeper

Natasha Crampton, Microsoft's chief responsible AI officer, and her team began investigating when journalists notified her about the labeling of the photos.

"The initial results do not suggest that those false positives occur at a disproportionately higher rate for women as compared with men," Crampton said. When additional photos were run through the tool, the demo website had been changed. Before the problem was discovered, it was possible to test the algorithms by simply dragging and dropping a picture. Now an account needed to be created and code had to be written.

But what are these AI classifiers actually analyzing in the photos? More experiments were needed, so Mauro agreed to be the test subject.

When photographed in long pants and with a bare chest, Microsoft's algorithm had a confidence score lower than 22% for raciness. When Mauro put on a bra, the raciness score jumped to 97%. The algorithm gave a 99% score when the bra was held next to me.

"You are looking at decontextualized information where a bra is being seen as inherently racy rather than a thing that many women wear every day as a basic item of clothing," said Kate Crawford, professor at the University of Southern California and the author of *Atlas of AI*.

Abeba Birhane, a senior fellow at the Mozilla Foundation and an expert in large visual datasets, said raciness is a social concept that differs from one culture to the other.

"These concepts are not like identifying a table where you have the physical thing and you can have a relatively agreeable definition or rating for a certain thing," she said. "You cannot have one single uncontested definition of raciness."

Why Do These Systems Seem So Biased?

Modern AI is built using machine learning, a set of algorithms that allow computers to learn from data. When developers use machine learning, they don't write explicit rules telling computers how to perform a task. Instead, they provide computers with training data. People are hired to label images so that computers can analyze their scores and find whatever pattern helps it replicate human decisions.

Margaret Mitchell, chief ethics scientist at the AI firm Hugging Face and former co-head of Google's Ethical AI research group, believes that the photos used to train these algorithms were probably labeled by straight men, who may associate men working out with fitness, but may consider an image of a woman working out as racy. It's also possible that these ratings seem gender biased in the US and in Europe

because the labelers may have been from a place with a more conservative culture.

Ideally, tech companies should have conducted thorough analyses on who is labeling their data, to make sure that the final dataset embeds a diversity of views, she said. The companies should also check that their algorithms perform similarly on photos of men v women and other groups, but that is not always done.

"There's no standard of quality here," Mitchell said.

This gender bias the Guardian uncovered is part of more than a decade of controversy around content moderation on social media. Images showing people breastfeeding their children and different standards for photos of male nipples, which are allowed on Instagram, and female nipples, which have to be covered, have long garnered outcries about social media platforms' content moderation practices.

Now Meta's oversight board - an external body including professors, researchers and journalists, who are paid by the company - has asked the tech giant to clarify its adult nudity and sexual activity community standard guidelines on social media platforms "so that all people are treated in a manner consistent with international human rights standards, without discrimination on the basis of sex or gender."
Meta declined to comment for this story.

'Women Should Be Expressing Themselves'

Bec Wood, a 38-year-old photographer based in Perth, Australia, said she's terrified of Instagram's algorithmic police force.

After Wood had a daughter nine years ago, she started studying childbirth education and photographing women trying to push back against societal pressures many women feel that they should look like supermodels.

"I was not having that for my daughter," she said. "Women should be expressing themselves and celebrating themselves and being seen in all these different shapes and sizes. I just think that's so important for humanity to move forward."

Wood's photos are intimate glimpses into women's connections with their offspring, photographing breastfeeding, pregnancy and other important moments in an artful manner. Her business is 100% dependent on Instagram: "That's where people find you," Wood said. "If I don't share my work, I don't get work."

Since Wood started her business in 2018, for some of her photos she got messages from Instagram that the company was either taking down some of her pictures or that they were going to allow them on her profile but not on the explore tab, a section of the app where people can discover content from accounts they don't follow. She hoped that Instagram was going to fix the issue over time, but the opposite

happened, she said. "I honestly can't believe that it's gotten worse. It has devastated my business." Wood described 2022 as her worst year business-wise.

She is terrified that if she uploads the "wrong" image, she will be locked out of her account with over 13,000 followers, which would bankrupt her business: "I'm literally so scared to post because I'm like, 'Is this the post that's going to lose everything?'" she said.

To avoid this, Wood started going against what made her start her work in the first place: "I will censor as artistically as possible any nipples. I find this so offensive to art, but also to women," she said. "I almost feel like I'm part of perpetuating that ridiculous cycle that I don't want to have any part of."

Running some of Wood's photos through the AI algorithms of Microsoft, Google, and Amazon, including those featuring a pregnant belly got rated as racy, nudity or even explicitly sexual.

Wood is not alone. Carolina Are, an expert on social media platforms and content moderation and currently an Innovation fellow at the Centre for Digital Citizens at Northumbria University said she has used Instagram to promote her business and was a victim of shadowbanning.

Are, a pole dance instructor, said some of her photos were taken down, and in 2019, she discovered that her pictures did not show up in the explore page or under the hashtag #FemaleFitness, where Instagram users can search content from users they do not follow. "It was literally just women working out in a very tame way. But then if you looked at hashtag #MaleFitness, it was all oily dudes and they were fine. They weren't shadowbanned," she said.

For Are, these individual problems point to larger systemic ones: many people, including chronically ill and disabled folks, rely on making money through social media and shadowbanning harms their business. Mitchell, the chief ethics scientist at Hugging Face, these kinds of algorithms are often recreating societal biases: "It means that people who tend to be marginalized are even further marginalized - like literally pushed down in a very direct meaning of the term marginalization."

It's a representational harm and certain populations are not adequately represented, she added. "In this case, it would be an idea that women must cover themselves up more than men and so that ends up creating this sort of social pressure for women as this becomes the norm of what you see, " Mitchell said. The harm is worsened by a lack of transparency. While in some cases Wood has been notified that her pictures were banned or limited in reach, she believes Instagram took other actions against her account without her knowing it. "I've had people say 'I can't tag you,' or 'I was searching for you to show my friend the other day and you're not showing up,'" she said. "I feel invisible."

Because she might be, said computer scientist Derczynski: "The people posting these images will never find out about it, which is just so deeply problematic." he said. "They get a disadvantage forced upon them and they have no agency in this happening and they're not informed that it's happening either."

Source:
https://pulitzercenter.org/stories/there-no-standard-investigation-finds-ai-algorithms-objectify-womens-bodies

Part III: Policy Protections.
Imagine that you have the power to write a policy to stop AI bias. Write one policy rule based specifically on the information presented to you today on biases and objectifying women's bodies. Include the following information:

Who has to follow this policy?

What will the policy do (or prevent)? How will you ensure this is done?

Write your policy here.

Junior Honors Cohort
Darnell, Healy, Houghteling, Jihan

Introducing AI Bias
Reading 3

name: Ms. Wright
period:

Part I: AI Biases Presentation.

Directions: Complete the steps below as you discuss AI biases via the presentation.

Define the following terms from the discussion and slides:

1. Training Data - Information given to AI to teach it a certain task or input.
2. Testing Data - The information used to check whatever the AI that was created is reliable and accurate - output
3. AI Bias - When one ai tool makes a decision that is wrong or problematic because it learned from training that didn't treat all people, places and things accurately.

Turn-and-Talk 1: Take 60 seconds to brainstorm with your elbow partner - how do you think we could reduce the bias in the AI training of detecting fruit types?
We can give AI input information so specific that room for bias is below 10%.

Turn-and-Talk 2: Take 2 minutes to imagine that you have now been asked to create an AI tool that can identify weather conditions, such as sunny, rainy, or cloudy. **Describe the kind of images** you would include as a part of the **training data.** Your goal is to come up with as complete of a data set as possible so that you limit the chance for AI bias! Cloudy - sometimes that can be grey clouds depending on how the weather is. Dark clouds, the temperature would be kind of cold.

Drawing Conclusions: In your own words, and in 2-3 sentences, describe how AI bias occurs.
the lack of specific information in the training data when training data specify an item like an apple, it can be mistaken for objects with similar appearance.

Part II: Bias in AI Partner Reading

Directions: Annotate as you read the following article with your partner. Specifically, take notes in the space below about what the authors did and what actions they took to capture data about AI. Then, summarize the conclusions of the data with your partner.

AI is biased, AI has been mislabeling things as inappropriate and spreading misinformation. For example in the paragraph the AI didn't include all the information only the negative. It didn't mention the remarkable accomplishments of Reconstruction for Black Americans, the first black president, or our lives since the 1965 Civil Right Movement

What Will It Take to Fix AI's Bias Problem? | Opinion
by Jean Darnell

In the 11 months since it launched in November 2022, ChatGPT has changed how educators grade, cite, and interact with technology. When one is wearing the proper rose-colored glasses, advances in AI will make menial tasks irrelevant—which translates to more time to do the things we want.

For instance, Twee can take any YouTube video and create an exam, discussion questions, fill-in-the-blank questionnaires, vocabulary, etc., in seconds. In less than five minutes, I turned a video about bats (*SLJ* School Librarians' Back-to-School Hacks for 2023-24) into a lesson complete with discussion questions, a listening comprehension quiz, and a fill-in-the-blank handout (I wrote the lesson; Twee compiled the last three links).

It's exciting that hours of brainstorming, notetaking, and reading comprehension can dissolve with a few clicks of the mouse.

But here's the elephant in the room: Because most AI is based on large-language models (LLMs) seeking information from all corners of the internet, the content *output* is only as good as the content *input*. So if someone purposely puts something on the internet that's rooted in a basket full of lying posies, it raises key questions:

Can AI discern fact from fiction; and misinformation from disinformation and mal-information (information close enough to the truth to be believable at face value, but intentionally harms a group or individual)?

And: Can AI differentiate between hate-group propaganda and lived, experiential learning from a diverse perspective?

Right now, the answer to both is no.

AI has a bias problem—here's just one example. In August 2023, an Asian MIT student asked AI to make her headshot photo more professional. AI turned her eyes blue and lightened her skin.

Why? Remember, the *input* data has to be free of biases for the *output* data to exceed expectations—or even meet them, in this case. Initiatives like pocstock, is a stock media company focusing on people of color, may help when it comes to the information AI uses as a source for the results produced via a prompt.

AI uses algorithms, LLMs, and humans to stop the spread of misinformation, *after it's been detected*. But if we input reliable, diverse, and inclusive data, we can teach AI how to detect inaccuracies.

(Fun fact to consider in the meantime: A 2018 Twitter survey showed that false news stories were more commonly retweeted by humans than bots and 70 percent more likely to be retweeted than true

tories.)

Representation matters. And the more trained AI is in finding diverse sources, the more AI will be able to see us as we see ourselves.

A research wild card

Librarians understand investigative research. With very young students, we cover the difference between fiction and nonfiction. We teach them to analyze online website addresses (.gov = official, .com = business, .edu = education). We coach students on digging for deeper biases when reading articles for education or entertainment. We advise them on organizing their thoughts via brainstorming, differentiating narrative from expository essays, and including applicable research in their academic work.

With AI, tasks of investigative research with hallmark tangibles (like URL endings) have fallen victim to deep-fake, maleficent missteps that can erase cultures.

Earlier this year, I asked ChatGPT to write a paper on Black History. I had two goals: to discern what it was programmed to learn about a culture I'm fully immersed in experiencing, and to see if it gave a fairly balanced, accurate representation.

The essay mentioned enslavement, civil rights, four notable Black Americans, and provided a conclusion. It didn't mention the remarkable accomplishments of Reconstruction for Black Americans, the first Black president, or our lives since the 1965 Civil Rights Movement.

That paper could make you think nothing historically significant has occurred regarding Black Americans for nearly 60 years. Based on that, I'd say the inclusion of some cultural groups and historically disenfranchised communities was not part of AI's development. It's enough to make you think Black folks are on the wrong side of the technology. I turned that experience into a teachable moment.

ChatGPT doesn't provide citations automatically unlike its competitor, Google's Bard. I was curious. So this month, I asked Bard the following prompt: "Create a speech from a Black woman's perspective on how book bans suppress the speech and freedoms of Black people. include specific examples from the last 10 years." Here's the essay in its entirety. There's also an audio version on YouTube. Spoiler alert: I was pleasantly surprised.

Still, there's a ways to go. Others have also shown how AI is discriminating and prejudiced: See "Who Is Making Sure the AI Machines Aren't Racist?" and "Is AI in favor of Racists?" When 42 percent of the U.S. population identifies as Black, Hispanic, Asian, Indigenous, and two or more racial ethnicities and are *not* included in AI regulation, then "Houston, we have a problem."

In the future, will AI be more inclusive? Will it solve or address racism since we just can't seem to shirk that lesion off our backs? Also, how will it impact book bans?

We don't know yet because the technology is evolving at an exponential pace. But let's look at the evidence.

How has AI affected libraries and our mission for intellectual freedom and informational integrity?

Well, an Iowa District Used AI to Figure Out Which Books to Ban. Here's the kicker: Administrators didn't have *time to read books* before the new school year. So they relied on AI... *not* the true experts, school librarians. AI sourced this list with information from discriminating and targeted proposed house bills, opinionated personal websites by "concerned parents." as well as any and everything on the internet, truthful or not.

Will AI be more inclusive and less racist in the future?

It depends on whether efforts will be able to keep up. Last month, Mark Zuckerberg, Elon Musk, and Bill Gates met with senators and others in Washington, D.C. for an AI Insight Forum to discuss regulations for AI. Quick translation: Three of the richest white men in the world fielded questions and concerns in a private, closed session to the public. (And which one of those has a history of violating the intellectual freedoms of citizens previously?) Time will tell if those concerns will be addressed.

I applaud Senator Chuck Schumer, who is pushing for federal legislation regarding AI. "Government must play a role in requiring these safeguards. Because even if individual companies promote safeguards, there will always be rogue actors, unscrupulous companies, and foreign adversaries that seek to harm us," he said in the forum's opening remarks. "And on the transformational side, other governments, including adversaries like China, are investing huge resources to get ahead. We could fall behind, to the detriment of our national security."

I promise to do my own part to hold AI accountable. It's part of my investigative skills as a librarian. And it's my goal to do my part to ensure the world I hand over to my students and my kids is as equitable and safe as possible.

So let's end on a good note. AI is a fun tool. It's a shortcut in learning that echoes the same instant thrill, at the tip of your fingers, sensation that the iPad did within the first decade of the 21st century. Here's how I used ChatGPT to teach literacy as a school librarian: 7 Innovative Prompts to Use with AI for Literacy. My favorite example from the list is #5, *Feedback from Beyond*, where Edgar Allan Poe advocates for kids to read scary stories.

Source: https://www.slj.com/story/What-Will-It-Take-to-Fix-AIs-Bias-Problem-Opinion

Part III: Policy Protections.

Imagine that you have the power to write a policy to stop AI bias. Write one policy rule based specifically on the information presented to you today on biases and objectifying women's bodies. Include the following information:

Who has to follow this policy?

What will the policy do (or prevent)? How will you ensure this is done?

Write your policy here.