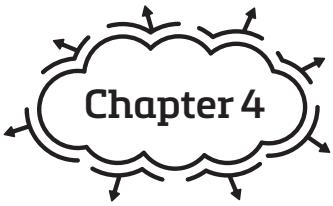

Human Bias: Can Artificial Intelligence Help Diminish Human Bias in Decision-making?

When I taught eighth-grade language arts, my students and I always read the Kurt Vonnegut short story "Harrison Bergeron," which is about an egalitarian society. In an egalitarian society, all people are considered equal and deserve equal rights and opportunities. In and of itself, the notion of an egalitarian society is ideal, but Vonnegut forces readers to explore a rather unconventional way of reaching the goal of equality. In the story, the government uses "innovation" in the form of physical handicaps to diminish each person's strengths and put everyone on equal footing. These handicaps include masks to hide beauty, large weights to slow down the most athletic, and distracting noises to interrupt the thoughts of the intelligent.

When I read this short story with my students, we spent a lot of time dissecting the typical elements of storytelling. We spoke briefly about whether people in this society were truly equal, and if equality made them happy. If I were teaching this story today, though, I think my focus would be a little different. I'd want my students to start considering our own society and whether the technology we have at our fingertips today could make a more just and equitable world for all of us to live in.

To help students have a conversation about the pros and cons of using artificial intelligence to create a more equal society, you will need to build their background knowledge. This chapter will give you a foundation to tackle some tough questions:

- What is the relationship between artificial intelligence and human bias?
- In what ways should race, gender, and representation be taken into consideration when developing a new tech product or service?

Building Background Knowledge

You've certainly seen Hollywood's portrayal of artificial intelligence in lovable characters such as Rosie the Robot Maid from the popular 1960s cartoon *The Jetsons* and the more recent Baymax, an inflatable healthcare robot that stars in the 2014 Disney movie *Big Hero 6*. Many of us have dreamt about the possibilities of robots who can serve our homes and ease our burdens while another subset of the population worries about those same automatons taking over the world.

According to *Merriam-Webster's Dictionary*, artificial intelligence (AI) is a branch of computer science dealing with the simulation of intelligent behavior in computers. This vague definition coupled with Hollywood's fascination for science fiction storytelling has left a large portion of the public ill-informed not only about what AI is, but what both its capabilities and limitations are.

In the most general sense, artificial intelligence relies on large data sets and algorithms to help analyze a scenario and take action that would help maximize its chance of success. We can think about a game of tic-tac-toe against a computer as a tangible example of AI in action. Using a large data set of previously played games, possible moves, and successful outcomes as well as simple algorithmic commands like "take the center square if it is free," a computer can make "intelligent" choices that would allow it to win the game against a human. Each time a new game is played, the moves

Vocabulary to Know

Algorithm – a set of human-developed, step-by-step instructions that computers follow to complete a task.

Algorithmic Bias – systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others.

Artificial Intelligence (AI) – the ability of a computer to modify existing or create new algorithms based on new data and inputs; AI uses human reasoning as a model but not necessarily the end goal.

Human Bias – a tendency, inclination, or prejudice for or against something or someone.

and outcomes are added to the data set, essentially making the AI even "smarter" as it has more information to rely on.

If we look to scholars for a more detailed definition of artificial intelligence, we can get a clearer picture of the types of feats AI is ready to tackle, where it is already in use in our everyday lives, and where AI still has its limitations.

Current uses of artificial intelligence

Today, artificial intelligence is used across various industries in ways that are both highly visible and nearly invisible to consumers. Innovations like text-to-speech, automated chatbots and online assistants, automatic email filtering, Google predictive searches, Netflix recommended content libraries, and GPS map estimated travel and arrival times are all examples of AI that you've probably encountered in your everyday interactions with technology.

Less visible uses of artificial intelligence include fraud protection services provided by your bank or credit card company and the ability for your

smartphone's camera to focus in on the people in the picture you're taking. And if you've applied online for a job recently, odds are there was some level of artificial intelligence scanning your resume before it ever made its way into human hands.

Limitations of artificial intelligence

Although AI is certainly becoming more prevalent in society, it still has its limitations. One major limitation is that AI is only as good as the data set it relies on for making decisions (Chowdhury & Sadek, 2012). Netflix might do an okay job recommending content to me based on previous Netflix viewing, but it might do an even better job recommending content if it also had data from what I stream from Hulu, Prime Video, and Disney+ as well.

A second limitation is that AI is typically limited to solving a singular type of problem (Lu et al., 2017). So while an algorithm might be able to determine the quickest route for you to get from point A to point B in your vehicle, it cannot take into account outside factors like the amount of gas you have in the tank, whether or not you have a fear of driving over bridges, or even if you prefer to pass your favorite coffee shop as part of your commute. Machine learning and artificial intelligence are excellent at observing and extracting patterns, but they cannot do the difficult work the human brain does as it takes so many varied inputs into account before deciding.

Artificial Intelligence in Decision-making

All people hold both explicit and implicit biases that subconsciously influence the ways they move through the world and interact with others (DiAngelo, 2018). These unconscious biases can unintentionally cause us to discriminate based on age, gender, race, or sexual orientation. The biases we carry with us can be a barrier to equal opportunities for all. Artificial intelligence is often touted as a solution for the bias that can creep into human decision-making much in the same way "innovations" are used in "Harrison Bergeron" to make everyone more equal.

Judgments regarding who should receive business loans, who should be hired for a job, which students should be accepted into a university, or even

which inmate is ready to be paroled can all be influenced by various types of explicit or implicit human biases. The promise of AI is that it can process data and make decisions based on previously charted successes without allowing factors such as age, race, or gender to come into play. These promises sound amazing. Who wouldn't want to be evaluated based on their merits and abilities rather than some obscure demographic details?

On the flipside is an argument that AI itself is biased because the human beings that create it may be inadvertently weaving bias into the programming. So the question is: can tech companies offer us the level playing field from which we'd all prefer to be judged? Or is it possible that we have become a society plagued by *technochauvinism*—a term that *Artificial Unintelligence* author Meredith Broussard (2018) coined to explain our collective belief that technology is always the solution?

Let's examine two examples of artificial intelligence aiding in human decision-making.

AI in hiring

An elementary school principal needs to hire three new teachers for the upcoming school year. She sits at her computer, logs in to the district's online application system, and scrolls through a list of potential hires. Much to her relief, the first ten applicants seem amazing. They are qualified, have intelligent answers to essay questions, and experience in other schools or districts. The principal begins a list of candidates to interview but pauses for a moment when she realizes that three of the five names on her paper belong to men.

A few thoughts cross the principal's mind: We don't have any male teachers in our building right now. How might a man fit in with a school full of female colleagues? How will parents and students feel about having a male teacher work here? Don't most men in education teach high school? The principal leaves the three men on the list, adds a few more female names, and continues scrolling through applicants on the second results page in case she finds even better candidates there.

Prior to online job applications, principals had to rely on paper applications, letters of recommendation, portfolios, or word-of-mouth recommendations from colleagues to identify good hires. Applicants whose portfolios resided at the top of the stack had an automatic advantage, as the process of simply looking through the giant paper pile could be daunting and time consuming. In the age of the online application, however, powerful tools allow HR departments to quickly identify candidates who can meet their organization's needs and automatically weed out the profiles who just don't fit the bill. This innovation saves time for people doing the hiring. But saving time is not the only motivation for the use of AI in hiring.

Although the principal in the opening scenario was a bit surprised to see so many qualified male candidates in her applicant pool, this was likely no coincidence at all. If the district has a goal to increase the number of male teachers at the elementary level, it would only take a few clicks of the mouse to prioritize male candidates in the online application system, putting them on the first page or two of the principal's dashboard.

Even if the principal is not opposed to hiring male teachers, her previous hiring patterns and hesitancy to include too many male names in her to-be-interviewed list indicate that she prefers to hire females, even if she does so subconsciously. In this case, AI was able to put male candidates on her list of interviewees that may not have made it there previously.

AI in law enforcement

A police officer in a large city arrives at the station to pick up his patrol assignment for the evening. The officer is told that between the hours of 10 p.m. and 2 a.m. he is to focus his patrol efforts within a five-mile radius of a nightclub in town that is popular with Latinx residents. Data from the precinct's PredPol software indicates an increased likelihood of crime occurring in that area on weekends between 10 p.m. and 2 a.m.

Early in his shift, the police officer parks his squad car for a while and monitors traffic activity near the outer perimeter of his assigned radius. After some time, the officer drives around a bit within the five-mile radius, keeping an eye out for any suspicious activity. As the officer gets closer to the

club, he sees two men outside in an apparent scuffle. The officer stops his car, intervenes, and arrests the men for disorderly conduct. When he returns to the station with the men, a female officer is sent out to resume patrol of the area. Toward the end of her shift, the female officer arrests a man leaving the nightclub for driving under the influence. All three arrests from the evening are processed and data on the perpetrators is entered into the PredPol database.

PredPol, Inc., the company used by the police department in the scenario above, produces real software that uses predictive analytics to support law enforcement. It was developed in 2010 by scientists at UCLA in conjunction with the Los Angeles Police Department. The goal of the project was to see if scientific analysis of crime data could help locate patterns of crime and criminal behavior (PredPol, 2018).

Today, PredPol's algorithms are in use by more than sixty police departments around the United States. The company claims that their algorithm is twice as accurate as analysis done by data scientists when it comes to predicting areas in a neighborhood where serious crimes are likely to take place during a particular period. No independent studies, though, have confirmed those claims (Rieland, 2018).

Although PredPol has come under scrutiny for their algorithms having racial and socioeconomic biases baked into them, the company claims this simply cannot be the case. Their database does not capture demographic data. It relies on only three points of data related to the crime: the crime type, the location it occurred, and the time that it occurred (PredPol, 2018).

In this story, however, it is easy to see how predictive software like PredPol may inadvertently be targeting Latinx residents. The algorithm indicates police should be present, so they are. Police make three arrests that evening—not because of calls, complaints, or accidents, but because they were in the area watching for crime. At the end of the shift, the PredPol database now contains three more data points that indicate crime is likely to happen in the area between 10 p.m. and 2 a.m. on the weekends. So, guess what happens the following weekend? That's right. Police are placed in the area

once more. And because they are present, they are likely to find a few more data points to feed into the system and further perpetuate the idea that the area around the Latinx nightclub is a hotspot for criminal activity.

Breaking Down the Arguments

In the two examples above, artificial intelligence is employed to aid professionals in their work. In the hiring example, male candidates for an elementary teaching job are elevated to bring more gender diversity into the school district. In the second example, Latinx men and women end up being targets of a policing algorithm that is meant to help officers more effectively reduce crime.

If we ask students to begin wrestling with the essential question of this chapter, "can artificial intelligence help diminish human bias in decision-making?" they are likely to make one of a few major claims with some popular, standard arguments to support their claim.

In the sections that follow, I will outline the competing claims, highlight some original research from experts in the field, and provide examples of each perspective in the headlines.

Claim #1: Artificial intelligence can remove bias from human decision-making.

Individuals who take this stance are likely to support their viewpoint with evidence like the hiring example above. Additional arguments to support this viewpoint include:

- Artificial intelligence can help people avoid common biases such as the similarity attraction effect, where humans tend to gravitate toward people like themselves, or confirmation bias, where humans favor information that confirms their beliefs.
- Predictive algorithms can ignore demographic data, such as gender and race, to make human decision-making more equitable and break cycles of oppression.

What the research says

In a study by Bo Cowgill, assistant professor in the Columbia Business School (2020), two groups of employees at a software development firm were tracked. One group was comprised of employees who were selected for an interview by a human; the second group of employees was selected to interview by a machine—even if their resume had previously been passed on by humans. Cowgill's findings are quite interesting.

First, a greater diversity of candidates was put forward by the machine than the human. This included women, racial minorities, candidates from "non-elite" colleges, and candidates without industry referrals. Once the candidates were hired, Cowgill found employees who were machine picked routinely scored higher than those screened by humans in measures of productivity that were already in use by the company. Finally, Cowgill found that the machine was also better at choosing candidates with superior soft skills like cultural fit and leadership.

Similarly, Kimberly A. Houser, assistant professor at Oklahoma State University, has written extensively about the successes of AI in reducing workplace bias (2020). Houser cautions that before we dismiss AI as flawed or biased, it is important that we have a clear understanding of how messy human decision-making is. She notes that "noise" in a person's day, such as when they've last had a meal or what the weather is like can cause humans to make completely different decisions than they might have hours or days before.

Coupling "noise" with unconscious biases and cognitive blind spots means that humans are not very good at making fair and impartial judgments. Houser cites multiple examples in which the use of technology to remove names and gender identifications has resulted in more women being hired, especially in the tech industry. Although Houser acknowledges that there are still improvements to be made in AI for decision-making, she asserts that the technology has come a long way, and that for many important decisions, including workplace hiring, the machines are already more reliable than humans at impartially selecting quality candidates.

The workplace is not the only one where artificial intelligence can help defeat bias. Another area where men dominate women is the start-up sector. Venture capitalists, who make decisions about which start-ups to invest money in, overwhelmingly support male-owned start-ups over female-owned ones. When asked about how they choose products to invest in, many of them openly admit to choosing entrepreneurs they think they can get along with and often rest on the laurels of a "gut feeling" that helps them decide who to give money to (Hernandez et al., 2019).

This "gut feeling" method has resulted in only 2.2% of venture capitalist funding going to women. Recognizing the need to locate and support more diverse entrepreneurs, major VC companies in the United States and Stockholm have developed and are now utilizing AI tools to inform investment decisions. The use of AI in venture capitalism is still too new to track major results, but venture capitalists themselves say that they are happy to have the data in front of them as they attempt to make more well-informed decisions about which entrepreneurs to invest in (Hernandez et al., 2019).

One thing these researchers have in common is that they find the greatest successes when AI is coupled with human intelligence. AI still has its limitations—it cannot detect things like body language, facial expressions, and general "human likeability" factors such as personality, all of which can be helpful in determining a good fit for a company. What AI can do, though, is aid in the earlier parts of a process and present decision makers with the most impartial set of candidates to choose from.

In the news

Want to be better at sports? Listen to the machines (Smith, 2020)

Artificial intelligence is making its way into the sports world, with everything from technologies designed to improve player performance to others that can detect and predict injuries. Some companies claim they can use technology to analyze a player's unique strengths and then match them with a team in need of their skill set. Any performance technology is only as good as the data set it must learn from, though, so many companies

interested in bringing AI into sports are in a race to collect, label, and code data as quickly as possible.

Facebook's AI for Hate Speech Improves. How Much is Unclear (Simonite, 2020)

In its most recent quarter, Facebook algorithms removed 9.6 million pieces of content deemed as hate speech. This was up from their previous quarter total of 5.7 million pieces of content removed. Facebook's chief technology officer attributes this increase to improvements in their artificial intelligence efforts. As Facebook collects more examples of hate speech, their AI becomes more accurate and is better able to identify more nuanced forms of hate speech. In this last quarter, Facebook's technology was able to identify 88.8% of the hate speech it collected before it was reported by human users.

Claim #2: Artificial intelligence cannot remove human bias from decision-making because it is created by humans.

Individuals who take this stance are likely to cite examples like the one of PredPol in policing to point out the flaws of using artificial intelligence in human interactions. Additional arguments to support this viewpoint include:

- Artificial intelligence and algorithm developers have largely been men. In the West, they have predominantly been white men. Without a diverse group of people creating AI, these algorithms may reinforce the stereotypes of their creators.
- AI can push its own learned biases forward. If a growing dataset says that men named Joseph get hired more than men named José, the AI may prioritize resumes based on something as irrelevant as a first name.

What the research says

In her 2018 book, *Algorithms of Oppression*, Safiya Umoja Noble explores how commercial search engines, largely created and maintained by a strikingly heterogeneous workforce, are reinforcing bias and racism. In 2011, Noble was disturbed that a Google search for "black girls" returned pornographic websites as the first ten results on the page. This is no longer the case, but Noble's research over the years continued to uncover some disturbing trends while using the popular search engine. In 2014, a Google Images search for the word "beautiful" turned up hundreds of pictures of white women even though the word "woman" was not part of the search term. In 2015, a Google Images search for "professor style" only returned images of white men. In 2016, a Google Images search for "three white teens" turned up tons of wholesome stock photos while the search for "three black teens" returned mugshot photos.

Although Noble (2018) acknowledges that there are many reasons for the issues she uncovered in search engines reinforcing stereotypes, including the commercialization of information, researchers from the AI Now Institute at New York University posit that bias within AI systems is a direct result of the lack of diversity in both the AI workforce and in academia (West et al., 2019). Their 2019 white paper "Discriminating Systems: Gender, Race, and Power in AI," indicates that only 15% of Facebook's AI team is female. At Google, only 10% of their AI team is women. In academia, 80% of AI professors are male and only 18% of presenters at leading AI conferences are female. The disparity in racial diversity is even more extreme. At Facebook, only 4% of their employees are Black and only 5% are Hispanic. The numbers at Google are even lower: 2.5% Black and 3.6% Hispanic. In academia, non-white professors make up only 19% of postsecondary faculty nationally, and minority faculty members are even more underrepresented in areas of science, technology, engineering, and mathematics (Davis & Fry, 2019).

In her 2017 book, *Technically Wrong*, Sara Wachter-Boettcher explores the ways that a homogeneous workplace becomes evident in technology through something as innocuous as the settings and options in many of the apps, websites, and tools we use. For example, the default settings on most

virtual assistants like Siri and Alexa are female voices—reinforcing the stereotype that women are more helpful than men. Snapchat filters meant to "beautify" typically slim the face, contour the cheekbones, and lighten the skin—reinforcing stereotypical beauty standards perpetuated in the media. In 2015, of the top fifty character-based games in the iTunes store, male characters were the default 85% of the time. What's worse, fewer than half of the games even offered a female option—reinforcing the idea that gaming is typically enjoyed by male audiences.

What these researchers have in common is their skepticism that most technologies, artificial intelligence included, can be developed in unbiased ways when the core group of people creating them go unchallenged. Whether programmers are bringing in their implicit biases or baking in rather explicit ones, it is difficult to establish a system of checks and balances within organizations that primarily employ people from the same walk of life.

In the news

Amazon scraps secret AI recruiting tool that showed bias against women (Dastin, 2018)

Since 2014, the team at Amazon had piloted a computer program intended to quickly scan resumes and flag top talent based on factors that had led to successful hires in the past. The problem was that most of the company's hiring over the last ten years had been male-dominated, and the resumes of those male employees became the data set through which the AI was trained. Luckily, humans picked up on the flaws within the system when they noticed highly qualified candidates were being graded lower on resumes that included names of historically women's colleges or even the word "women" (as in "women's chess club president").

Insurers want to track how many steps you took today (Jeong, 2019)

In 2018, insurance company John Hancock offered its customers the option of wearing a fitness tracker. If customers showed evidence of living healthy lifestyles, they became eligible for discounts. Insurance companies have long used data to determine the risks of taking on a client and adjusting

Additional Questions for Students to Explore

- Can computers help humans make better or more fair decisions?
- Is it possible for humans to create artificial intelligence free from bias?
- Should artificial intelligence projects be regulated by some form of governance?
- What makes humans "smart"? Can those characteristics be replicated by machines?
- How might a tech company's workplace culture contribute to a lack of diversity in the field?
- How do scientists define bias as opposed to those in the humanities? How might these different definitions be considered in the creation and use of AI?
- What steps can a technology company take to prevent bias in their products?
- Should computers help decide who gets admitted to college? Released from prison? Secures a loan? Pays more for insurance?

their prices accordingly; it's the reason younger drivers have higher car insurance premiums than seasoned ones. With artificial intelligence and larger data sets, insurance companies have even more predictive powers. Using data collection tools such as fitness trackers, connected cars, smart appliances, and even personal home assistants like Alexa, many more insurance companies are getting into the business of "personalized rates." Skeptics worry that more data could lead to greater discrimination based on age, race, geographic location, or even genetic makeup.

Curricular Connections

Michelle Ciccone (@MMFCiccone) is a technology integration specialist in Massachusetts who has intentionally found ways to engage both middle and high school students in conversations around digital ethics. Michelle has found that middle school students are not only capable of learning about how the internet is structured but are excited to do so. She says that "by de-personalizing the conversation and focusing on infrastructure and the way things are built, I am able to communicate that the Internet is not this natural, ephemeral 'cloud,' but in fact is built by other humans and plays out the biases of those humans." Once students understand the basics, Michelle can ask them to "reimagine a different way of doing business."

At the high school level, Michelle collaborates with content area teachers to bring tech ethics conversations into the classroom. During a series of lessons with eleventh and twelfth grade engineering students, Michelle asked how one of them would feel if a product they created was harmful or discriminatory in some way. There was a little bit of debate around the room, but the general consensus of the class was that "if the engineer/inventor didn't mean for the impact of a technology to be discriminatory, then there's nothing that the engineer/inventor is obligated to do once the impact is known."

The students felt they could remove themselves from the ethical question given that they were only a small part of a final product. The overwhelming opinion of the class was that "an engineer's job is to build the product their client is asking them to build, and if it turns out that their client is asking them to build a product that has a problematic impact, it's not really the engineer's place to raise concerns."

To help the engineering students in her school consider another perspective, Michelle invited Ruha Benjamin, Princeton professor and author of *Race After Technology*, to a virtual meeting with the class. Benjamin spoke with students about her studies of encoded biases and was able to field their questions about the role of engineers in ethical design. Over the course of the lessons, Michelle understood that most engineering students struggled with the concept of intent versus impact in their designs and would need to

revisit this topic often as they worked through various design projects over the course of the school year.

Try this

A team at University of Colorado Boulder (CU), led by Tom Yeh and backed by a grant from the National Science Foundation, has been working on a series of lessons for middle and high school students designed to target ethical concerns over AI through storytelling, simulation, debate, and chatbot design. Matthew Turner, a member of Yeh's team, told me about two activities you could use with your own students.

In "The Undone Activity," students are introduced to a dystopian future where "The Society" is running out of resources and must figure out a way to cull the weak from its population. An AI system is used to determine who is "undone" based on character traits such as health, athleticism, hobbies, careers, and goals. Students are tasked with designing a variety of fictional characters to live within The Society. Each of their characters has a mini-bio outlining positive and negative traits and values of the character. The class is then tasked with determining which of the fictional characters should be undone. They talk about how they made their decisions and how they would quantify those decisions into rules that AI could follow. The project is used as a catalyst to help students have conversations around self-driving cars making life and death decisions.

In the "Your Ethical Code Lesson," students are tasked with creating their "personal ethical code" that could be transferred into an intelligent machine. Students create lists of everything they consider good, neutral, and evil. The teacher prompts students with ideas of what can go on the list—everything from guns to pencils, happiness to anger, technology to nature, and more.

After completing their list, students compare notes with a neighbor. The teacher points out that the lists probably look different and asks students to consider how these differences might pose a problem when humans begin designing AI systems. The teacher then provides the students with more items and asks them to sort the item into a column on their list. When

ISTE Standards Addressed

Student Standard 3d: Knowledge Constructor – Students build knowledge by actively exploring real-world issues and problems, developing ideas and theories and pursuing answers and solutions.

Student Standard 4d: Innovative Designer – Students exhibit a tolerance for ambiguity, perseverance, and the capacity to work with open-ended problems.

Educator Standard 5b: Designer – Design authentic learning activities that align with content area standards and use digital tools and resources to maximize active, deep learning.

Educator Standard 6c: Facilitator – Create learning opportunities that challenge students to use a design process and computational thinking to innovate and solve problems.

students are asked to place a beautiful painting of a mountain on the list, most chose to place it in the good category. When asked to place Hitler on the list, most chose to place him in the evil category. But did the students know that Hitler loved painting mountains? The purpose of this exercise is to show how complex and gray humans truly are, making it incredibly difficult to create an AI that is perfect in its design.

Matthew says that the "difficult debate topics in this curriculum foster incredible student discussion." He and the rest of the team at CU is "determined to bring a humanities approach to computer science in the hopes of fostering a more well-rounded student experience."

More resources

Scan this QR code for additional articles, resources, and lesson ideas around this question: "Can artificial intelligence remove human bias from the decision-making process?"