IBM AI Ethics(mod2)

Module 2

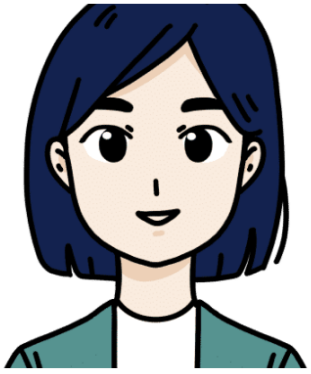
What is fairness?

Meet the team

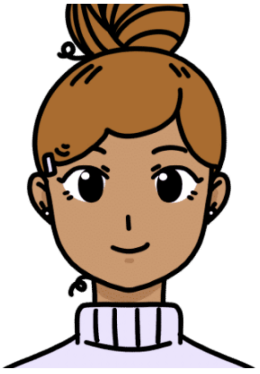
What’s meant by fairness in AI systems? This story looks at a large national banking company and some of the issues with fairness that must be considered and dealt with when deploying AI systems.

As the company gets ready to deploy an AI system to help them identify high-value candidates in their promotion pool, Priscilla notices that most of the candidates belong to one race. They launch a review of the system to understand what is driving the result and find that the system is biased. This story covers basic concepts of bias within the context of fairness, and how bias can enter a system. The story also provides a set of questions for you to think about.

The team



Jordan (they/them)
Chief Data Officer (CDO)



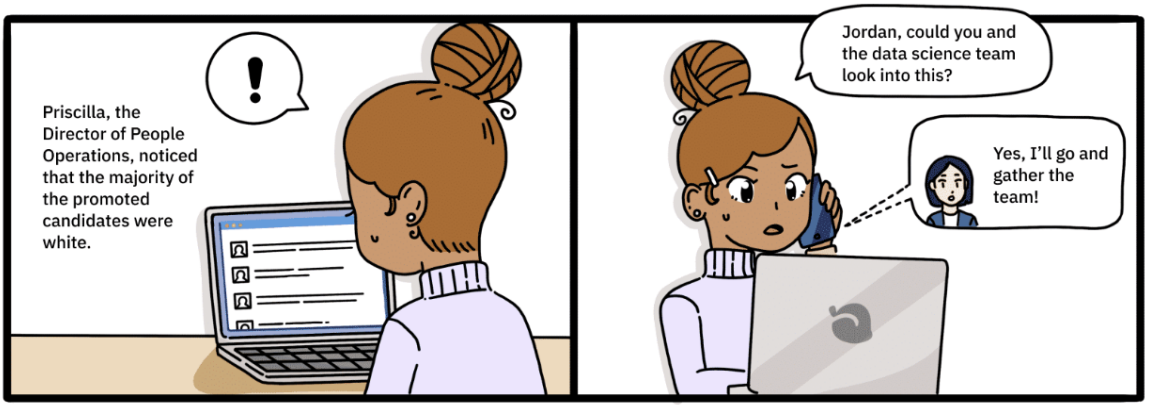
Priscilla (she/her)
Director of People Operations (PeopleOps)

Nischal (he/him)
Director of Data Science

Identify the issue

Jordan has taken a new role as the banking company's Chief Data Officer. Before Jordan joined, the company had been testing a new AI system, the first of its kind, to help their PeopleOps group identify high-value candidates in their promotion pool for the current year.

The system seems to be working, so the team shares the promotion list with Priscilla, the Director of People Operations. First, Priscilla looks at the list alongside the candidates' demographic information. As she goes through the list, something catches her attention. Priscilla notices that a disproportionate number of the promoted candidates are White. She is surprised because she had specifically recommended a non-White employee who was already performing the next-level job and who sounded like the perfect fit for a promotion, but that person is not on the promotion list. She calls Jordan to start an investigation.

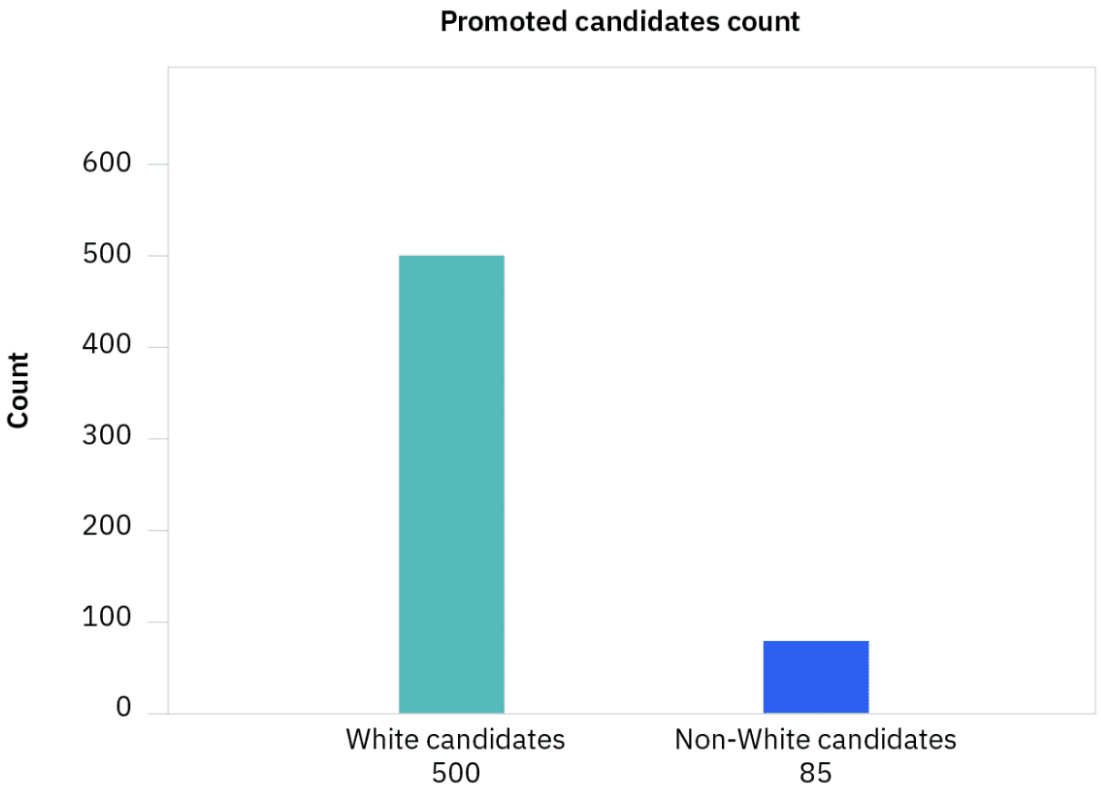


Jordan asks the data science team to look into Priscilla's concerns. As a first step, the data science team quickly analyzes the 5-year promotion data used for training the AI system.

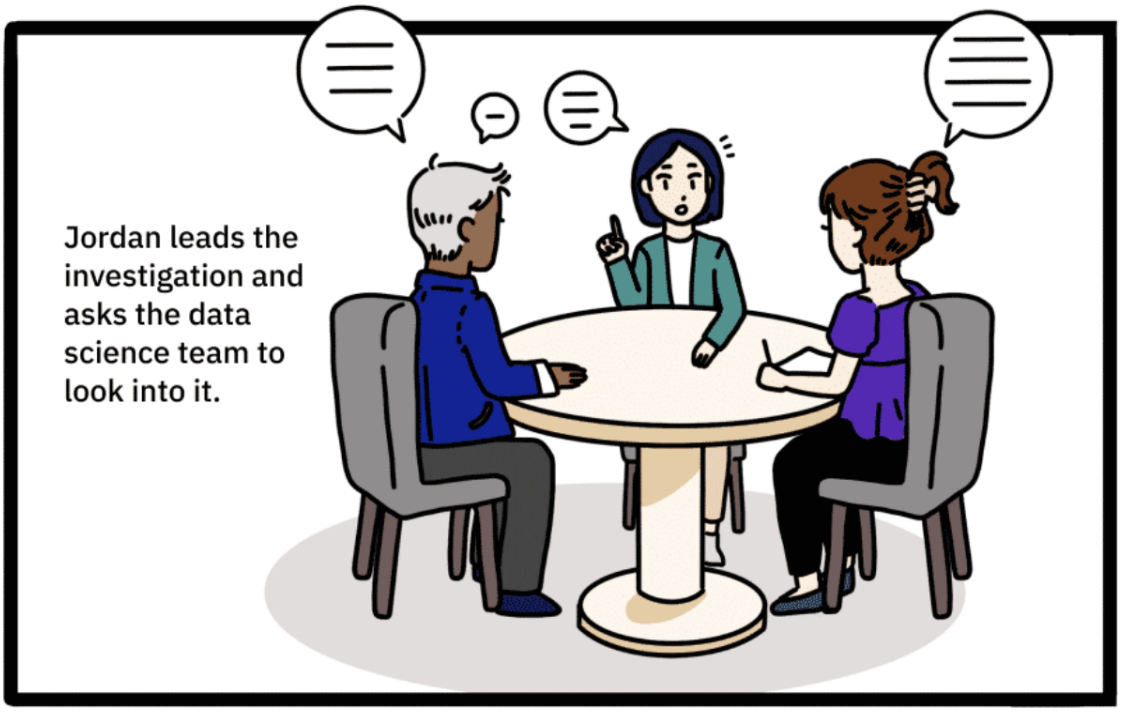
The following table shows a sample of the data they collected:

Employee attributes	**Employee 1**	**Employee 2**	**Employee...**
Employee ID	1	2	...
Employee department	Data & AI	Data & AI	...
Employee qualification	Masters	PhD	...
Employee race	White	Non-White	...
Years of employment	10	12	...
Average rating			...
Promotion decision	Yes	No	...

The team compiles their initial findings in a graph that shows promotion data and candidates' race:



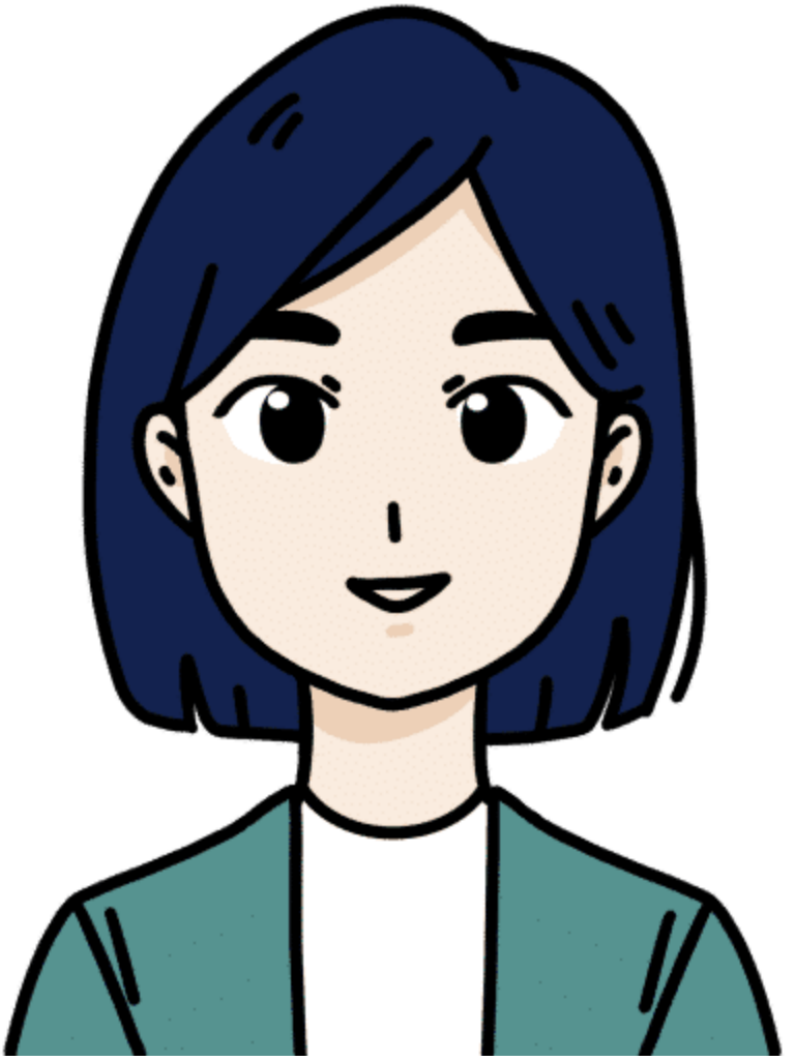
By looking at the graph, Jordan notices that historically there were disproportionately more White candidates than non-White candidates in the final candidate pool for promotion. It occurs to them that the system could have a fairness problem resulting in bias based on race.

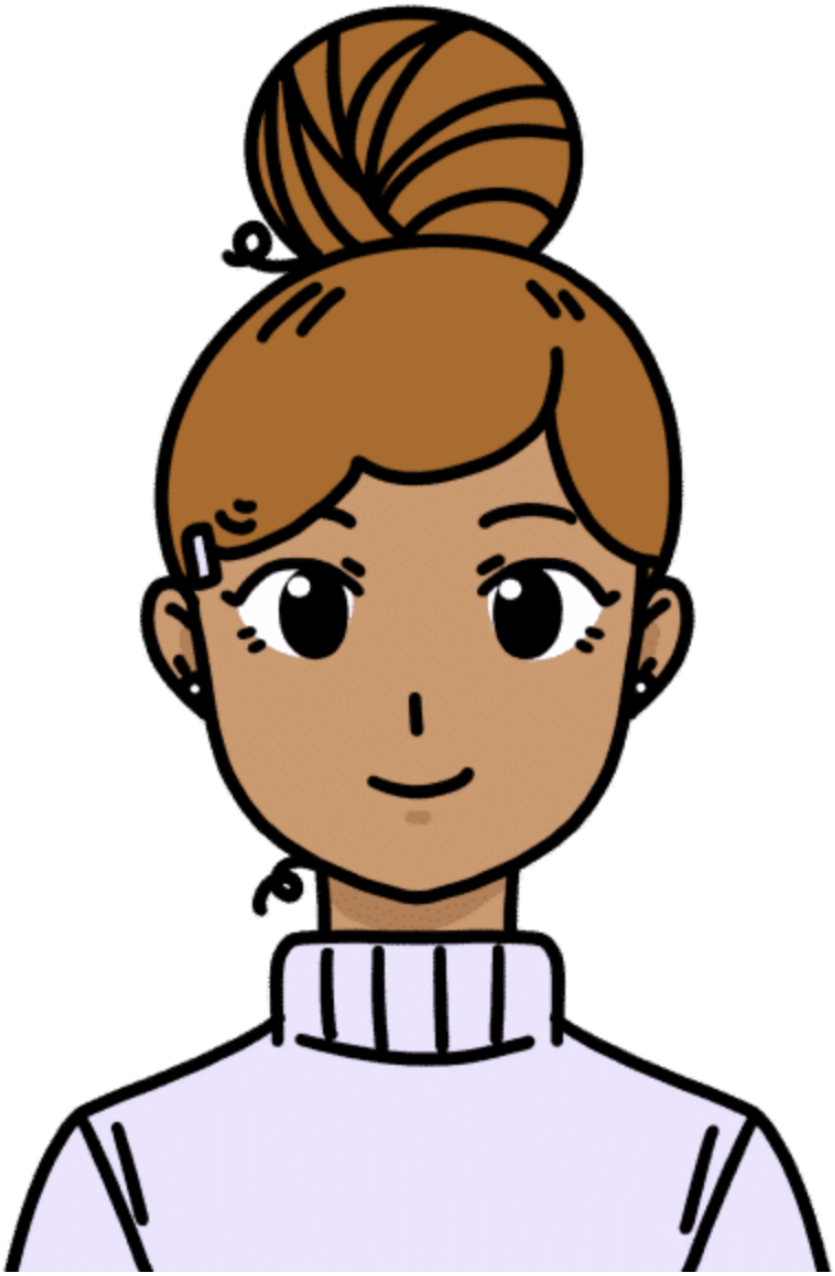


Jordan gets the team together to discuss the first issue they found and to plan what they need to do next.

As a first step to address the dilemma, Jordan asks the team if they think there is business value in keeping the AI system.

Jordan asks, “Do you think we should continue using AI to help with the promotion process?”





After taking a moment to think, Priscilla nods her head.

“If done well and in a way PeopleOps team members can trust, it will save us time and could help us make our promotion process fairer,” says Priscilla.

Jordan, Priscilla, and Nischal all agree that the value of AI is there, if implemented correctly.

Explain the issue

Jordan asks Nischal to work with the Data Science team to prepare some materials to help the PeopleOps team understand what could have caused the problem.

Address the issue

The team has identified the issue as one of **fairness** in the AI model. To achieve fairness, unwanted bias needs to be reduced. In AI, bias is a systematic error that, intentionally or not, can influence an AI system in a way that might generate unfair decisions. Bias can be present both in the AI system and in the data used to train and test it. Based on the data that was given to the AI system, bias had crept in and was affecting the system's results.

Reflection: Fairness in AI

Imagine that you are part of the team trying to deal with fairness issues. Think about the following questions. Take a few minutes to reflect and type your responses in the following text boxes. (Writing an answer is a good way to process your thoughts. These answers are for your use only. You have the option to download your response and save it. It will not be saved in the text box when you move on in the course.)

For question 1: **Is there a step to analyze the intended and unintended consequences of the application using a design-thinking approach? Do you think that such a step is necessary?**

Yes, it is very important to understand known and hidden effects of the applications to mitigate significant harm. One way to try to better understand known and hidden effects is by considering layers of effect. When you consider layers of effect, you think about the application's primary effect (its intended impact), its secondary effects (known or predictable unintended impacts), and its tertiary effects (potential unpredictable or unforeseen unintended impacts). Considering layers of effect with a diverse and inclusive team helps identify a wider range of potential impacts, including potential harm.

For question 2: **Which attribute in this story's data set has the potential to introduce unwanted bias?**The "Employee Race" attributeFor question 3: **Is there a way to mitigate bias at every stage of the AI lifecycle (from development to deployment)? How would you do it?**Yes, there are many ways to mitigate bias throughout the AI lifecycle. Throughout the AI lifecycle, it is critical to work with a diverse and inclusive team whose collective wisdom will help better identify potential

bias issues. An AI models extract key patterns by looking at training data in order to make decisions and predictions. So, selecting high-quality data that is relevant, accurate, complete, and representative is important because using high-quality data will help reduce bias issues later in the lifecycle. Once an AI model goes to production, using tools to continuously detect, measure, and mitigate bias is also very important because it enables you to identify, understand, and remediate issues proactively and on an ongoing basis.

For question 4: **What are the ways observed bias can be dealt with?**There are many ways to mitigate observed bias. The first step is to investigate where and why the model is exhibiting unwanted bias. Then, you can review the data and data labeling and fix any observed issues. You can also check to see if model retraining is needed.