IBM AI Ethics(mod4)

Module 4

What is explainability?

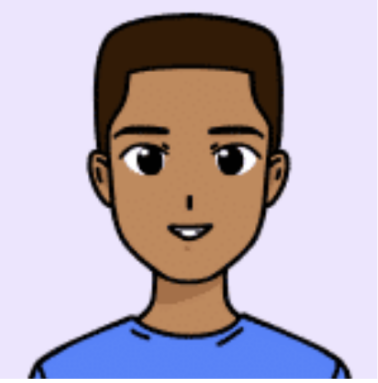
Meet the team

After deploying an AI-based product recommendation system, an online retailer notices that customers want to understand why and how they are getting the product recommendations. The Data Science team dives into how to explain the models in different ways that are relevant for different users. The team also generates a set of questions to think about when solving this kind of problem.

The team

Olivia (she/her)
Head of AI Lead

Clara (they/them)
Data Scientist

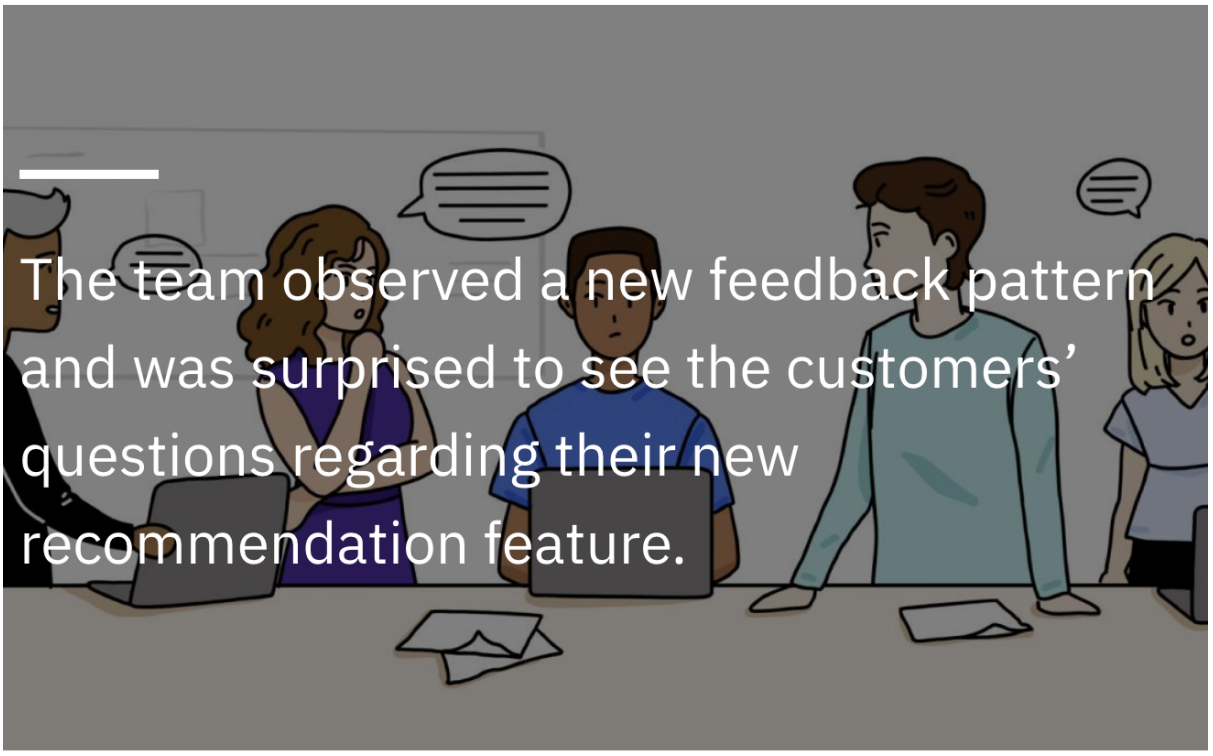


Luan (she/her)
Data Scientist and Model Validator

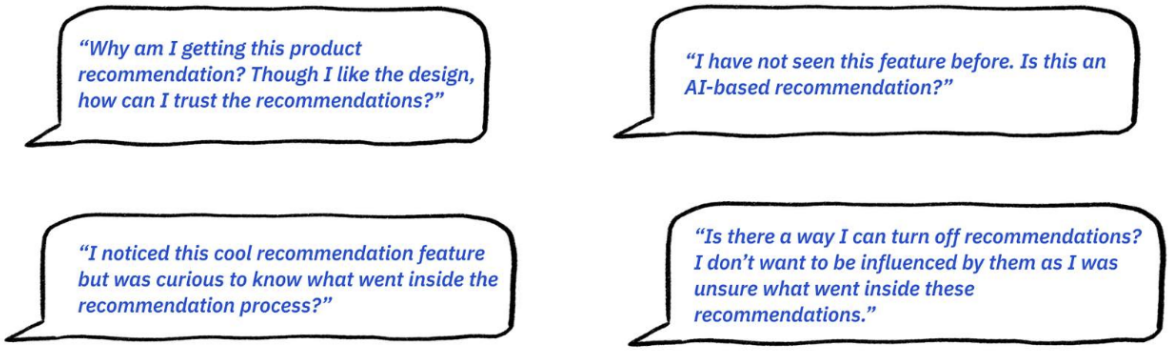
Identify the issue

An online retail company has recently integrated an AI-based product recommendation system to improve the customer buying experience.

After releasing an AI-based product recommendation feature, the Data Science team gathers data and starts the feedback analysis.



Here are some of the customer questions about the recommendations:



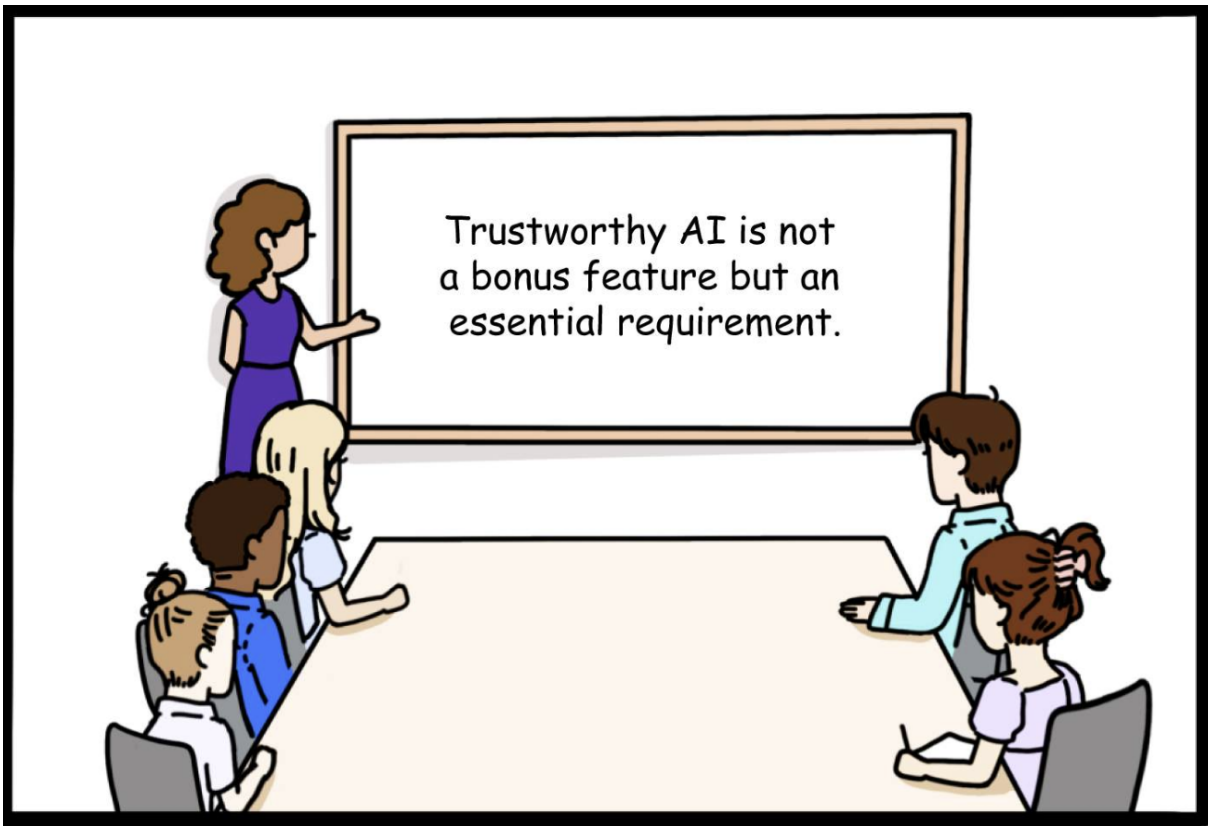
The Data Science team creates a report consolidating the received feedback and emails the report to Olivia.

Coincidentally, Olivia is reading an article about the importance of building trustworthy AI systems so that companies can stay competitive in the market, earn customer trust, and build confidence in trustworthy AI.

Olivia receives the team's email and is pleasantly surprised with the timing of the feedback report. She wonders, "Should the Data Science team spend time adding a new feature to explain where recommendations come from?"

Olivia sends a meeting invitation to the team, calling for a brainstorming session to discuss the issue.

Olivia begins the meeting by thanking her Data Science team for the report. She explains the importance and the need for building trustworthy AI systems. Then she writes on the board at the front of the room: Trustworthy AI is not a bonus feature but an essential requirement.



The team starts to understand that AI's trustworthiness might be more important than they originally thought.

Explain the issue

As Olivia turns back to the team, she hears two voices at the same time.

Clara asks, “So, should we focus on making our model interpretable?”

Luan asks, “Should we focus on the explainability of the model decision?”

Olivia believes these questions are a great way to lead the team in the right direction.

“Great questions!” exclaims Olivia. “But first, let’s understand what each term means. Both **explainability** and **interpretability** are ways to understand how the model works.”

“**Interpretability** is the degree to which an observer can understand the cause of a decision. It is the success rate that humans can predict the result of an AI output, while **explainability** goes a step further and looks at how the AI system arrived at a result.”

“Now let’s look at a scenario that will connect those terms with our system so we can understand them better.”

Scenario: Retail recommendations

Model explanations

Now, let’s see how explanations about recommendations change for each persona Luan listed.

Embracing Discovery

Users might like to understand how the AI system makes specific recommendations to them. They might also like to know what actions they can take to get similar or different recommendations in the future.

> ***P2*** **was recommended as you have previously purchased sugar-free products or products with the vanilla flavor.**
>
> ***P4*** **and** ***P5*** **were recommended based on your search for products with vanilla flavor or products that are not sugar free.**

Approvers and Auditors

Approvers and Auditors might like to understand the overall model decision-making steps based on the features. Both Approvers and Auditors can end up checking this against random users' explanations. The goal may be to evaluate the model against regulatory requirements or internal policies.

```
The product will be recommended based on
the sugar-free indicator and flavor features.

Personally identifiable information is not used
for building the AI recommendation system.

Users with similar purchasing behavior
receive the similar product recommendations.
```

Data Scientists and Model Validators

Data Scientists and Model Validators might like to understand the overall model performance, the effect of features on the performance, and other explanation types to evaluate the model.

Model features:
- **Sugar-free indicator**
- **Flavor**

How will the recommendation system be affected if the feature, Flavor, is removed?

Olivia and Clara thank Luan for the explanation and agree that the mapping will be helpful in drafting the next steps to make the model explainable. Together, the team wraps up the brainstorming session and plans the next meeting to lay out the development and implementation plan.

Address the issue

Thanks to Luan, Olivia and Clara are ready to get to work on increasing the explainability of their AI model. The team now understands the importance of making sure that everyone, not just customers, can understand how and why AI systems generate particular predictions or recommendations.

Reflection: Explainability in AI

Imagine that you are part of the team trying to make an AI system more explainable. Think about the following questions, then take a few minutes and use the space

provided to record your answers. (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.)

Expand for some more thoughts

For question 1: **What is a way for the team to be aware of who is involved in development and deployment of the AI system and each of their roles?**

It is important to define roles before starting to work on the project, after the business discussion stage. Remember that building a diverse and inclusive team — including a diverse community of stakeholders — helps to build systems that are more trustworthy.

For question 2: **How can the team gather the type of explanations each persona will aim to get from the model?**First, the team should define the personas and what kinds of explanations each one needs. Then, the team can think together about how to share those explanations with the different personas without detracting from their user experience.For question 3: **Which explanation method will better match with each persona's expectation?**Explanation methods are different for different consumers. Understanding the needs and goals of the consumer helps in choosing the appropriate method. Design thinking can help the team consider the needs and goals of the consumer.