Video Transcript
Trustworthy AI is crucial for business
(0:00 – 4:07)

So, when we're talking about trustworthy AI, we hear about these five pillars, right? Fairness, robustness, privacy, explainability, and transparency. So, what is all of this?

You're right Aishwarya. There are you know, at this point in time we usually talk about five different pillars, but keep in mind that this is a fast-evolving space. This field is changing rapidly, but at this point we usually talk about fairness, robustness, privacy, explainability, and transparency. Let's maybe talk about each of them quickly.

Fairness is probably obvious it is to make sure that the models are not behaving in a biased way. Now it may actually start, the challenges may start way before a model is built. It might be understanding if the data itself is biased. If it is how do you deal with that? When you build a model how do you make sure that the model is not systematically giving an advantage or a disadvantage to a certain group. And the definition of the group varies by industry, by use case, it could be based on sensitive attributes like age and gender and ethnicity but may not be limited to any of those. You want to make sure that the system is not consistently favoring one over the other in an unfair way.

Robustness, you want to make sure that your models behave well in exceptional conditions. How do you make sure that the model performance is good over time? What is happening with the effective data drift? Or for example, in the context of the of the pandemic, we know that customer behavior has changed you know customer patterns have changed, customer touch points have changed. Is your model still behaving as expected, or if it is not can you at least have an understanding of how the model behavior is changing, how data is drifting, how accuracy is drifting, etc.

Privacy, can you make sure that the model, the data, the model that is built off of that model, the insights from that model, they are all that the model builder owns and retains control of those insights. And how do you do this not just as in terms of consumption of the output of the model, but across the life cycle. How do you make sure that data protection rules are in place through the model building testing validation and monitoring stages.

Explainability is probably pretty obvious. How can you explain the behavior of a model. Why was someone approved for a loan, why was someone rejected. When somebody applied for a job and that person was selected but someone with very similar qualifications applied that person was rejected, can you explain the behavior to the end user or to a decision maker.

Transparency, you want to be able to inspect everything about a model. Can you understand all the facts surrounding the model. Who built it, what data is being used, what algorithms, what packages are being used, who approved it, who validated it. All of these aspects of the model, facts about the model, should be easily available. Just like you know, you have you buy a food product and there is a, there's a label on it, you know, it has the nutritional facts, when was it manufactured, where was it manufactured, all of that. Just like that for a model, you should be able to get the facts of that model very quickly.

So, these I would say are sort of the fundamental pillars of Trustworthy AI. The challenge is making sure these can be done in a systematic way regardless of what tools are used to build the models and where the models are deployed.