**CRAIG:** Hi, I'm Craig Smith. And this is Eye on AI. This week. I speak to Seth Dobrin, chief AI officer at IBM, about the company's tools to increase the trustworthiness, fairness and explainability of AI models. Public confidence in AI systems is critical as decisions affecting us all become increasingly guided by machine learning. I hope you find the conversation as hopeful as I did.

**CRAIG:** Seth, could you introduce yourself?

**SETH:** Yeah, sure. So I am chief AI officer for IBM. It means I'm responsible for developing IBM's strategy and execution against that strategy for AI

**SETH:** this is critical for our success as a company, because as many of you are probably aware our strategy focuses on two things: hybrid cloud and AI. Our hybrid cloud strategy is centered, solidly around Red Hat, OpenShift, and multi-cloud delivering capabilities on any cloud for our clients.

**CRAIG:** We're going to talk about trustworthy AI. It's something that is increasingly in the news and concerns a lot of people. IBM has a product called FactSheets 360. That I understand is going to be integrated into products . Can you tell us what FactSheets 360 is?

**CRAIG:** And then we'll get into the science behind it?

**SETH:** Yeah. So let me start by laying out what we see is the critical components for trustworthy AI. At a high level, there's three things. There's AI ethics, there's governed data and AI and then there's an open and diverse ecosystem. An AI ethics is fully aligned with our ethical principles that we've published with Arvind our CEO, co-leading the AI initiative out of the World Economic Forum and I'm advisor for essentially open sourcing our perspective on AI ethics. From a governed data and AI perspective. It falls into five buckets.

**SETH:** So first is transparency. Second is explainability. Third is robustness. Fourth is privacy and fifth is fairness. And so the goal of FactSheets is to span multiple of these components and to provide a level of explainability that is needed to drive adoption and ultimately for regulatory compliance.

**SETH:** And you can think of it as, a nutritional label for AI. Where, nutritional labels are designed to help us as consumers of pre-packaged foods to understand what are the nutritional components of them. What's healthy for us. What's not healthy for us. FactSheets is designed to provide a similar level of visibility for AI.

**CRAIG:** I'm curious about the name FactSheets. Where did that come from?

**SETH:** This project came out of IBM research and it was one of these efforts that was put in the open source community.

**SETH:** So part of it is an open source package and it's really providing facts on AI and and I like it and it stuck because it's a simple name.

**CRAIG:** Beyond the label aspect, there are tools associated with it. Is that right for providing explainability or adjusting bias?

**CRAIG:** Mitigating bias.

**SETH:** Absolutely. So FactSheets is going to be integrated into a cloud pak for data, which is our premier data and AI offering. And it's, gonna deliver information about the AI from the data at rest. So from the data layer all the way through to, as you mentioned, capabilities that exist in Watson studio and Watson open scale around bias detection, bias mitigation.

**SETH:** In a privacy preserving manner. And it actually integrates with our GRC system, OpenPages With Watson, so that you can fully document that workflows are being executed against them and so that you'll basically have an end-to-end explanation, in FactSheets, and then an end to end documentation in OpenPages With Watson.

**CRAIG:** Okay, so let's break this down a little. On the label on the sort of nutritional label as you describe it, there has been a lot of talk about data cards. I think they're being called that would travel with a data set so that users of a data set could quickly see provenance, how the data was collected warnings, about how it could be misused and instructions for its intended use.

**CRAIG:** Is that what you're talking about here? Or is it broader than that?

**SETH:** So this is broader than that. We within enterprises and within our offerings in cloud Pak for data, we have Watson knowledge catalog. And Watson knowledge catalog is essentially the brain of a data environment.

**SETH:** That's where you can search for data. You catalog it. It applies policies. Within Watson knowledge catalog that's where you get all of those pieces of information you were talking about the. Requirements that automatically masks individual elements that people should see based on their entitlements.

**SETH:** It provides lineage and provenance. What FactSheets does on top of that is it ties that detailed level of understanding to the AI that's being delivered, and it allows you to develop policies. So it comes with specific set of templates, but you can modify and customize .

**SETH:** You can see how, your AI service was created, how it was tested, how it was trained. How it was deployed evaluated. What data was used, what regulations apply to it, or company policies need to be accounted for. And it delivers those through a set of model facts. And this allows organizations to automatically capture facts about the models. The capture of the model facts as we call them will be defined in the FactSheet template and it's across the entire AI life cycle.

**SETH:** Today, you have to manually do that without FactSheets, and because the process is automated, you get automated reporting.

**SETH:** And you can automatically generate a set of FactSheets. They are shareable. They're stored in a single location for all the information you need about the AI and everything that's been done to it. So when you think about provenance and lineage of data, you also need provenance of the model.

**SETH:** So it also keeps track of different versions. When you get into federated learning, keeping track of where different components of the learning came from, FactSheets will do that.

**CRAIG:** Can you describe how someone would use this? Is this a platform that you import your model and data into?

**CRAIG:** Or is it a tool that you run on the side that accesses some model and data?

**SETH:** Yeah. So today, there's AI FactSheets 360, and that is a set of notebooks that you can call and use in Jupiter notebooks and Python. And so you can do that today.

**SETH:** What we're doing, is the integration of these capabilities into our entire data and app portfolio that provide some of these value adds around policy creation and automated data capture and automated reporting by tying all of these capabilities together. So you'll have it out of the box across the entire life cycle from, as I said, from data at rest, all the way through to inferencing and scoring, even when we think about federated capabilities at the edge.

**CRAIG:** The part that Most interests people is fairness and bias detection and bias mitigation because a lot of these applications have, real world impacts on people's lives. Last time we spoke, we talked about deciding who gets a loan, for example, at a bank. Can you describe the bias detection and mitigation part of this, how it works?

**SETH:** Yeah. So when we look at bias detection and let's start there, there's a set of AI within, in this case, Watson open scale, that FactSheets will read. And within Watson open scale, you have the capabilities to identify undesirable distributions. You have the ability to identify what are the protected features of the data.

**SETH:** So protect the data elements. So for instance, gender would be quote unquote protected because you don't want bias and gender, race, ethnicity. Those are out of the box capabilities that we provided. You can also customize for what types of bias you want to detect for. And so the detection is fairly straightforward.

**SETH:** When you look at how we remediate the bias, it's done a couple of different ways. So one is as your model is being deployed and scored and monitored. It actually creates multiple different versions of the output so that you can look at which ones are within your bias parameters, if you will, and which ones are outside of it.

**SETH:** So it tunes the hyper-parameters it adjusts the different weights of the features, and it allows you to be able to see, okay, when this model falls out of the parameters for fairness, for a specific bias - it doesn't just check for all biases because I don't know that we'll ever be able to do that, but for specific predefined biases, it creates these multiple versions of the model.

**SETH:** And it looks at, which ones are within the distribution that you're desiring and which ones are out of it.

**CRAIG:** And then whoever's managing the model would pick the optimal model within the distribution, and then discard the other ones.

**CRAIG:** Is it simply a matter of balancing the data set? So on gender, if you see the distribution, the top of the bell curve is on the male side and you want to shift it more so that there's an equal number of male and female on either side of the curve.

**CRAIG:** Is it as simple as that?

**SETH:** So I guess a little more detail on how we detect the bias. So I said it creates, maintains multiple versions. It also flips those protected elements. So were I to run the model as a male. And then I flipped features and known correlated features to be female.

**SETH:** Do we get the same answer, same thing for, race and ethnicity. And so when we talk about how is the bias remediated or mitigated? We actually don't remediate it because remediating means fixing the underlying problem, which is the data, right? And we're not fixing, we're not doing anything to the data at this point. We're mitigating the bias.

**SETH:** And the way we do that is we adjust the feature weight and sometimes we adjust the hyper parameters. So the tuning of the model to account for the bias at enough of a level that you mitigate it to what is acceptable for you as a company and over time, and this is why it's important to keep monitoring the models, over time the data that is being trained on will change how the model is scoring. And there are likely some trade-offs you're making when you're tuning these hyper-parameters or adjusting these feature weights, and that will allow you to stop having to account for them in the model itself because the data is getting better.

**SETH:** And you also will have some instances where the data is for whatever reason, getting worse. And so you want to make sure that you're monitoring and make sure that models that were fair and balanced, don't become unfair and unbalanced.

**CRAIG:** When you're talking about the data, you're talking about the training data or the data on which the model is operating

**SETH:** the training data.

**SETH:** So it's important to remember that some people might argue, but the math itself is not inherently biased, right? There is some bias in the opportunity for bias and the type of model you select but the math itself is not biased. The math learns the bias from the data that it's trained on.

**SETH:** And so the short answer, your question is yes. I mean that training set and, the training set is the data and the bias came from historical decisions of humans that were biased. And so as we make less bias decisions over time, that data will become less biased.

**CRAIG:** Okay. Because the data is being fed back to the training set.

**SETH:** So as you retrain data, over time, you will dilute out the bias data with the unbiased. And eventually there'll be a point in time where you may just eliminate the part of the dataset. The time period prior to a certain date, when, there was biases, affecting your model, but you need enough data to move forward that you feel comfortable is unbiased, or de-biased.

**SETH:** So when we initially spoke and we talked about, the mortgage example where, and this was in New York city, people in certain zip codes in New York city were getting denied mortgages because the data had picked up historical artifacts of red lining.

**SETH:** And so people of color were denied mortgages at a higher rate in New York city. And a lot of other areas, people of similar backgrounds tend to live in the same kind of area. And so there was a correlation between in this case, race and ethnicity and zip code. And so the model picked up the zip code.

**SETH:** And two things. One is, red lining was banned, right? And so if I have enough data to train a good model today that doesn't use data from when red lining was a thing, I would certainly do that but we also know that they're still biases that are inserted into decisions that individuals are making.

**SETH:** And so you may eliminate the red lining data, but humans still had biased decisions. And so as you're using AI to drive decisions, you'll reach a point where you have enough data that you don't need to use that data before a certain date that has this bias in there. So you could essentially eliminate some of the data in your training model.

**SETH:** You may decide not to do that. You'll still have to account for the bias, but there is a point in time where you're generating enough, less bias. We'll call it data that you don't need to rely on that previously bias data.

**CRAIG:**

**CRAIG:** And then why not go in and try and rebalance the data itself.

**CRAIG:** Why do this in the parameters?

**SETH:** So you essentially do it in the data. By tuning the feature weights, you're accounting for the bias in the data just by correcting, by eliminating data elements in the biased data, you may introduce other types of bias. And you may essentially eliminate an entire race or an entire gender from a dataset, because how do you know which specific decisions were biased that was on that race or ethnicity or that gender and how do you know which ones aren't biased. And so you're actually would be creating a new bias where you're training data only on white men.

**CRAIG:** And these are neural nets. You're talking about. How do you know that in adjusting these weights you're not introducing a bias that isn't obvious at the beginning.

**SETH:** So that's why we approach it the way we do where we create multiple essentially versions of the same model.

**SETH:** And you would pick that up. So if you were trying to adjust for say gender bias and somehow by adjusting for gender bias, you created an ethnic or racial bias. You would pick that up because you'd be monitoring for it and you would work to retune the model, so it didn't have that, or that version of the model just wouldn't be selected.

**CRAIG:** Yeah.

**CRAIG:** And how do you monitor that , is that another tool then that scans the inferences for distributions?

**SETH:** Yeah, it's all in the same tool. So it's all in Watson open scale. And so Watson open scale evaluates the model before you put it in production. Think of it as a lifecycle management tool.

**SETH:** So it continues to monitor the model as it gets retrained and redeployed over time. And so it's not a one and done type capability. It's an ongoing part of an AI life cycle or ML ops, or even broader it's an AI governance tool, essentially, especially when combined with open pages.

**SETH:** Today, at least, and probably in the foreseeable future. We can't eliminate bias period, full stop, we can monitor for and mitigate specific types of biases that are predefined.

**SETH:** And and because where we're monitoring the output of the model, so we're actually monitoring as the model is inferencing or scoring the underlying model is irrelevant for the detection. If you define hair color as a potential bias and you're tuning for one, and you create an uneven distribution for hair color, then you would not use that version of the model.

**SETH:** You would look for another one that didn't contribute to those predefined biases,

**CRAIG:** but discovering whether there is a bias at that you haven't accounted for is just a part of society's developmental process, right? The initial racial and gender biases of, neural nets were not recognized until they were in production and people started discovering them.

**SETH:** So , social and ethical construct, right? What is bias is not consistent over time, right?

**SETH:** What a society would consider being biased is not consistent over time. And We don't claim to be able to, and I don't think anyone does to be able to identify any kind of bias without having it predefined. So we can detect and mitigate any kind of bias , but you have to define it upfront, right?

**SETH:** So it's not just randomly looking forward that here's the biases that I care about as an individual. Here's the biases I care about as a company. And here's the biases I care about as a society. We want to make sure a specific decision isn't driven by an undesirable bias, right?

**SETH:** And the reason that these algorithms are so good at selecting based on biases is because what you're asking a machine learning model to do is discriminate multiple things from each other and so in this case, if you're selecting based on bias, the easiest way for the model to segregate again, back to the mortgage case, a likely person to pay their mortgage had some racial and ethnic component to it. Not to say that people of a certain color or ethnicity, we're less likely to pay a mortgage. They were just less likely to get a mortgage in the past.

**SETH:** And so the algorithm pick that up because that's what it's doing. It's saying this group has not historically gotten mortgages. It's this much less likely. And so we're going to make it this much less likely. And those are the types of things that we correct for

**CRAIG:** I want to talk about transparency and explainability, in neural nets that obviously is a difficult thing to do.

**CRAIG:** What can FactSheets explain,

**SETH:** So the simpler, the model, the easier it is to explain. So some models you can certainly explain just by looking at the features weights other models, especially when you start getting into ensemble models or deep learning models, get harder to explain because there's different contributing factors or with ensemble models, each model you have to explain individually and then daisy chain it altogether. In the case of deep learning models or neural networks, as they're sometimes called.

**SETH:** Those are what we think of as black box models. And so being able to explain how you do it as an impossible, just by looking at the feature weights, cause you don't see the feature weights at each layer, right? And so it evaluates the models again, based on the outcome. And it uses a set of proprietary algorithms that were developed in IBM research to interpret based on.

**SETH:** Couple of different methodologies, what are the underlying features that are contributing to the data? So it essentially extracts the feature information from the models. And in this case, the hardest ones are the deep learning models.

**CRAIG:** And in the case of healthcare applications, Where explainability becomes important for a variety of reasons, not least, which is liability is that level of explainability adequate,

**SETH:** I think it's better than what we have today. Or what we've had before. Any level of explainability is going to be more valuable than none, is it perfect yet? I would say

**SETH:** no.

**SETH:** Do I want a model making a decision about my health independent of a physician just because it's explainable? No. Even when it's explainable I still want my physician and I to have a conversation about what we think is right for my healthcare. Having this level of explainability allows us to have a more informed conversation, cause my physician, she and I could have a conversation and she could say, okay, here are the three recommendations that we have for you. And it's based on a model. And here is why we think each of these three things are the right one to do. And we firmly believe that, the goal of AI is not necessarily to replace humans, it's to augment humans and help us make better decisions.

**SETH:** Explainability only helps us make more informed decisions.

**CRAIG:** How do you expect this to fit into an eventual regulatory framework? There are developing standards around this that then will inform regulations that will guide the use of AI for different applications. I remember you've done some work on that. Can you talk about that?

**SETH:** Yeah. So our government affairs group is definitely working with teams internally that I'm part of to help define recommendations that we will.

**SETH:** We'll make an, a series of white papers. What we think AI regulation should look like

**SETH:** separate from that. It's very likely, especially looking at the regulations that are already in the works there will be some level of explainability that's required. The depth of that explainability is probably going to vary from industry to industry. So for instance, you'll have banking, specific rules, you'll have insurance specific rules, you'll have healthcare specific roles, and I'm sure many others around what explainability means for those regulators

**SETH:** you will also probably need to document that certain types of biases again, predefined biases do not exist and explain what you did to account for them. And a big part of this is data privacy too, right? We've always said as a company that we believe your data is your data, your models are your models.

**SETH:** And that goes for both companies, as well as the individuals that those companies interact with. We firmly believe that tying back to regulations, like GDPR and CCPA, you need to make sure that you're protecting that data that's going into the models and following the consent that was given by the individual in terms of how that data is used to help us make business decisions.

**SETH:** And so some people opt in for their data being used in certain types of use cases. And they opt out for others and you want to make sure that you can document that their data wasn't fact excluded if they asked it to be excluded and what's not known yet is in both GDPR and CCPA and other similar regulations I can opt out after I've already opted in.

**SETH:** And so does that mean that okay, Craig has opted out to allow iBM to use as data to make marketing recommendations, but he previously opted in we've since built models. Does that mean I need to remove your data from the model or can I keep that data in the model? Because it was trained before you opted out.

**SETH:** Those are the types of things we don't know what's going to happen. If the decision is that we need to remove your data from a model that is not an insignificant feat to ask companies to do. It could require that you entirely retrain a model and for simple models who cares, but as you get into some of these highly complex models, like neural networks, there's a significant amount of time and compute that went into just the training part that you would have to either figure out how you extract individual data elements data fields, from, for an individual.

**SETH:** So a cell essentially.

**SETH:** How you would account for that, if you wanted to just remove it versus completely retraining it. So I hope we don't get to that level of control, but I think since I opted in, at one point up to that point, the company should be able to use my data for the models.

**CRAIG:** And does FactSheets include some privacy protecting tools and encryption tools?

**CRAIG:** Or you mentioned federated learning.

**SETH:** Yeah. So as I mentioned, FactSheets is fully integrated into our cloud Pak for data offering. And that goes across the data management and data governance, policy enforcement, and FactSheets captures that information across the whole life cycle from data at rest, through training, through scoring through different iterations of the model.

**SETH:** And so FactSheets captures all of that information. It doesn't actually stop you from using the information, right? The data inappropriately. That's what we use Watson knowledge catalog for, but a documents that you did use it appropriately.

**CRAIG:** And to that point with the managing the bias of models or the bias of the inference from models.

**CRAIG:** Is there a danger that people could use models in abusive ways to target, to actually create biases?

**SETH:** Yeah, so that would be picked up in, in our case Watson open scale, when we're doing the bias detection.

**SETH:** And again, what the definition of the bias is set by a set of policies or a template in the case of fact sheets. And then you define what is acceptable and what's not. And if I were to create a model and I was biased against New York times reporters.

**SETH:** And so I always excluded New York times reporters from a mortgage decision. If that was a predefined bias that we don't want this model discriminating against journalists, it would pick that up in Watson open scale, cause it would see an uneven distribution where journalists are less likely to get a loan.

**SETH:** And FactSheets would document that. And so even if I had a bias against, in this case journalists, that would be documented that it exists and it would automatically be corrected by the model. And even if I was the one kind of, controlling whether or not we decided to mitigate this bias, it would be documented that I didn't.

**CRAIG:** And that's where then the regulatory framework comes in. There would be some reporting requirement presumably

**SETH:** even within a company, right? Even if it's not a regulated area within a company, companies should have policies around what's acceptable and what's not right. We all have business conduct guidelines, whatever we call them.

**SETH:** This would clearly be a violation of a business conduct guidelines. In my mind, if someone. Intentionally created a bias model, right? And so even absent regulations companies should have an AI governance process where they define what's acceptable and what's not to some extent tied to their overall governance of the company and make sure that it's crystal clear to people what's acceptable and what's not. And what the consequences are

**CRAIG:** That's it for this week's podcast. I want to thank Seth for his time. If you want to read the transcript of this episode, go to our website. I on AI that's eye-on.ai I encourage more listeners to reach out to me.

**CRAIG:** My email is craig@eye-on.ai.

**CRAIG:** We're hearing from more and more listeners around the world. And I'd love to hear from Nepal and Kazakhstan. Remember the singularity you may not be near, but AI is about to change your world. So pay attention.