AI and You

Transcript

Guest: Michael Hind

Episode 74

First Aired: Monday, November 15, 2021

Hello, and welcome to episode 74!

We've got some terrific learning about an important and very current topic in this episode, because we're getting into explainability. My guest is Michael Hind, and the reason he's here is that I set out to find an expert on the explainability of AI, that led me to IBM, and that led me to Michael. I contacted him and he was happy to come on the show and is in fact a listener.

Michael called in from the IBM Research AI department in Yorktown Heights, New York where he is a Distinguished Research Staff Member. His current research is in Trusted AI, focusing on governance, transparency, explainability, and fairness of AI systems. He has authored over 50 publications, served on over 50 program committees, and helped launch several successful open-source projects, such as AI Fairness 360 and AI Explainability 360.

Now, we're going to explain… explainability… that sounds very meta, right? in the show, so I'm not going to set it up much now, just enough for you to want to keep going.

We think we know what explanations are, and how to give one, and it seems very obvious and intuitive, but it isn't. Think about a grade schooler who's always asking, why this, why that. Suppose they say, "Why is fill-in-the-blank president of the United States?" So you explain the electoral college, the constitution, the vote tallying process, okay, you're feeling like you've got this one nailed, and then they ask "But why did they vote for him?" and you realize this is going to take longer than you expected. Context is everything, right?

I saw a neat video on Wired recently where computer scientist Hillary Mason explained what machine learning was to five different people in succession: a child, a teen, a college student, a grad student, and an expert, and at each level she adjusted to the background and experience and most importantly, the needs of the person she was talking to. It's a great example of how we fit our communications to the context. I pitch this show to a certain audience – smart, interested, but not expert. We can't please everyone, but I'm always thinking, "Who is going to be able to take this and make a difference with it, and what would help them the most?"

Explainability in AI is about decisions that it made. If you asked someone to explain how they recognized a friend's face, it's unlikely that the answer would be useful, in the sense that you could take it and then go out and recognize that friend yourself. Not unless that person has a coat hanger through their nose or something like that. We don't learn how to recognize faces like that, we ask them to show us a picture of their friend and then we learn it as a pattern. And AI can learn to recognize faces the same way, through patterns. It can learn how to make decisions that are really important, like whether to grant someone a loan or where a crime is likely to be committed, but now it's no longer good enough

that it learned that as a pattern, because people are going to want explanations of those kinds of decisions, and you can imagine how important the answers will be.

This is one of the most important and one of the hardest problems in AI, and it's pivotal to how it's being used today, because without explainability you can't responsibly use AI for important decisions like what treatment plan to follow for someone with cancer. That puts Michael Hind front and center of the bleeding edge of applied AI, so let's hear what he has to say.

Michael, welcome to the show.

Thank you. It's a pleasure to be here.

So, I have been wanting to talk about explainability with an expert for ages. So, if I get enthusiastic and obsessive here, just remember I told you that was coming. And you've been in this field now for long enough that there must be something pulling you towards explainability. Can you just - because we're going to drill down into the weeds fast enough - let's go up to the 100,000-foot view and tell us why you got into this. What appeals to you about this field in particular?

Sure, sure. So, for several decades, I've worked in the research field around software engineering, and programming languages and program analysis and things like that really geeky stuff. And then about four years ago, I heard about this problem of how do you get AI to explain itself? And I thought it was a fascinating question. And I wanted to dig deeper and understand it more. And what I found is that, you know, usually, as a scientist, you want to have a nice, crisp problem statement. So that you know when you solve it, you've solved it. And in this particular case, there wasn't a crisp problem statement, which to me is fascinating, and we can go deeper. But that's what really drew me to this space. And then as I got into the space, I learned about things like bias and fairness, and transparency and governance and the whole bunch of topics that you've been talking about in your podcast.

You just described the entire Ph. D. program. And so explainability is what Pamela McCorduck described on the show as a suitcase word, I think she said Marvin Minsky came up with the term: we throw them around all the time, but when you try and unpack them, there's a huge amount in there, like consciousness and thinking and creativity. So, let's just start out at the 50,000-foot level with explainability. How would you describe that?

So, it's interesting, because there's what how the masses view this problem, and then what the sort of technical community defines it. So, the technical community defines explainability as: A model has made a prediction, and you want to know, why did it make that prediction? That's sort of the vanilla version. A slightly more general one says, "I would like you to tell me in general how the model makes decisions." So an example would be a flowchart, if you can show me a flowchart of how a model works like a decision tree and they call that a global explanation. But it turns out, that's not usually the definition that most people use. So, for example, NIST just had a workshop a few months ago, it was three days long, and it was about explainability. But I would say 80, 90% of the people in the audience weren't using a definition I gave, they were using definition, what I would call more *transparency*, tell me everything about how the model

was constructed? What was the training data set? What's the error rates for different groups, and so on and so forth?

Right? Let's unpack that distinction in a little bit. But this idea of an explanation, are you pointing out that the people who are working on this, who have to produce results on this are treating it somewhat differently, and I want to pick out the differences between what people outside the field expected means and those people inside but I think we have to do a little more due diligence or exploration first. Explanation of something is, if there's a demand for it, it's not an inherent property. It's something that we want out of system for a reason. And if, for instance, Netflix says, watch this movie next, I think you will like it and you hate it, you're not going to sue. But if it's making a medical diagnosis, if it's denying an insurance claim, then explanation has a lot of weight attached to it. And humans can construct explanations for those things, but - oh, you mentioned flowcharts in the good old fashioned AI ers a flowchart would have been the way that the thing was written. Machine learning today - describe how it's created and how much that lends itself to explainability?

Yes, so it's quite interesting situation we got ourselves into, right? So, on one hand, the appeal of AI is as we don't need to create those flowcharts, or those programs, right? In particular machine learning, right? You give it a bunch of examples, teach it that, given these examples, the prediction should be either yes or no, approve the loan or deny the loan. And then, through some interesting math and computer science, you're able to learn some patterns. But what's actually learned is not necessarily the same way we think about things. So, the challenge is that when you then go to that model that was built, that's learned these patterns and say, well, why did you think that, it could be things that are totally unrelated to the way we think so some examples. There are some famous ones where I think there was an image recognition system that was trying to recognize polar bears. And what it was discovered - it was trained on a bunch of pictures - and it was recognizing things that were pictures of animals that were based in with snow backgrounds, right? And so, it wasn't picking up on the bear, it was picking up on the background. And so that as a human, we would say, well, that's, that's kind of silly, but it's just looking at these pixels as pixels, not as you know, as animals. So that's the challenge is that it's operating at a different level than we are. And then when we make the requirement that it needs to explain to us, then there's this big disconnect.

Yes, is there this entire section of explainability, that sort of like, *Here Be Dragons* that we've fenced off for now? You mentioned, the polar bear example, reminds me of a similar one about sheep, that the algorithms for recognizing sheep in pictures found that because they were always appearing in landscapes, they would often identify pictures of landscapes that had no animals in whatsoever as sheep. So clearly, if we asked for an explanation of why did you say there were sheep in this picture, it wouldn't look anything like what we would say, well, there were white fluffy things with four leg. It would be related to something inscrutable that didn't even include parts of the image that were only inside the sheep bounding box. So when, so when we talk about explainability, in, in a practical sense, like it *has* to work, we're not talking about sheep, we're talking about insurance claims, how do we fence off or compartmentalize it

to be able to do something useful, as opposed to being bogged down in what sounded like an unanswerable question with the sheep?

Right, so let's take an example. Let's say, again, back to the loan example, you fill out an application, a bunch of questions that gets fed into this magic model, and outcomes, an approval decision, I guess, more likely, the one you'll care about is if you get rejected, you want the explanation. Typically, people don't want to know why they got approved, they're happy. So, you get rejected, and you want to know why. And currently, there's different approaches to how to answer that question. One approach, very, very popular, is to say, I'll give you an explanation based on the answers you filled out in the application. So, when you train the model, you show it a whole bunch of filled-out applications, you tell it for all these applications, which one should be approved, which one should be denied, out comes a model. And now you give it a new application, it has the answers filled in, and the model says Deny. And the question is, well, why did you do that? And so, sort of a basic intuitive assumption is to say, well, it was probably based on the answers that you put in the application, because that seems to be the only information available, the features in a technical sense. So, then the question is, okay, what can we say about those features? And now it gets challenging, because the question is, I could show you equations. But that's not gonna be comprehensible to anyone, often even a data scientist. So, what I can do is I can probe the model, what I can do is I can say, I'll take your application, and I'll tweak some of the answers. And I'll send some, do some tests and see if it changes the decision to from denied to approved. And then based on that I can build up sort of a boundary condition to say, okay, it seems like these are the important features, and then it you know, can give you an "explanation." The problem there is it's not the model that's telling you why it's sort of like almost a third party. I mean, it's a piece of software. It's a third party doing a probe. And based on these investigations, this game of sort of yes or no and then builds up what it thinks that's a good explanation.

I like that answer. It hadn't occurred to me that by tweaking the parameters in the answers to the questions, you could find where it's sensitive, and where it changes, so you can see which answers were pivotal. And there's so much more to this of course… If we were to go back to the days of expert systems, this would be relatively easy: You write a PROLOG program that got umpteen rules in it, and it's trivial to get it to explain its answer. You may not *like* the answer, but that's exactly how it got to the answer is by following that chain of reasonings, which [are] basically a giant set of if then else blocks with maybe some fuzzy logic in them. Now we have models that build up parameters. And one of the things inherent in what you said was that we identify the reason for the answer as being in the data was input. But it could also be in the training data that built the model. In essence, it is in the training data. But now that brings up bias, because what if the real answer is well, I denied this claim, because you're in a red line district, which is illegal, but that was what it learns from its training data, right?

Right. Yes. So that this is an example of where the explainability and sort of bias or fairness sort of sub expertise, you know, domains kind of overlap a little bit. But what you said, I think, is actually interesting, because another class of techniques is one, typically called Prototype-based. And what that actually says is, it doesn't try to explain why I denied this person based on their

answers, it says, here are three people you train me on, let's say, three, for example, here are three people, you trained me on that I think look a lot like this person. And for each of these three people, you said they should be denied. So therefore, that's why I made my recommendation that you should be denied. Now, clearly, there's issues with privacy, you don't want to give out the training data to customers and so on. But this could be useful, for example, to the loan officer, who's trying to understand what's going on, why is this recommendation of the model to say, hey, this person should be denied?

Do those explanations satisfy? I'm reminded of the principle that we judge people by their intentions and machines by their actions. Because we don't think machines have intent. So, in asking a human for an explanation, we're probably trying to see if they had some hidden intent there, can they provide an explanation that satisfies us that they were doing the job properly? In the case of a machine, do people look for the same kind of explanation that they would have gotten from a person if a person was making the decision?

Yeah, so another fascinating question topic. I think there's a big disconnect. And it depends on who the people are, that are asking for it. So, I'll give you some examples here. Let's say I had an AI model that was making a decision on whether someone should be convicted for a crime or not. That's one example. Another one is whether someone should be hired or admitted to college, or their results on a big exam? Okay. In all those cases I just mentioned, no one gets an explanation. Right? They're done by humans, right. But when they're done by an AI system, there's concern that, hey, there could be biases, or whatever, it could be a bug or whatever it may be. So, we need an explanation there. So, I'm not saying we shouldn't say AI systems, high risk systems, should give explanations. But just going back to your analogy of comparing to what we do with humans, and not it seems like even there right now, there are some cases where we seem to be okay. I mean it's shocking that I could be convicted of a major crime and not be told why except that 12 of my peers in the US felt it was reasonable.

You make a good point there. But this is where the bias overlaps with the explainability because the things you described have all been subjected to revelations that they were not being done fairly. So, if we can apply some kind of statistical or objective analysis of those decisions and determine that they were not fair, then we can do something to the system, even though that may not look like asking for an explanation. Like for instance, orchestras used to have 6% female composition in 1970, that went up to 21% in 1993, and they moved to blind auditions, so even though none of the people who were doing the auditioning would have said or thought they were sexist, and no explanation would have satisfied, the statistics leads to no other conclusion than that there was unconscious bias. So, can we satisfy explainability requirements by just analyzing the heck out of the system for bias?

So, I think that's one concern, why people wanting explanations without a doubt that they want to make sure that wasn't a bias. Actually let me take a step back. So, if we talk about, I think this was mentioned on an earlier episode of yours, there's a good bias and the bad bias, right? There's a mathematical bias, which in the case of a loan example, a loan approval example would be, maybe the model has a bias towards people who make a higher salary. That seems

appropriate, right? It seems like if you're looking for credit worthiness, higher salary is probably a good indication or outstanding debt or whatever it may be. But if a decision depends on a person's gender, or age or something else, that shouldn't be the case, that's illegal, immoral, and so on. So, in terms of explanations, yes, you can be interested in explanations to see if there was a bias, like if the explanation comes back Because you were male or female, that's why you didn't get the loan. And clearly, that's not good. But there also could be other reasons, right? It could be the case that you wanted to know Why didn't I get into the school? And maybe I didn't get into the school because my dad didn't go to the school. Okay, is that illegal bias? I don't know. But there are other reasons that you probably would be upset about that aren't necessarily what we would call sort of societally biases. But things that are, are just don't feel right to you. Or maybe it's a situation where it's a job interview, and you want to be you want to apply again in six months. So, you want to know, what can you do? What kind of actions can I take to address or tell me why I was denied? And I'll go and get more training or whatever it may be.

Just to chunk this up a level because we've been talking about this well-known set of high stakes examples of insurance and job applications and so forth. There are other perhaps classes of explanation that I want to see whether they're part of your radar or not. I'm thinking about the hypothetical question of asking a self-driving car Why did you turn left here? And the answers might be My front wheels pivoted for 1800 milliseconds, or, That's the way that the routing algorithm told me to go or that You told me to, as you know, because it's going up to the root of everything, and, and other possible explanations. But the context for that might well be because those answers are all correct, but probably useless. But the if that was an unusual way for it to take to get to its usual destination. And that was an unusual question to ask, then probably the answer you're looking for is, the usual route was blocked by an accident. And now that's a different kind of explainability. We're not talking about anything where bias enters into it at all. Is that part of your radar? Or are you only focused on the high stakes kind of things we were just discussing?

No, it definitely is part of the radar. It's, I would say any model, any machine learning model, or AI model that's making a decision or prediction, you want to know why? That's, you know, that's under explainability. Right. And in fact, let me just give a third example, which is one that I worked at, I was involved in approach to solving this problem. And this is what my original motivation to get into the space led me to. And it may be an example where I was sort of naive and new to this space, and either came up with something that was not brilliant or just silly, because I didn't know the space. So, when I discovered that it wasn't a well-defined problem, right? So, for example, if you asked me like you said, why did you make left turn you know, the answer that depends on who you are like, if you were the engineer, that I probably will give you some technical terms about that. If you're a regulator, maybe I'll give a different kind of answer not saying none of these are wrong. It's just what's the appropriate technical jargon and level of detail that I give? And so given that, in general, the general problem was, how do you give an explanation that satisfies all these different personas and users, to me, it just seemed impossible. Right? That, you know, you're saying, it's kind of like when you get into an elevator, you see a random person? And they say, what do you do? You kind of look at them and say, first of all,

how much time do you have? And how much interest do you have? And what level of detail and so on? Right? And so that motivated me to say, I don't think we can actually solve that problem. Because I need more information, right? And luckily, with humans in the elevator, we are able to, to kind of say some things and see how they react. And based on how they react, we can say some more. So there's our interaction aspect. So that led to this this algorithm that we call Ted explainability algorithm called Ted. And it basically says, you need to teach the explainability component, what you mean by "good explanation." So just like you teach it, here's a bunch of here's a loan application, and here's whether it should be approved or not. The training data in that training data, also put Why. And it's like that when you have a new employee, just dealing with humans, new employee comes, you're doing training, you're say, Okay, let's do you read all the books. Now, let's do some examples. Okay, this person should be denied. Why? Well, because blah blah, this person should be accepted. Why? Because blah, blah, whatever that blah blah blah is, you write down and give that to the system. And then when it's time for it to make its own predictions, it can give you explanations, similar to what you gave it. So, it's kind of a different, a very different approach that we've been working on more recently.

Hmm. And as you say, if you're in the elevator, and some random person comes in, you have no idea who they are, and asked you that question, then it's this enormous space, the sort of space that, say, a librarian deals with people coming up to a reference desk with a question. And I like to use the example that we don't treat them the way that we treat Google and just bark at them, "Avocado fertilizer" and expect them to know what we're looking for. But somehow Google does, but with a librarian, they would, number one, say, that's kind of rude, and number two, they would ask questions, there would be a dialogue, and they would get rapidly to the kind of thing that we wanted to know. The systems you're describing have this idea of engaging in a dialogue to find out more about the context? Or are they defined for context to be executed in a context that's already well known?

Yes. So, I would say we're very early on in this field. It's one of those weird situations where society answers now, which is great, great problem to work in. But the technology is I'd say still early on. So, the way that the community, the scientific community has looked at this is more of this sort of one-shot kind of explanation, we are talking about as a dialog interactive, you know, Star Trek computer is a great example. And we're not, there's not much work going on in that space yet. But that's something that that is certainly attractive.

So, you deal with customers a lot in your work. And so, these are people that have to get an answer that works. There's money at stake, they can't afford to dabble, in theory, they want to make sure that they don't get sued for only hiring white people ,or something like that. So, in that, in that general sense there, hat approach do you take for providing them with solutions that are going to satisfy that critical demand?

Right. And so, it kind of gets back to, you know, what's their problem statement? Right. What are they asking for? And it's quite interesting. Sometimes they're motivated by regulation. So, for example, in GDPR, there's a statement that says an automated decision-making system needs to give a "meaningful explanation." Another example is the State of Illinois that passed a law

about video interviews, and particularly AI-scored video interviews. And they were concerned that these can have issues. And they specified a number of safeguards, and one of the safeguards was, the provider needs to explain how the system works. Right? So, I mentioned these two, because here's regulations, but they're not really well defined, what is meaningful explanation, what is explain how it works, and so on. So that's some of the motivation from companies comes from there. But most of them is, you know, they want to be able to have assurances to their customer that decisions were made in a reasonable way, or more likely, the AI system is assisting them, right. So, think of again, back to the loan example. It's not that the AI system is going to decide who gets the loan or not, it's going to give feedback into the loan officer, who's a human who's gonna make a final decision, and that loan officer may want to know, hmm, that's interesting, you know, I would have approved this, you know, why ,AI system, do you think that this is should not be approved? So, in that case, something like, you know, give me other examples might be acceptable, because it's only going internally within the bank. And so, we had an engagement with a major company that was trying to build up their own expertise and explainability. They realized the maturity is not totally there yet. And they realize that depending on the use case, depending on the kind of AI system, different techniques may be more appropriate. And they were building up basically their own center of experts to augment the data scientists that say, okay, when a data scientist creates a model, we want to sprinkle on this explainability dust, you know, literally pull out another explainability technique to be able to augment the model to give some level of explanations. But it's still, I would say very early. There's a lot a lot of interest in this as you can guess. And it's still, you know, I think where we are is, is we need the science to be able to, to be played with right, to be experimented with to get feedback from the pragmatic, you know, with customers, to back to the scientist to then help solve the problem in advance things and to find out that maybe that you know, that the *Star Trek* version is the one that we need to put more focus on. And so, we're in that experimental stage, I'd say right now.

And the *Star Trek* version being what?

So interactive, where you get to have a conversation with the system. It's not just, you know, here's the explanation. And that's it, you can't ask for any clarification.

So, you describe the scenario there, where the human is still making the decision, and responsible for the AI is a tool that's helping with that. And that reminds me of chess computers. You're not a thousand miles away from Deep Blue, and some history there. I actually just had a chess grandmaster on the show, Jonathan Rowson, and we were talking about this aspect of technology and how much computers are now part of the life of chess players. So, then that's setting a different bar, and the AI hasn't made a decision that it's accountable for. But it's providing assistance, that doesn't have to be a decision - or is it a decision and then you ask it to explain it, but you don't have to accept it?

Right? Yes. So yeah, you know, those, all those scenarios are important. And one of the things you, I think, pointed out, is, I think I'm not a chess player, but from what I've heard is, a lot of great players use computers to help get insights for new moves, new techniques, and see how

things play out. And that's an important thing I should have mentioned earlier is that the earliest work on explainability was motivated by data scientists who were trying to improve their model. Right? So, it was more like debugging use case, can I get some kind of insight into what's happening here, to make it more effective, more accurate, whatever it may be. And that was a line of work that was going on for many, many years before society started to say, hey, now that we're deploying AI, more mainstream high-risk cases, I need an explanation. And that's one of the reasons why there's this disconnect between a lot of a science that happened, which was focusing on that sort of debugging use case and the need from society, the average consumer, maybe, what do they need, what kind of explanations they need, as well.

Okay, we're going to finish this interview next week so you don't get your download or your attention span overloaded.

I *really* get engaged with this topic, I mean, we're at the dawn of a new era in the use of cognitive machinery, and we're learning about it from someone on the front lines. One of the reasons I love doing this is how engaged I get with it. It's like the flow state that Mikhail Csikszentmihalyi talked about. I've got an example. You know how you can drive from one place to another and not realize how you got there, because you were on automatic? Well, we've got a road near us called the Humpback. Used to use it a lot on certain journeys. And you cannot drive the Humpback on automatic. It's a single lane – not in each direction, I mean for many sections it's only one lane wide altogether, including where it goes over a hump – hence Humpback – and curves around, and there are trees right up to the road so you can't see past the curve, and in some places there are signs telling you to honk before you proceed, as the only way of warning someone coming the other direction. You've got to be a hundred percent present just to survive the Humpback. I don't drive it for the heck of it, but it certainly lets you know you're alive when you do.

The same thing happens to me when I'm coaching, because I'm a coach as well, right, and the coaching is going well, and I go into a flow state, and everything disappears, including time, it's just me and the other person. I love that. And I find that talking with our guests, I get into the same state because I'm so engaged with the dialogue and the questions. I hope that a little bit of that at least transfers itself to you, the listener.

In today's news ripped from the headlines about AI, Facebook has developed an AI system that it claims to be a big leap forward in machine vision. In a set of blog posts titled "Teaching AI to perceive the world through your eyes," which is clearly leveraging their big bet on augmented reality and virtual reality, they say, "Imagine your AR device displaying exactly how to hold the sticks during a drum lesson, guiding you through a recipe, helping you find your lost keys, or recalling memories as holograms that come to life in front of you." Their Ego4D project is by a consortium of 13 universities and labs across nine countries, who collected more than 2,200 hours of first-person video in the wild, featuring over 700 participants going about their daily lives. The Ego part of the title is because it's all about the viewpoint of the user, and they've developed five benchmark challenges centered on first-person visual experience to gauge the progress for future AI assistants. Those are:

1. Episodic memory: What happened when? (e.g., "Where did I leave my keys?")
2. Forecasting: What am I likely to do next? (e.g., "Wait, you've already added salt to this recipe.")
3. Hand and object manipulation: What am I doing? (e.g., "Teach me how to play the drums.")

4. Audio-visual diarization: Who said what when? (e.g., “What was the main topic during class?”)
5. Social interaction: Who is interacting with whom? (e.g., “Help me better hear the person talking to me at this noisy restaurant.”)

This is big because as they point out, the first-person point of view is very different from the external point of view – think about how different the view of a roller coaster ride is from the front seat versus looking at it from the ground.

Next week we’re going to finish the interview with Michael, starting with him explaining the TED algorithm, which he mentioned there, which stands for Teaching Explainable Decisions, and how an AI can actually learn to give explanations, plus some of his history with the Watson project, which was IBM’s Jeopardy-winning AI, the distinction between transparency and explainability, and much, much more. That’s next week, on *AI and You.*

Until then, remember: no matter how much computers learn how to do, it’s how we come together as *humans* that matters.

http://aiandyou.net