AI and You

Transcript

Guest: Michael Hind

Episode 75

First Aired: Monday, November 22, 2021

Hello, and welcome to episode 75! Last week we started interviewing Michael Hind about AI explainabilty. That term itself bears some, erm, explanation – all my spellcheckers flag it – and it's broadly about how to get an AI to explain why it made a certain decision, satisfactorily. The *satisfactorily* part is key. It's not hard to get an explanation of how a computer figured out the square root of 42, because we accept that square root functions are a thing, right, but getting an AI to explain why it denied someone a loan is not only hard, because it likely learned from a lot of training data, but it's also a much more demanding environment. That explanation had better be *good*. And maybe also have to stand up in court.

Michael is a Distinguished Research Staff Member in IBM's Research AI department in Yorktown Heights, New York. 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 is an ACM Distinguished Scientist and a member of IBM's Academy of Technology.

One of the terms Michael refers to in the interview is NIST, and that is the National Institute of Standards and Technology, and they are like the guardians of how we measure things accurately. All kinds of things. Started with things like how to weigh stuff accurately and for everyone to agree that they were weighing things the same way, say, by sending out precise reference weights that everyone could trust. So if NIST develops a standard for something, they do it so carefully that everyone who needs to refer to a standard for that can breathe a sigh of relief and just refer to the standard. So a sign that NIST is getting involved in an aspect of AI means that it's important that everyone agree on some definitions and metrics of that aspect.

Last week we defined explainability, talked about why it's important, and the relationship between bias and explainability and context and some of the directions this bleeding-edge work is starting to head in. Today we're going to talk about the Teaching Explainable Decisions project, some of Michael's experience with Watson, the difference between transparency and explainability and a lot more. Here we go.

You mentioned earlier, TED, what does that stand for?

Teaching Explainable Decisions.

That's a very accessible kind of explanation there. So, what does a day in the life of TED look like?

So, it is a lot of upsides. And there's one major downside, right. So, the one major downside is that subtle requirement that I said, when you have a training dataset, in addition to having, let's say, again, the answers to the loan application, and whether or not it should be approved, please also tell me Why. And I would argue that if you were making a data set from scratch, that shouldn't be too hard. It's kind of like grading an exam. And you've decided, you know, a question, oh, this is wrong, you put an X through it. And then just asking the grader to say, tell me Why. And you see this certainly with, you know, like English essays, for example. So, but the big but is, most people don't make datasets from scratch. They either get them from a third party, or they take their historical decisions. Maybe it may be loan, hiring, whatever it may be. And then they use those. So here was the information and here was the decision, and the person who made the decision is no longer with us or available, we can't ask them why, right? So that's the major downside of the technique is that you need to have the explanations in the training dataset. Okay, there's some caveats there, but let me just park it for there for a second. The upsides are a lot. One of them is the explanation that comes out should be relevant to you. Because it's the one you train the system on, right. So, if a loan officer is making explanations, it should make sense. You know, to them, it's not going to be some numbers, or some combination of the features, which really aren't accessible. Another aspect of it is you can have sort of higher-level concepts, right? So, take an example of let's think of a model to predict who might leave my company, because they're looking for better opportunities. So, attrition, right. And you may have this high-level concept of an employee that is disgruntled because they're not getting pay raises. And it may be the case that depending on which organization they're in, what's not a good pay raise may differ. So, if you're in a real high-flying organization than just getting an average pay raise, maybe it's bad, right? But if you're in one where there's not a lot of attrition, and maybe it's not so bad. And so, there's this high-level concept that I mentioned, of not being compensated well enough for appreciated well enough, and how it maps to the actual features will depend, depending on let's say, what organization? Okay, so that was a lot of words, to net it all out is, with this technique, you can specify higher level concepts, higher level explanations that don't necessarily directly map to the features. Like, you know, when I first started out, I said, you look at the features, and if the salary's too high or too low, then that's your answer. Here, it's these higher-level concepts that can be encoded. And so, it's getting closer to the way humans hopefully think about the space, as opposed to what the machine is churning away on.

So let me see if I, if I heard part of that right. You're saying that the training data can include explanations, and then we'll learn what an explanation is. Is that accurate as far as it goes?

Yes, yes.

And that's fascinating, because I can see how, given data of examples, supervised learning can learn how to reproduce the pattern that led from the question to the answer in each case, this is different; it irresistibly reminded me some for some reason of the IBM's cognition thrust with Watson, which is doing all kinds of things that went far beyond what I thought those pattern algorithms were able to do. So… well, how do you encode an explanation in that? It can't just be unstructured text.

Okay, so Yeah, great question. What does it look like? So my view is it can be it could be anything, right? It could be like, you know, the FICO credit scoring system has these reason codes or has numbers, you know, FICO thing that a lot of companies do have these codes. So, it could be a number, if that makes sense. But more often than not it, I think it will be text, it'll be text that that that you think is relevant for that explanation. You'll want to think about that text to maybe be consistent, right? If you have two different applicants, and they're being rejected for the same reason, it would be convenient to write the same text string. And then the next question, I think is, the way this system works is it just treats that text essentially as a number, right? So, it's not doing anything deep in the sense of understanding what that text is. So an example may be when we do prediction, in general, we're mapping the features to a label, or class, right. So, let's say in the case of loan, it was approve, or reject, that's two different classes. Approve is one class, reject is another. There are some problems called multi class classification problems, where there could be multiple things, you could be, you know, looking at a grade on an exam, maybe it's an A, B, C, D, E, F kind of thing would be one example. And so, what we actually do in the technique, that there's many ways of implementing the TED technique, the very simplest one, is we call it a Cartesian product, it's basically saying, you gave me an explanation, I don't care what it is, it could be a number could be a string, it could be a video, right? Whatever it is, I'm going to map that thing to a unique number, every explanation you give me, I'm going to number from one to whatever number of explanations you give me, okay, so it sits there on the side. And now I take that number, and I take the class of, let's say, approve, or reject. And I basically encode them together. So now I have, you know, reject. And I now have, let's say, 10 reasons to reject. So, I have reject, R1, R2, R3, R4, and so on. And then I actually take that, and now I have a new problem, a new machine learning problem, I've got my training data set, and that new label, which is R1, R2, R3, for reject reason, one, reject reason two. And I give that to any classifier, any algorithm to create a to create a classifier, and now it's a multi class problem, and then out comes up with a new model. And that what that new model will do is rather than spit out, approve or reject, it'll just spit out reject seven or reject number three, and then I go and look up what that maps to, and I give the answer. That's the simplest implementation.

Oh, okay. And that I get, so it's treating each reason, as part of an explanation as an atom.

Yes.

It's not going to be divided, it's going one or more of those are going to pop out the other end. In response to a question. Okay, that's comprehensible. Thank you, what I've been thinking of explainability, in the general sense is something that requires artificial general intelligence. But where you're helping me understand how we can carve off useful parts of it at the moment, talking about useful parts, which customer sectors most interested or engaged in this at the moment? Finance? Insurance? Can you characterize that?

You did a very good job. Yes, it's finance, insurance. Maybe health. Generally, it tends to be regulated industries. Places where if they mess up, there'll be consequences. But it's not limited to that we've had, you know, we've had interest from, you know, other places as well, retail, and so on.

So, were you in the sphere of the Watson project and what that turned into?

Yeah, I shared a building. Certainly, I wasn't part of the project. And I was also one of the early testers of the Jeopardy system and lost very badly.

And well, everyone lost very badly, I think; but I'm interested in its later career in medical diagnosis, and wherever that intersected with explainability.

I haven't worked with that part, not too much. I guess we had some discussions, but there's nothing significant there. Okay, but there is a related topic that we could branch into if you want, which is the bigger topic of transparency.

Yes, please.

So, in that case, transparency, what does that mean? It generally says, you know, people want to know a little bit more about a system and AI system, rather than just you know, deny me from a loan Why because maybe I don't even want to use it. I don't want to, you know, don't want to apply for this place. If it's using AI system, I want to know more about it. Right? So, you can think of this exists everywhere. It's not specific to AI; when you go to an online market, you're looking to buy something, you want to know more about it. You want to know reviews, you want to know features, you want to know price, so you want information about it. So, this started to be a pretty hot field, as you can imagine. And we in my group introduced an idea called the factsheet, which is a very simple idea. And it's not new to this, in general. It basically says, can you document for an AI system the relevant information for the target audience who's asking for it? So, if I'm a regulator, what does a regulator want to know? And that's the information you would record. Right? So, if I'm a consumer, now you can think of *Consumer Reports* where you're getting information, let's say about appliances, what do they want to know? How was it tested, things like that? And so, there's a lot of demand for this. I was just talking to a cabinet officer in the UK, who, you know, in the past, they've had some, they weren't unique, but they had some situations where an AI system was making decisions. This was related to exams, exam scores. And that there was some concern about what was going on there, was it right or wrong. And, and so there's a really strong desire to say, well, okay, what kind of information do citizens want? What kind of information can we make available, just for more for better transparency? So, so we've written a bunch of papers on that, and there is a rich website that gives examples and so on. But what was interesting in our journey there is going back to talking to customers, certainly, some customers cared about transparency. But they also cared about sort of documentation for their own internal purposes. Data scientists are brilliant at creating models. They're not necessarily brilliant at documenting what they do. And like most humans, and also there's a lot of turnover, and there was a question of how we can actually capture what happens. And so, in some cases, particularly like in finance, there's already elaborate processes, to write documents, sometimes over 100 pages long, to please model validator and regulations, and so on. But these are very, very painful to do, because they're all done by hand. And so, the thing that we've been working on very closely now with customers in our product division of IBM, is how can you automate this process? How can you, as you're building AI, and even thinking about building AI, how can you capture the breadcrumbs so that you can produce a document at the end that people can learn from? Right? So, it's, it's if you think of the whole pipeline of building AI, it's like software development, there's a process, there's many players involved, and they're all making decisions, and measuring things, and producing things. And if you can capture that information, then you have sort of the resumé or something of the model, what is, you know, what's their story? And how that can be used for many different things.

I'd like to drill down more into the definition of transparency, because I've been thinking it was what's the architecture of this, this thing, and that people would not be satisfied by being told well, it's a BERT transformer with 150,000 parameters. That's of interest to another computer scientists, but probably not to whoever wants to know. And so you're describing something that's a bit more abstract, for me, I don't have a good handle on it yet. What is the goal? What itch is transparency trying to scratch?

Yeah, so I'll make an analogy that I think will turn on some light bulbs, and then I'll tell you why it's not perfect. So, the analogy is a nutritional label. Right? So, for a long time, consumer goods were produced and different brands were using terms that weren't necessarily meaningful. And so, in the US, they actually took about 20 years to figure out what kind of information do consumers want? How can we make it consumable, understandable, and then require manufacturers to produce that information? Right. And so, you could think of this definition of AI transparency or factsheets as sort of like a consumer label for AI. What would that be? That's the hopefully light bulbs go off there.

So that suggests to me – and this is probably carrying it too far that you can deconstruct AI into components that are as meaningful as vitamin C and saturated fat, and a small enough number to fit on a label is that am I going in the right direction there?

Yes, so far, until the caveat comes. So yes, so the end those components would be things like bias, have you measured bias? What's the value and what protected groups did you look at? Did you measure the ability to for adversaries to attack the system? And what was the score on that? Privacy leakage and so on; there's a bunch of these dimensions. That would be probably relevant to most people.

Although if I'm being picky, with a computer science hat on, that doesn't sound like a description of the components. That sounds like a description of the user acceptance testing.

Right? Yes, yeah, actually, this is an important point. So, in this view of transparency, transparency doesn't mean explaining how, how chocolate chip cookies were actually made, like you think about you think about a nutritional label, it doesn't tell you the recipe of how they made it, it tells you information that should be relevant to your consumption of that. So, I would say here, it's your consumption of AI, what do you want to know about it? You probably don't even understand how it was made? Because, you know, a lot of people don't understand that. But what's relevant to you? You want to know, for people like me, is it fair, for example?

So, is it like, Good Housekeeping Seal of Approval, that we have tested this thing? And it fits this standard?

Yes, you’re going in the right direction. Absolutely. We’re far away from that. Ideally, it would be nice to have someone say, or third party whoever say, you must test the following things. And then hopefully, you get you know, some level of consumable label, thumbs up thumbs down, or more details or so on.

So, then you want standards that are third party and, and not per vendor. And you mentioned NIST earlier? Is NIST working on standards like this? Or do those standards already exist in some international body?

There’s a strong desire for this, and people are working on this. We’re working closely with NIST on this. The EU is coming out with new regulations on AI. And they talk about this, and they gave examples. And here’s where the caveat comes in. Right now, we don’t know exactly what should be measured, right. So, I said bias before. And I probably got away with saying that. But in reality, there’s many different ways of measuring bias. And some of them make more sense, depending on your use case. So, it’s not like with a car, you measure the speed of the car; we don’t yet have the agreed upon metrics on what makes sense in this case. So, this is where, right now, the food label analogy kind of breaks down. Because we don’t know the calories and the sodium content, we don’t have the equivalent just yet of what they should be. We have ideas, and there’s been proposals. But for now, what we are saying is, it should be customized to whoever decides this, or in the case of you know, we haven’t figured it out yet, maybe it’s for a particular company, what do you want to see? What are you concerned about? Maybe your model is, is deployed behind a firewall, and you don’t care about someone coming in trying to attack it, right? Or maybe your model is doing prediction on manufacturing defects. And so, there’s no humans involved there, so maybe bias is not as much of a concern, human bias, not much, and so on. So, the idea is, it’s customized, it’s sort of like a template where you figure out what information you need, and then you instrument your AI lifecycle to collect that information, and then produce them. And we expect - hopefully, in a steady state, there won’t be a lot of customization needed. But at least right now, we’ve seen a large variety with companies wanting different kinds of information.

You mentioned the EU there, and they came out with a proposal in June, that I haven’t read yet. But its laying title is “Laying down harmonized rules on artificial intelligence.” I’m wondering which players and whether it’s them, or NIST, or anyone else, are capable of, and have it in their wheelhouse to, construct these standards, or is it something that’s going to emerge from industry and some consortium of IBM, Google and so forth?

Right, so both are happening. So, the EU, for example, I think it was two years ago, they got together a whole bunch of experts that had this higher-level expert group, they called it. And they looked exactly at this problem. And they came up with a draft proposal, which was a set of over 100 questions that they were recommending an AI developer or an AI company would have to answer about a system. And then they opened it up and they said, “Okay, here’s what we came up with.” And it opened up to any EU company to give them comments and feedback, right? Because there’s two sides of this coin. Right. There’s the one that I think is the most, you know, it’s probably most important is what rights should society have and citizens have to be protected and so on and so forth. And that's where you get the over 100 questions. The flip side is what's it going to take to fill out those answers? Right? So, you need the companies to figure out, first of all, can I even answer some of these questions? Hopefully they can. And then two is, how much of a burden? And you might say, okay, ignore that, you know, people have fundamental rights and burden, they should just deal with it. So, they got, I think, several 100 companies replying to that draft, and so iterating on that. So, the state of the art right now is very active research in the space. Maybe research isn't the right term; very active development in the space trying to figure things out, and groups coming together. There's an organization called Partnership on AI, that was bringing together basically, worked done at IBM and Google and Microsoft and others that have had similar ideas, like the factsheet idea, and trying to say, is there some consensus, and our view at IBM at least has been that we don't think that in the near term, there's going to be one magic set of questions, right. There may be a magic minimal set of questions. But then particularly use cases are going to want to add even more set of questions. And so, it's still a very active area to figure out what exactly are those two sets.

What about academia? This is the sort of thing where it sounds like it's at the cutting edge of computer science research. So, it is the government throwing money at someone in Stanford or Carnegie Mellon or other places to put a bunch of postdocs on this?

Yeah, it's certainly an active area, there's a major conference called fairness, accountability and transparency. It's an ACM conference that meets every year. And there's a lot of participation from academia there.

The FAT conference, that's going to be easy to remember.

Yeah, they changed its name, FACT. Yes, for good reasons.

That was amazing stuff we're talking about here; our time is drawing to an end. But it is such a rich field, what do you see yourself doing in this space five or 10 years from now.

So, one of the things I wanted to mention is, earlier, I talked about a gap and a gap in many areas, but one in explainability. So, where you know, the scientists were doing good work on a particular problem in society cared about, you know, a variant of the problem, you know, explainability for society, as opposed to explainability for data scientists. What we've done, my group and a bunch of dozens of other researchers with me, have put out a bunch of open-source toolkits in a lot of different spaces. So, we focus on explainability. So, there's a toolkit called *AI Explainability 360*. that's out there. It's been out there for two years now. And it has a lot of these algorithms. And the idea and motivation were, let's put them out there, so that people can play with them and get feedback. TED, for example, is there, the code is available, people can play with it, and get feedback and understand what you know, what works, what doesn't work, which, you know, so on. And we've done similar things with fairness called *AI Fairness 360*, another topic called uncertainty quantification, which is talking about sort of the risk level of a model, one on privacy, one on adversarial robustness, and so on. So, for those who are interested in going deeper, those are all available, you can find them links on my webpage, or just google them.

We will put links in there. And the transparency one talked about factsheets; can you explain fact sheets briefly?

So yeah, so fact sheets are what we were talking about before, it's the sort of this transparent document, it is a document for transparency, it collects facts about how a model was constructed. And things like what was the reason? What's the intended purpose of this model? Why did you do this? And that's important, because often models are reused, and they may not be appropriate to be to be reused. So, a concrete example: I develop a loan approval for a rural area in the United States. And one of the things I use in my model is, you know, does a person own a house? does the person want a car? Those are probably good predictors of credit worthiness, right? And then a friend of mine down the hall says, hey, your model works really well. I want to use it in Boston, Massachusetts. Okay, fine, whatever. But it turns out in Boston, Massachusetts, many people don't own houses or cars, but they may be creditworthy. And so, it's trying to capture that kind of use case of, let's clearly say, what's the intended purpose of this. And then, you know, help reuse things appropriately and not have problems. And then other things as well, it captures.

Kind of like establishing the pedigree for a product in the way that aircraft have to trace the pedigree of all the parts that are in it back to the manufacturer.

Exactly, yes. And things like the training data set and, and information about the training data set. A lot of it is a lot of the information that's used during the whole process, then you can decide which of this information is relevant to your audience, right. So, if it's you know, the data scientist, they probably will eat it all up. If it's your consumer, maybe they don't necessarily care about some of the details. And so the website, does it. It's called Factsheets 360. It actually has examples of fact sheets and ways of sort of tailoring and looking at subsets of information and so on.

So, if you could put on prediction hat and imagine where this field will be in 10 years, what would it look like?

So, I think we'll have in the transparency part, we'll have a better idea of what kind of information should be required with every model? I think we'll see regulation; I think a combination of, you know, the question you asked before between industry, and regulations, we'll figure out what's the appropriate level of documentation that's required for transparency. The other aspect is this topic of AI governance. And what I mean by that is governance of the lifecycle, the construction of AI. A natural follow on from what I was saying is, if I'm collecting the breadcrumbs during the process, then maybe I can actually have some rules that say, Hey, if you measured bias at this point in the lifecycle, and it was not good, don't go any further, don't allow the system to go any further. So just like you have with traditional software, you know, testing and different checkpoints, before it can go any further, you can have the same kind of thing here with AI, but not just for accuracy, but for these other attributes as well.

Hmm. Kind of like going to that aircraft analogy. It's all the regulations about testing for metal fatigue, and so forth, that yes, make them safe enough. And really, that level of safety industry around AI or even computers in general is very, shall we say, embryonic. So let's say there's someone listening to this, who's learning about this, maybe in university, and they're excited because of what you've described, it's a fascinating field, and also, they can imagine that it would employ them for the rest of their life, then, where would they start? Should they send you their resume? What would you suggest?

So, the best place to best thing to do is, you know, figure out which of the many little sub areas we talked about, they want to focus on, you know, explainability, transparency, privacy, whatever it is. And pretty much for each of those, we have the website and the toolbox. And each of those also have Slack channels as well, a community where people can come and ask questions and interact, not with me, but also other people as well. So, for example, the Fairness 360 toolkit has over a thousand people subscribed to this Slack channel, which are that are engaging. And some of these are students, I'm sure, I know, there are classes that are using toolkits. But there's some of them are also companies, people, practitioners, who are using it as well. So, it's a great opportunity, not only to talk to me, a fellow researcher, but also opportunity to interact with the pragmatics of this, which can really, I think, inform people's solutions.

Well, that's great to have that level of openness and accessibility. We'll put up a link to your page. Is there anything else you want to tell our audience in terms of, do you want to issue some kind of assurance that we've got a good handle on this? They're feeling nervous about this, the state of AI being adopted for things, and maybe that it's too much Wild West, can you tell them that we're going the right direction?

Sure, I guess what I would say is speaking to many, many companies, and their usage of AI is more in the space that we're talking about that there with having a human assistant with them. And so, I think because of a lot of great work by others who have raised the issue of what can happen here, and of course, you know, bad publicity and brand reputation, that's happened to some companies who have messed up, there's a huge awareness to make sure that they don't rush into anything and they think things through and so on. So, it's not a guarantee that certainly things won't happen. But the fact that I'm speaking to dozens and dozens of companies, I'm a researcher, I'm not a product person means that they really want to know and really want to do the right thing and educate the uneducated.

Great. Well, it's terrific hearing from someone who's on the pointy edge of something so important to our adoption of AI safely. So, Michael Hind, thanks for coming on the show.

Thank you. It's a pleasure.

That's the end of the interview. Did we… explain… explainability well enough for you?

I was fascinated by the idea that explanations are a thing that you can put in the training data for an AI so it actually learns how to give them along with the problem it's learning to solve. I know that's a gross simplification, and it's obviously limited in how far it can go, but still, it blew my mind as an answer to the problem, opened up a whole new line of thinking for me.

Maybe the next frontier – and I only just thought of this, or I would have brought it up in the interview – is in getting a decision-making AI to ask for more data where it would make a difference. After all, we can imagine a human loan officer being on the fence about a decision and asking for some data or being influenced by some data that wasn't on the form, that tilts their decision. On the one hand that's very human and who would want to give that up; on the other hand it's potentially unmanageable and an auditor or a general counsel would have a hard time with it if it came to either of them. So whether we train an AI to do that might have less to do with whether it's possible and more to do with whether we want to do it.

I also found Michael's analogy of a food label to explain transparency really illuminating. Stealing that one, okay, Michael?

Anyway, I came away from this more hopeful about the future of explainability in AI. Going to be keeping a close eye on this.

In today's news ripped from the headlines about AI, you know, we've talked about AI in law before on this program, like in the interview with Ted Parson of the UCLA Law School we discussed what if AI replaced a judge or jury? So, people are working on this. Back in 2016, a team of British and American researchers created an AI system that could predict rulings by human judges by reading the transcriptions of real-life cases from the European Court of Human Rights.

The AI was trained with 584 court rulings by extracting the relevant sections, and then it was used to make a prediction regarding the judgment, and it was 79% accurate on average, which is pretty good.

So in December last year, the Ukraine introduced AI to its justice system, and in February this year, their High Council of Justice approved an electronic court where AI would be used for legal advice and the resolution of disputes that are less complex.

The Ukraine Ministry of Justice also intends to use AI software, called Casandra, as a decision support tool in criminal justice to assess the potential risk of recidivism and also help judges in making custodial decisions.

We want to be circumspect how we talk about this, because otherwise people will start talking about robo judges and the cartoonists will get going and common sense will flee the building. This is a very limited application in a very limited context. We're not going to see a computer put on a black hat and sentence someone to death for an unpaid parking ticket because there was a bug. I know that laws in some countries are referred to as the Criminal Code or the Civil Code and so there's that temptation to think that they can be turned into computer code, surely they're supposed to be unambiguous and rational and fit together like subroutines in a program, right? Well, lawyers aren't computer programmers, so, no, it doesn't work that well yet. And don't get me started on the Tax Code.

But nevertheless, this is a very interesting development. Let's see what happens.

Next week, I'll be talking with Alexandra Mousavizadeh, who has launched several indexes of great interest to us, which rank countries and companies according to objective standards. She's founded the Global AI Index, the Responsibility 100 Index, and the Global Disinformation Index. You can imagine how much we'll have to talk about. 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