Rashmi Mohan:

This is ACM ByteCast, a podcast series from the Association for Computing Machinery, the world's largest educational and scientific computing society. We talk to researchers, practitioners, and innovators who are at the intersection of computing research and practice. They share their experiences, the lessons they've learned, and their own visions for the future of computing. I am your host, Rashmi Mohan.

Artificial intelligence is rapidly becoming one of the most powerful forces shaping modern life, influencing who gets hired, who receives medical care, who gets loans, and even just how justice is delivered. But what happens when the systems making these decisions can't explain themselves? Today's guest has spent her career asking that question and challenging one of the biggest assumptions in modern AI: that accuracy must come at the cost of transparency.

Cynthia Rudin is the Gilbert, Louis, and Edward Lehrman Distinguished Professor of Computer Science at Duke University. She's a pioneer in interpretable machine learning and the director of the Interpretable Machine Learning Lab at Duke. Her path-breaking work confronts the industry's long-held belief that transparency and performance cannot coexist. Among her many accolades, she's the recipient of the esteemed 2022 Squirrel AI Award for Artificial Intelligence for the Benefit of Humanity. Her work has won her countless awards and celebrated papers, making her a very coveted speaker and a distinguished member at most conferences and committees. It is our honor to welcome you, Cynthia, to ACM ByteCast.

Cynthia Rudin:

Thanks for inviting me.

Rashmi Mohan:

Of course. Yeah, I'd love to hear from you a question that I ask all my guests, which is if you could please introduce yourself and talk about what you currently do as well as give us some insight into what drew you into computer science in the first place.

Cynthia Rudin:

So my name is Cynthia Rudin. I work in interpretable machine learning. There's the introduction kit. So I'm a professor at Duke University and I've been working in interpretable modeling for a really long time. I started working in this area when I was working on a project with Con Edison in New York City. I was employed at Columbia. There was a group there that wanted to completely change the way that power grids are maintained and they wanted to do it with machine learning. And so I said, "Okay, I've got all these tools I learned in graduate school. I know how to do machine learning. I'm going to take all this data, I'm going to throw it into these machine learning models and it's going to tell me which manholes to inspect to figure out where there are problems in the power grid." We were trying to predict fires and explosions and smoking manholes and things like that.

The machine learning algorithms did not help. They were just a disaster mass. They were telling us to go look at manholes where there was almost nothing wrong with them, just because they were near other manholes that had problems. So while we were troubleshooting those, I realized that if we could actually show the machine learning models to the power engineers, they could help us troubleshoot them and they did. They helped us troubleshoot the data and the algorithms on the models. From there, we were able to get much more accuracy than we could ever get with the black boxes. The interpretability made all the difference in the world in that case. And that's how I realized, hey, there's a

different path for success in machine learning than just throwing black boxes at everything because it doesn't work.

Rashmi Mohan:

That's super accurate. I mean, I think the one thing that I understand from what you said also, Cynthia, is in these domains where you're working with people who are not... I mean, they're experts in their own areas, but not necessarily computer scientists, it's so critical for you to be able to explain how you're making these decisions and for them to be able to provide you input to refine your model.

Cynthia Rudin:

Yeah. Especially when you're working with really noisy data, like data that's not perfect. Like a car. A car is a giant black box. We don't understand exactly how everything works, but we don't need to because we know that there's physics behind it. We know that this thing, every time we press this button, it does this. But with data, if you're building a model from data, well, that gets really noisy because you can't trust a lot of these databases. They're not trustworthy. And so when you're making decisions based on them, especially when you're involving domain experts, it's really helpful to be able to understand what's going on. What's in the data? How is it translating into a decision?

Rashmi Mohan:

That makes a lot of sense. It was a similar process that, you know, I was doing a little bit of research in your previous work. The building of Series Finder, the work around crime series detection.

Cynthia Rudin:

I was a faculty member at MIT before I was at Duke and I was sitting in a room and there were a bunch of police officers from Cambridge Police Department that had invited themselves to meet with us. It was actually a really funny story. So they walked in there with their uniforms and their guns and here's all us professors sitting at the table cowering and they said, "Look, we need your help with something." And then one by one they talked to the professors around the table and the professors were like, "Well, I don't want to actually... I'll supervise, but I don't want to actually collaborate." And I was like, "I'll collaborate with you guys." And I was like, definitely the youngest one there.

And they actually had a really interesting problem, which was they wanted to figure out which crimes were committed by the same individuals. Because they wanted to figure out crime series. So a crime series is a set of crimes that are committed by the same individual or set of individuals acting in concert. They were thinking of things like housebreaks in Cambridge, where you have these groups of people who go and rob houses in Cambridge and they do it while people aren't home and they break into houses the same way. They have a modus operandi. They pick a certain neighborhood and have a certain way of doing things. And if you can figure out the modus operandi, then you can figure out which set of crimes are committed by the same people. And so it turns out to be a really interesting subspace clustering problem because the modus operandi is actually the subspace and the clusters are the crime series.

And so if you could figure out at the same time what's the modus operandi and what's the crime series, what's the subspace and what's the clusters, then you can help solve crime because they can't do anything about a crime series unless they know that it exists. So me and my student, Tong Wong, and this is with detectives, Rich Sevieri and Dan Wagner was the one leading the whole thing. He was the detective who came up with the idea. We worked on this piece of code and we made it public, which is the series finder algorithm. And then we found out years later that the NYPD, they had been asking us

questions. They had a new data analytics team and the person there was asking us questions about how you implement this thing and we were like, "Okay, sure, we'll help you." And then it turns out they had implemented it at the NYPD since 2016. And so it figures out whether a new crime is related to past crimes so they can find series much more easily.

Rashmi Mohan:

That's amazing. I mean, these are such impactful projects. Do these seek you out or do you actually go looking for them?

Cynthia Rudin:

It's some of both. I mean, the crime series one just fell into my lap. Like I said, I was in a room and police officers were like, "Can we work with you?" And I was like, "Yes." But it doesn't always work like that. For example, I have a long-term collaboration with some neurologists at Massachusetts General Hospital, so in particular, Brandon Westover. And that started because my student, Berk Ustun, was looking for a collaborator. And so he was like, "There's this guy at MGH, he's teaching this class. It looks like right up our alley, maybe we could contact them." And so we sent them an email and I'm still working with them and it's over a decade later. So yeah, sometimes it happens that way.

I have this other problem that sometimes I form a relationship with someone and then I can't get out of it because they still want to keep working together and then I'm like, "I have too many collaborations." So it's really hard to end things. So I'm now really careful about what I start because then I know I can't end it. But I've been loving working with Brandon Westover all these years. It's been great.

Rashmi Mohan:

That's incredible. Thank you for sharing that. I know we dove straight into your projects, Cynthia, but I'd love to sort of go back even further because oftentimes, I mean, the audience that we have are folks that are early in their career, could be considering computer science, folks that are in college. I'm curious as to what drove you into computer science in the first place. Were you exposed to it early on?

Cynthia Rudin:

I was an applied mathematician. I wanted to do applied math modeling. I think my earliest goal was to try to run one of these giant weather models. I'm so glad I didn't do that because that's definitely not the right thing for me. I found out about machine learning because Gary Flake, he was a really important influence on my life. He was a researcher at NEC Labs and he handed me a book and I read the book and I was like, "Oh, you can predict the future with data." So that was kind of my first introduction to statistics and computer science was this book. That's how I found out about machine learning. So I read this book and I was like, "That's what I want to do. I want to predict the future with data." So yeah, it wasn't hot at the time. It was not like it is now. It's hard to believe the difference between what it was then versus what it is now. Now you go to a big conference, like a machine learning conference and there's 16,00 people before it was maybe 200 people-

Rashmi Mohan:

Wow.

Cynthia Rudin:

... and you knew a lot of the people. So it was really different. It was just like a nothing field. I mean, nobody knew about it. My husband was in biology and everybody was doing genetics and biology. That was the cool thing at the time. That was everybody was like, "Oh, this is the wave of a future. They're going to change the whole world." And now machine learning's taken over as being the really important thing like AI. So I guess it just goes to show you that you don't have to go into the thing that's the most popular. Go into the thing that you really like to do, think for yourself and don't just follow the crowd. Yeah. I mean, I've never followed the crowd, right? Working in interpretability when everybody thought black boxes were the way of the future, that's totally different than what other people were doing.

Rashmi Mohan:

I think you hit the nail on the head, which is really, I mean, I can tell from just the way you're describing your work, how passionate you are about it. And to your point, I think pursuing what you're most interested in, because it's so hard to predict how these fields transform over a period of time and what becomes relevant and critical and important. Can you describe to our audience what exactly is an interpretable ML model and how do you contrast it to a black box model?

Cynthia Rudin:

Let me elaborate a little bit on the previous question before I do that. I don't want the young people who are listening to think that all of this was easy and that I just made up my mind that I was going to do this when I was really young. That's not true at all. I went through many years where I was like, "This field just doesn't make any sense. I don't know why I'm in this field." I was pretty depressed and it was pretty upsetting because I was like, "These black boxes are not working for me." So it's definitely hard. It just takes a lot of work and a lot of thinking to get to the other end of it. Just can't give up.

Okay, back to things I actually really love, which is interpretable models. An interpretable model is a model that a person can understand. Interpretable machine learning is actually a pretty big field. You can actually work in a lot of different areas of interpretable machine learning. The field was named at least as far back as 2001 in the Breiman two cultures paper, because he used the word interpretability in that paper. A lot of people have co-opted that word to mean explaining black boxes. That is not interpretability, that is explainability, that is explaining black boxes. And the terminology there is really important because you do not want people to get confused between a model that's actually interpretable that can be used in high stakes decisions and a black box which can't and that you're just explaining it. You're just poking at it to see what it might have in it. So it's really important to keep that terminology straight and no matter how many times I say that, the explainability people ignore me. They want to use this term that's used for important high stakes decisions.

Now, interpretable models can be a lot of different things. For tabular data, so data that can fit on a spreadsheet, I like to say that the models can be fit on an index card. You can just write them down like a little simple formula that can fit on an index card or a PowerPoint slide or maybe a piece of paper. These can be like medical scoring systems that you might have calculated for you in a hospital. You give them a little bit of your medical history and they give you two points for this, three points for that, four points for this. And medical scoring systems are something I've worked extensively on and the collaboration I was mentioning earlier with Berk Ustun and Brandon Westover and also another neurologist Aaron Struck led to a scoring system that's widely used in ICUs, intensive care units, now for predicting seizures in critically ill patients. And the model we designed is called the 2HELPS2B score and it has just a few things that the neurologists look at in the patient's EEG and they calculate the score and that score determines the likelihood that the person is going to have a seizure and they use that

information to try to figure out how to treat the patients and whether to move EEG equipment around and so on. These are like tiny little models that you could memorize potentially, like a little formula.

If you're working on different types of data than tabular data, like you could be working on say images, then my lab works heavily on interpretable neural networks for computer vision. And so these models, for the most part, they use case-based reasoning. So case-based reasoning is a kind of like this looks like that type logic. When you're analyzing a new image, you would compare parts of it to other images that you've seen before because they're in the training set. So you could say, "Well, okay, I got to classify this image of a bird. What kind of bird is this?" "Well, the head of this bird looks like the head of that bird. I know what that bird is. That's a clay-colored sparrow. The belly of that bird looks like the belly of this other bird, this is the same texture on the belly, things like that. So these models are called ProtoPNets. They've become really, really a very popular type of interpretable neural network for computer vision. Interpretable machine learning is really broad. So you can work on all different model classes.

We work on GAM, so generalized additive models. We also work on decision trees. We work very extensively on sparse decision trees, which have if then rule-based logic. We also work a lot on visualizing high-dimensional data using dimension reduction. So taking a high-dimensional data set and projecting it down to two dimensions so that you can get a bird's eye view of what's in it. You could see all the clusters and you could see all the manifolds and how they connect to each other. And so that's another area that we work in to try to understand high-dimensional data in an interpretable way, in a nice way. Yeah. So that's just some examples of what the field is. We've written some review papers that kind of talk about it. It's basically models that are constrained, that actually have constraints so that humans can better understand what they're doing.

Rashmi Mohan:

Thank you for that explanation. That was very detailed and it really helped. One question, are there a certain set of applications or particular scenarios where interpretable models are more useful? Are they broadly applicable across most use cases that you'd imagine a layperson using? I know you mentioned high stakes earlier, so I was wondering if you could qualify that a little bit.

Cynthia Rudin:

Yeah. So interpretable models are really useful when you need to troubleshoot. If you don't need to troubleshoot, then maybe they're not that useful. If the model's 100% accurate, then maybe you don't need to troubleshoot it so you don't really need it to be interpretable. It depends on the situation there. What my lab has found is that when the data are somewhat noisy, meaning that there's a non-deterministic relationship between X and Y, then interpretable models tend to be very competitive with the black boxes in terms of accuracy. In other words, things like recidivism prediction, like criminal recidivism prediction. You're trying to predict whether somebody is going to be arrested for a crime of a certain type within a few years of when they're released from prison. That's something you can't predict very well in advance because you don't know what's going to happen to that person in the next three years. And so for those kinds of problems, black boxes tend to do just about as well as interpretable models. So those are cases where there's no reason to use a black box because these are high stakes decisions, like they're decisions about people's freedom. And so you don't really want to leave those to a black box anyway. So I would say non-deterministic cases, high stakes cases like that. Interpretable models are really good because you can troubleshoot them and they're just as accurate as black box models.

Rashmi Mohan:

Got it. Okay. And in general, do you get a lot of questions? What is the biggest myth you get in terms of trade-offs between these two?

Cynthia Rudin:

The biggest myth is the one you just mentioned. So there are a lot of people who are still wedded to the idea that when you add more complexity, you get more accuracy. And for a lot of these problems, the data sets just don't admit more accurate solutions when you add more complexity, they just overfit. And so people, they really don't like that idea because a lot of machine learning is... I mean, this is my theory, that a lot of machine learning is founded off the idea that you're working with super clean data sets. For instance, if you're doing image recognition, you don't really have as much noise in the data for some of these problems. If you have 10 identical images, then either they all have a chair in them or they don't. Whereas if you have 10 medical patients, they have the same medical record. Then it's possible that half of them can have a stroke next year and the other half might not. So these are like very different problems. And so people try to use the mentality of these very clean data sets on realistic data sets and it just doesn't work. It just doesn't apply. And so you have to think about the statistical considerations when you're talking about interpretability and a lot of people just can't do that. They just say, no, black boxes are more accurate no matter what. That's where it's really hard to change people's minds.

Rashmi Mohan:

Is it fair to say that when you work with interpretable machine learning models, the domain experts that you're working with, one, have a better way of providing feedback. Is that necessarily only to computer scientists that are working with that model or is that passed on all the way to what I would call the end consumer or customer?

Cynthia Rudin:

So the domain experts I work with are neurologists and radiologists and power engineers and police officers. So they're the ones who need to understand the model. And if they can use that model, if they can do better, it definitely gets passed onto the end.

Rashmi Mohan:

Yeah. No, fair enough. That makes a lot of sense. How should we think about accountability in these situations, Cynthia? When an algorithm makes a decision, these are critical decisions that are being made as you explained.

Cynthia Rudin:

Yeah. So I'm working on decisions that are high stakes and they're generally made by people. You don't really want to outsource a lot of these high stakes decisions to AI unless there's serious time pressure and a human couldn't do nearly as well in that situation. But the cases I'm working on are cases where you're assisting a human. You assisting a radiologist to analyze an image or you're assisting a neurologist to make a decision about a patient. So these are high stakes decisions and they're aids for humans. And so the accountability rests with the human and we're just trying to help them.

Rashmi Mohan:

Understood. Yeah. Are these models harder to train? Are they more cumbersome or is it more expensive to train them?

Cynthia Rudin:

I'm not sure because it's hard to weigh the different costs of training these things. So for us, there's a lot of troubleshooting of data. There's a lot of algorithmic development. There's a lot of talking to domain experts. Those are things that a lot of black box developers don't have to deal with, but they have other challenges. They have to obtain very large data sets. Ideally, they should obtain them legally. They have to deal with a lot of hardware. I mean, we have to deal with hardware stuff too, but not nearly what they have to deal with. So I think the challenges are just different. For example, you really can't compare us building a radiology model to OpenAI building ChatGPT or something like that. They're just very different problems, different goals, different data.

Rashmi Mohan:

I know in some of your other work, you also talk about allowing users to interact with models and be engaged in picking the right models for the use cases. Could you tell us more about why that is significant?

Cynthia Rudin:

Yeah. So this is actually something I'm really excited about. When we first developed the 2HELPS2B score with Brandon Westover and Berk Ustun and Aaron Struck, Berk printed out like a hundred models that were all about equally good for the data. He just literally handed Brandon and Aaron this package of models and they had to look through those models and figure out which one of these is going to be the one that we're going to use on the patients. This involved not just the data because the data was limited, but it also involved their domain expertise. So what they knew about the important variables and stuff. Because there were so many equally good models, we could give them a big choice, what to choose from, but giving them a giant packet of paper is not, that's one way to do it, but it's not really ideal. And to be honest, before what people normally do, they don't even hand the domain experts a packet of paper. They just hand them one model because that's what machine learning algorithms return. They just return one model at a time and that's not great. If you hand a domain expert one interpretable model, they will find problems with it. They're not going to want to use it. They say, "Oh, this model is not good." And you say, "Why not?" And then they try to describe to you what it is and you're like, "Okay." And so you try to reformulate the problem and this is just a giant mess. And so we call this the interaction bottleneck. So it's the bottleneck that's the interaction with users.

And so to try to avoid that, we developed a new paradigm for machine learning that doesn't just return one model at a time, it just returns all the good models all at once. And so you got maybe several million models or something like that that you're storing and then you have to provide that to the domain experts, but you've got to provide it in a way that they can look through those models so that it eases the interaction bottleneck. And so we worked with human computer interaction experts to develop interfaces to these models and those interfaces are what the domain experts can use to search through all of these good models to find one that doesn't just agree with the data but also agrees with their domain expertise. And so we have these beautiful visual tools that kind of index the Rashomon set, the set of these good models so that they can pick models from there. So that's what we've been trying to do. It's a lot of fun. It's really rewarding and I think it's going to totally change the way that domain experts interact with machine learning, scientists and algorithms and data.

Rashmi Mohan:

ACM ByteCast is available on Apple Podcasts, Google Podcast, PodBean, Spotify, Stitcher, and TuneIn. If you're enjoying this episode, please subscribe and leave us a review on your favorite platform.

Yeah. I mean, it's an incredibly powerful way to, one, give more agency to the domain experts to be able to pick the model. I'm sure that to your point earlier that rather than poking holes at one model, they now have the ability to decide between multiple and determine which works best. When you were talking about visualize, do each of them come up with their own metrics for how they decide which is the most applicable model and does the tool that helps them visualize these models, how does that adapt to their decision making?

Cynthia Rudin:

The answer to that question is not fully determined yet because we're still throwing this out there. We've done some user studies with domain experts, but the answer to your question hasn't been completely resolved and I think different people work differently, but we want to make it look a little bit like an encyclopedia so that people can look up models of this kind and then dig down into those models and then look up models of a different kind and dig down into those. And so what we envision is that people will go, "Oh right, we really don't want models that look anything like this or this or this or this or this." And then there's a whole portion of the model space that they've excluded and so they've narrowed it down a lot. And so that's what we're hoping is that they can really help us narrow down the model space quite a lot and then deciding between those models, well, that's either something they can do themselves or we can help them with it. They can say, "Oh, you need more data about this." And then we can say, "Oh, okay." And then we can build that. That's the ideal is that they can reduce the hypothesis space tremendously just themselves just by looking at this thing.

Rashmi Mohan:

Got it. Yeah, that makes sense. A lot of these domain experts that you speak of come from fields outside of computer science, but they seem to be very open. They understand the value of using machine learning to solve some of these problems. Has there ever been a time given that you've been working in this field for a while where there was more skepticism?

Cynthia Rudin:

I think there's a lot of skepticism in general and there should be because you've got a ton of people selling explanations of black boxes and you should really be skeptical of those because like I said, they're just like a poke at the black box and these guys are selling these things as actually interpretable and you can't use those for high stakes decisions. I'm actually happy with the skepticism. I admit the people who select me as their collaborators and the people that I like to work with, they're people who know what I'm doing. Like Brandon Westover, the neurologist, he's trained his own machine learning models. He's got this team that trains neural networks now and my radiologist colleagues know exactly what they're doing. They also train their own models and I work with people who are experts in heart monitoring. They also have their own teams, now, that train machine learning models.

So I work with a lot of domain experts who are very well-educated and they know what they're doing. The power engineers that we work with even had statisticians on their team. In terms of the people that I work with, it's often people who are quite educated. And the police officers that I worked with, they didn't even have... Dan at the time didn't have a college degree. I mean, now he has a master's degree from Harvard, but he at the time didn't have a college degree, but he was able to try to read scientific papers. He was just an unusual guy. He was just super, super smart. So I think it's part of it self-selecting,

but that's okay because if we can develop tools that other people can use, even if they're not machine learning experts, that's fine. You see what I'm saying?

Rashmi Mohan:

Yeah, no, absolutely. And there's clearly, I mean, some folks that you've worked with are who are just exceptional in the ability to spot the opportunity and then come seek you out to actually work on these problems. Thank you so much for sharing that.

I'm going to pivot a little bit to talk about ethics in AI. So there I know is another area of interest for you. What does that mean to you? What does trustworthy AI mean in practice?

Cynthia Rudin:

Trustworthy AI I think is a huge field and it encompasses a lot of different things and I think it does encompass interpretability because you don't really want to be using a black box for these high stakes decisions if you don't need to. If you can use an equally accurate interpretable model that you can troubleshoot, then you should be doing that. It's such a broad field that it's hard to, trustworthy AI encompasses the data and its provenance and the code and it's quite broad. I like to sit in my narrow little corner of it just because it's a huge thing. Actually, let me take that back. My corner of it is not narrow and tiny. My corner of it's important and people don't understand how essential interpretability is to trustworthy AI. This is something that I've had problems with the last 20 years. I mean, like I said, people used to... I used to give talks and people were very skeptical and sometimes people would come up to me afterward and they would start yelling at me. "Why do we need this? We don't need to see what's inside the black box. We just want it to work. We want it to be more accurate." And I would say, "But my models are accurate. They're just as accurate as the black box models and now I know why they work." I think people just didn't really understand how essential this is.

Rashmi Mohan:

What was the hesitance there? Was it speed? What was the objection? I'm trying to understand. I mean, if you have a model that is equally accurate but also interpretable, what would be the objection to it?

Cynthia Rudin:

Exactly. They just don't believe that that can possibly happen.

Rashmi Mohan:

I see. Okay, got it.

Cynthia Rudin:

This idea that an interpretable model can always be replaced with a more accurate black box model, that pervades everything. I mean, even my own friends who work in interpretability sometimes believe this. I'm going to pick on my friend because I know she doesn't mind. So I have this friend who's a famous computer scientist. She's a famous AI expert and she is one of the smartest people that ever existed on this earth. Her name is Regina Barzilay. She's a breast cancer survivor and she was designing a model that predicts breast cancer five years in the future and she also really cares about interpretability and she told me, "Cynthia, this model, you can't replace it with an interpretable model. It's a black box. Nobody knows how it works, but it can predict breast cancer five years in the future." I said to her, "Really? Are you serious?"

And she said, "Yes, and I've built it on MGH data, tons of data. I tested it on Emory data and it works." And she published it in radiology and it's amazing. And so I said, "Okay, I'll go look at it." I brought my team of radiologists and students to take a look at this model and luckily Regina made it public and so we could actually play with it. And within a short time we figured out what was going on and we built an interpretable model that was just as accurate. And so it turns out that her model had been detecting very subtle asymmetries between the left breast and the right breast and the mammogram. Her algorithm was a classic black box algorithm. It's like a bunch of convolutional layers and a transformer. And transformers, they just churn up data like it's a smoothie. I mean, there's just no way you can reverse engineer what's going on in the transformers. But once we figured out that these models were detecting these subtle asymmetries, we actually were able to remove the transformer altogether and just create a symmetry detector for the mammogram. And so we got a model that was just as accurate as her model, but it's actually interpretable. And so because of that, it's actionable. We can pinpoint exactly where in the mammogram this asymmetry is between the left side and the right side. And so we can actually... What we want to do is take this model and be able to predict breast cancer in advance and tell people when they need to come back more often and tell people when they could come back less often and so on. That's an example where somebody like one of the smartest people in the world thought it's got to be black box only. And it turned out that that wasn't even true.

Rashmi Mohan:

The models that you talk about, Cynthia, are they broadly deployed in the field to a lot of organizations, hospitals or clinics, do they adopt these models and are they starting to be used more widely?

Cynthia Rudin:

Well, the 2HELPS2B score that I talked to you about earlier is used in most intensive care units in this country that have EEG monitoring. This is brain monitoring. This is brain monitoring for critically ill patients. This is a very common model that we published and you can just look it up. You can just look up 2HELPS2B score and there it is. Yeah. So if you end up with a brain injury and you end up in the ICU, there's a decent chance you'll get scored with our 2HELPS2B score. Hopefully that won't happen to you, but still to the point.

It is harder to adopt other methods. So the deep learning methods are harder to, because you got to get them approved by the FDA for healthcare. And so a lot of those models take a while to get it approved. A lot of models have been approved by the FDA, the black box models, and then they didn't work out. So you do want to be careful about launching things a little too quickly. And then like I told you, our crime series detection method has been used by the NYPD since 2016 to try to figure out which crimes were part of a series. It is easier to get interpretable models used in practice generally than black box models, I would say. We're doing our best. We're doing our best trying to get things into practice. Yeah.

Rashmi Mohan:

I hear you. Based on all that you've described so far, I mean interpretable models are just, it feels like innately it would be more trustworthy simply because you can see what's going on or you can understand what it is and you can provide feedback and tweak it. That sounds like something that irrespective of what your field of expertise is, I think you'd want to be involved, especially in these high stakes situations.

Cynthia Rudin:

One big challenge in getting a lot of things implemented is the lack of data for high stakes decisions. So for example, what I've been hoping is that the US government starts producing data sets because the government already has proven itself to be really good at producing data sets and creating challenges. So every year they have this facial recognition challenge and they evaluate facial recognition algorithms from all over the world. I feel like NIST, National Institute for Standards and Technology, they've been a real driving force behind the quality of facial recognition methods because they created a data set. And I don't see why they can't do the same for health monitoring. I don't see why they can't revolutionize heart monitoring by providing a giant data set of like ECG or PBG signals, which are the signals that come out of smartwatches. Heart monitoring, for instance, it could make a huge difference for a ton of people if we can do heart monitoring better. But right now, the only places you can get really big data sets are if you work at Apple Watch or something. And since I don't work at those places, I can't access those big data sets and the public data sets are terrible for heart monitoring. And so the lack of data has really been a major challenge for designing any kind of models, black box or interpretable.

Rashmi Mohan:

And is collaborating with industry on some of these initiatives, Cynthia, what's been your experience there? If you went to Apple with a proposal, Apple just being one example, but I'm just curious if you've had those experiences as well.

Cynthia Rudin:

I'm not really interested in working for a company and taking their proprietary data and producing a proprietary model. I'm interested in designing models that the public can use that's owned by the public, like models you can publish. I can't imagine that a company would want to hire me to design a model that I then release. Collecting that data is expensive, right? That's their secret sauce. They're not just going to release it. Yeah.

Rashmi Mohan:

Yeah. Fair enough. I would love to also understand from you, Cynthia, given that we have a lot of young professionals who could greatly benefit from advice on how to navigate their careers and maybe an additional lens of, say, women in computing. What has been the single or a few incidents that have shaped your approach towards problem solving or the choices that you've made with regards to your career?

Cynthia Rudin:

I had some really good role models. Ingrid Daubechies and Rob Shapiro were my PhD advisors. I could not have asked for better influences on my life than those two people. They're truly amazing. Ingrid, obviously a free spirit. She does not care what anybody thinks. She has her own notion of what's beautiful and she's going to pursue that. Rob Shapiro is the same way, but Ingrid is, she's out there, and she involved me in the Women in Math program at the Institute for Advanced Study. That program was a major influence on me, I guess, being around all these women mathematicians and you don't feel like you're being judged. There's a lot of confidence issues that come with being in a field and not being the majority group. There's a lot of imposter syndrome and things like that that people experience. You really think you're the dumbest person in the room all the time. It's important to get over that at some point and just like fake it until you make it. I don't know how else to say that.

I also don't think I'm maybe the greatest role model in the sense that I did get pretty down on machine learning for many years before I discovered the area that I really cared about, which is interpretability. And that came from actually working on a real problem with domain experts and realizing that what was in the field just wasn't doing it for me for solving this problem. It took me a long time to figure that out. So I didn't have like a direct route to getting where I am and it took me many years to get there. So not sure I would recommend people follow that.

Rashmi Mohan:

Oh, but that's real. I mean, I think a lot of us do have that journey. It's probably a more regular or more likely to happen than not. And so thank you for sharing that. I think that's very helpful and encouraging. And it's also nice to hear that you had these role models. I mean, do you have any advice on how to seek out role models? How do you find that person who'd be open to investing time and energy working with you?

Cynthia Rudin:

I'm not really sure exactly how you find the right people. I just know that the first few people I found, well, I mean, I went through at least two PhD advisors before I found Gary Flake and he was fantastic, but then he moved and he introduced me to Rob Shapiro. So that was just a coincidence. There's something to be said for choosing carefully and then realizing when it's not going to work out and switching to someone else. Yeah. The two people that I originally chose, they just weren't going to work out for various reasons. Their fields weren't right or their personalities weren't right. Yeah, I'm glad I didn't end up working with them because I wouldn't have found the people I found.

And my goal as an advisor, I think about this from the advisor side. My goal is to make sure my students don't go through what I went through. I'm really proud of my students who, a lot of them are professors now. They went straight from grad school to being a professor, whereas I didn't. It was many years between when I graduated and when I became a faculty member. And I'm really proud of what I've been able to accomplish with my students. They're amazing. These people are so smart and I'm so glad to have gotten a chance to work with them. I'm talking about my students, my former students, my current students. Just honored to be able to work with them.

Rashmi Mohan:

That's wonderful. Thank you for sharing that. That's a very, very positive way to reflect on that problem or that situation. For our final byte, Cynthia, what is it that you're most excited about in this field of interpretable machine learning, say, over the next five years?

Cynthia Rudin:

What I'm most excited about. Okay, so I'm really excited about Rashomon sets right now. So I was telling you, a Rashomon set is a set of equally good models. And because there are a lot of equally good models, there are a lot of simple models. And so finding as many of them as we can is what I've been trying to do the last few years. So I'm really excited about Rashomon sets and what they can do for machine learning.

I think a big question that people are asking right now, it's an obvious question is, how do you make a large language model interpretable? And nobody knows the answer to that. There are entire fields of people who are poking at the insides of these models trying to figure out what they're doing, but that's not the same as actually building a full model that's actually interpretable, like a model with interpretability constraints.

And it doesn't help that you can't train LLMs. That's really something that you can only do when you're at certain companies and you have certain resources and the time to do these experiments, these are very time-consuming experiments. People don't know the answer to that question.

And it took us years to even get to interpretable computer vision models. So from 2012 when AlexNet came out to maybe 2019 when ProtoPNet came out, we didn't know how to do it. I think that's an outstanding question that people are asking right now. We've been trying to build agentic models or LLMs that use tools and the tools have to be, if they're equally good solutions, we want to have the one that has the most interpretable or reliable tools. So we wrote a paper about that. And so we've been trying to get at it from that direction, but I think there's a kind of more crucial role for interpretability that it hasn't yet played.

Rashmi Mohan:

Wonderful. Cynthia, it's been an absolute pleasure to host you on our show. Thank you for taking the time to speak with us at ACM ByteCast.

Cynthia Rudin:

My pleasure.

Rashmi Mohan:

ACM ByteCast is a production of the Association for Computing Machinery's Practitioner Board. To learn more about ACM and its activities, visit acm.org. For more information about this and other episodes, please visit our website at learning.acm.org/bytecast.