1:14

Alexey

**This week we'll talk about bringing together research and industry and how explainable and interpretable machine learning and AI fit into it. We have a special guest today, Polina. Polina is a data scientist at SAP. She's passionate about bringing current machine learning research to business. In her PhD dissertation, she created a framework for churn prediction and this framework uses organizational trust theory and explainable machine learning methods. We will be mostly talking about explainable ML, but I'm also very curious about organizational trust theory. I have no idea what it is. Maybe we will talk about that. Welcome to our podcast!**

1:52

Polina

Thank you so much for the warm welcome. Very excited to be here.

1:56

Alexey

**We're pretty excited to have you here. And as always, the questions for today's interview were prepared by your Johanna Bayer. Thanks, Johanna, for your help.**

# Polina's background

2:05

Alexey

**So let's start. Before we go into our main topic of interpretable/explainable AI and ML, let's start with your background. Can you tell us about your career journey so far?**

2:19

Polina

So it's maybe also going to be further covered in the industrial PhD. But basically, I think the biggest element of my, or the earliest element of my journey as a data scientist is that I did an industrial PhD with SAP at the University of Mannheim, merging together sociology and data science. I also identify myself as a computational social scientist. Then after the PhD, I started as a data scientist at SAP. What I do day-to-day, I think – I used to always cover it, or call it end-to-end data science. Basically, from the point where somebody comes up with an idea to bringing a productive machine learning model to life and then maintaining it.

This is kind of what I can work on and that's what I cover in my job at SAP. I'd like to say, actually, that's an interesting point – I listened to one of your talks recently, I think at the Arize:Observe conference, and you were calling it a full stack data scientist. So I learned a new thing. [chuckles] And now I think I can identify with that as well. And one very important point that I have to mention – I described my job and this is something that I can say publicly, but anything further, I should not talk about. Everything I talk about here is me as a private person, not SAP, not an opinion of my employer. Just wanted to say that.

6:37

Alexey

**My Zoom crashed. Apologies for that. I don't know where it stopped. I remember we talked about a full-stack data scientist, and you told me that you learned that in the talk I gave the other day. Then after that, I think it dropped.**

6:58

Polina

Okay. After that, I mentioned that all of the opinions that I voiced are my own and not that of my employer and that I'm not going to be talking about the actual job that I do, because obviously, I'm here as a private person. So that's what I have to mention, just for the clarity of the situation.

7:19

Alexey

**Yeah, it's funny that you mentioned this full-stack data scientist term. When I first gave this talk like two or three years ago, it was a thing because the role of an ML engineer was not yet that developed. It wasn't that common. And MLOps wasn't that common either. So data scientists often needed to do everything end-to-end. There were, of course, data scientists who specialized more into I know the business side or specialized more in software engineering. But now, with ML engineers, data engineers, and other people working together on a team, I see that there are fewer full-stack data scientists, yet you say that you are one. I'm wondering how many of these people like you see in the industry.**

8:17

Polina

How many are still left? [chuckles]

8:18

Alexey

**Yeah.**

8:19

Polina

Yeah, I think it's also kind of my impression of what I observe on LinkedIn and how people's career journeys are developing, I think it's getting more specified. Maybe I would be then going closer to the business side and to understanding the business problems, compared to some people who develop more towards this MLOps specialization. So I do see that. But I think it's kind of where my roots were, in my early career, especially in the PhD project as well – I was the only person managing my own PhD project, and I was facing a lot of points, so it was very hard to bring many things together. So that's, I think, a simple explanation of how I first came to be a full-stack data scientist. [chuckles] In a PhD, you're supposed to do many things independently.

# How common it is for PhD students to build ML pipelines end-to-end

9:19

Alexey

**Is it a common situation when a PhD student actually needs to do everything end-to-end? Because I think it is, right? That's kind of the point. Or is there usually help?**

9:30

Polina

I think it really depends. My PhD project was actually an applied project for the company. There, of course, I faced this duality of, on the one hand, I have the scientific elements to it – I have the research – and then on the other hand, I have actual stakeholders and people who would, in the end, be using the model and benefit from it. So of course I was also facing all of the things that don't necessarily belong to a PhD project. But there are so many PhD projects – so many industrial PhD projects – so I think I cannot say it's 100% a must, but it was a big learning for me.

10:16

Alexey

**So the answer is, “It depends,” right? [Polina agrees] It depends on the project, it depends on the group where you work, and many, many other things.**

10:23

Polina

But really, I didn't want to phrase it this way, because “it depends” is an answer to everything.

10:28

Alexey

**[chuckles] Yeah. Right. [chuckles]**

10:29

Polina

So I still want to try better.

# Simultaneous PhD and industry experience

10:34

Alexey

**Actually, I wanted to talk more about your PhD at the end. But I think now, it's a perfect segue to actually talk more about that. You said that it was an applied project for the company – for SAP, as I understand. There was a scientific element and in addition to that, there were stakeholders that you needed to manage. For a usual academic PhD, the output would be like three papers, and then a dissertation based on these papers. But for you, it was different, right? In addition to the papers, you also needed to deploy a project, to show the business value. Am I correct?**

11:15

Polina

Yeah. Exactly.

11:16

Alexey

**Does it make it more difficult because you kind of have two goals?**

11:22

Polina

Yeah, I think this would be a situation or a problem that every industrial PhD student faces, where on the one hand, you have the research interests, and maybe your university supervisor, who pushes you into bringing the research further on. And on the other hand, you would have the industrial benefit. There might be a bigger or a smaller overlap as well, if it actually is the same project, or if it's two completely different ones. So for me, the overlap was rather huge but, of course, there were elements that were only in the research part – only in the output text, basically, in the dissertation. And, of course, a lot of the things that I learned during pushing things in production, for example, were not covered in my dissertation, so they were industry-only.

12:22

Alexey

**So you did not describe in your dissertation how exactly you used Kubernetes, or whatever framework, for deploying models – you did not write anything about that.**

12:33

Polina

Exactly. Maybe to get closer to the topic, for everybody to understand more what I did – my dissertation was focused on studying the relationship between a software as a service provider and their customers. Basically, I looked into how trust is built up, theoretically, and then tried to see it in reality. I also wanted to understand how this trust element impacts the continuation of the relationship. Basically, if you phrase it in machine learning problem terms, then it's a churn problem. So basically, predicted churn was the model.

For the organizational trust things, I think, of course, a lot of things were just describing the theories and not exactly immediately relevant to the industry. If you put it this way, actually, for the text of my dissertation, how I bring it to production was not necessarily relevant. It was a tiny part of my defense. During the defense, I mentioned that it is actually live, but it's not something that has a chapter or something like that.

13:52

Alexey

**So it was a tiny part in the dissertation, but I imagine the actual work... I don't know what your experience was like, but from my experience, this is usually what takes most of the time.**

14:02

Polina

It was a big learning curve for me, because I came from this kind of idealistic perspective, maybe, that whatever insights I generate – anything could be very, very beneficial to the industry. Then I realized that there is just so much more to it and that a real data science project is actually more complex. Also managing it completely is a big challenge. Yeah, so that was a big learning for me. I think maybe that's why it was more work as well. Maybe a fun story: The first time I was discussing pipelines around machine learning engineering, I was basically at the stage where it's like, “Data comes in, I want it there,” but I was not understanding maybe... no, I do, but I wasn’t understanding the complexity of how things in the industry can actually work.

15:11

Alexey

**Also, it means that there is an additional risk factor. For a usual PhD, the typical risk factor could be that you do not find what you wanted to find and then it's like, “How do you write your PhD if you could not prove that (in your case) organizational trust theory can be applied to churn prediction.” So that's one element. Then another element, even if it theoretically works, will it work in practice? So you have two things that could go wrong, as opposed to a usual PhD, where it's maybe less risky. Am I right?**

15:56

Polina

I think so. I had a lot of research showing that it can work. And there are also a lot of interesting studies showing other data and translating it into industrial terms. So I think, for the PhD, there is a lot of risk anyways. I would not say that there is more risk in this case. Probably, at the level of a PhD, it's as much risk as it can get. It either works out and you get the title or it doesn't. I think that's actually kind of the same for all PhDs. Maybe, to turn it into a positive direction, it's also double support – or was a double support for me – because my supervisor at the university was very understanding of the setup and was also helping me to find the elements of value that I can get to the company.

My team in the company was also very, very supportive of my PhD. I think that's kind of where, if it all fits together nicely, then it's actually a lot to learn and it gives you two things. First of all, there's the research experience, but then still a lot of work experience as well. There's a reality check, to be honest, because I think my original motivation for doing it in the industry was that I didn't want to just write and then put it on a shelf, or just publish something and never know if anybody got anything out of it. Here, I can actually see that it is applied and there is a lot of learning for everybody, including me.

# Support from both the academic and industry sides

17:57

Alexey

**You mentioned that you had a supervisor from the university side. Was there also a supervisor from the company side?**

18:06

Polina

Not immediately. It's also different in different industrial PhD setups – sometimes there is a supervisor. In my case, it was basically that the teams that I worked with were interested in the project, but there was not necessarily somebody who would just supervise me completely.

18:28

Alexey

**But there were some stakeholders, right? Still, somebody from the company side would guide you and say, “Okay, this is not what we need. This is not what we meant. We want a different thing.” It was like that, right?**

18:41

Polina

Yeah, exactly. Basically, my stakeholders were helping a lot.

18:46

Alexey

**How did this actually work? Were you talking to them every day? How often did you get feedback from them? Were you talking to other team members? How did it look day-to-day?**

19:05

Polina

Day-to-day, I think it was just a data science project. Just the data science work that you can imagine – regular calls with stakeholders. I think that's not that much different from what every data scientist who has business facing roles does every day. Of course, there was a lot of exchange in the team. Again, very typical for data science teams to just talk to each other. [chuckles] No secret information revealed here.

I think what was different for me was that I also had a chance to attend conferences or summer schools. Again, that was this merging together of the industry and the university lifestyle. Basically, I could go to a conference or to summer school, and then I could bring my ideas from my project and then learn on the go from the research. That was a very interesting experience.

20:11

Alexey

**And by conferences, you probably mean academic conferences, not industry, right?**

20:16

Polina

Exactly. Academic conferences.

20:18

Alexey

**The ones where people publish papers, print posters, and then discuss research.**

20:25

Polina

Exactly.

20:28

Alexey

**Summer schools are a lot of fun. I wish there were more things like that. They're not very common in industry, I guess. They're usually for students, right?**

20:37

Polina

Yeah. Also, my PhD was partially during COVID. The first year I started, it was 2019, so I did summer school, then. [chuckles] And then 2020-21, it was virtually impossible to have any real-life summer schools. But I'm also kind of thankful for the situation of being locked down because it forced me to write a lot. [chuckles]

Balancing the PhD and industry sides of the projects

21:04

Alexey

**Okay. But did it mean that, in addition to your work as a full-time data scientist, you had some extra work because you also needed to write it all down? Or was it balanced? Your team knew that this is actually your PhD project, so then you could spend time saying, “Hey, this week, I'm actually working on a paper. I'm not working on the project.”?**

21:31

Polina

It was very balanced. It, again, actually depends on the setup. In my setup, it was balanced, but for other people... Also, in academia, I think some people are working on projects, and then on top, on their PhD. When there are deadlines, you're just on your own with the deadline and it's you and your personal time management. I think that's probably shared among all PhD students – not specific to this industrial setup.

22:03

Alexey

**I remember when I was getting my Master's degree at TU Berlin, I saw the PhD students all of a sudden realized that the deadline is soon and then they would spend 24 hours, 7 days a week on just writing the paper. They did not look very good when I met them at the university. [chuckles] They were very stressed.**

22:30

Polina

Yeah. The stress level, I think, for any PhD student... At some point, it's just your project and you want to get it out in the best possible way and share the experience. [chuckles]

22:45

Alexey

**So it was an explicit decision that you made, that you wanted to do an industrial PhD project. You didn't want to do just a PhD, you wanted to immediately apply this research at the company. How did you come up with this realization? How did you learn that this is what you want to do?**

23:07

Polina

In my Bachelor's and my Master's, I was already very interested in doing data analysis, basically, but for companies – or with data that can bring something valuable to someone. I did a couple of projects that were also for companies during my Master's in seminar work. Then I realized that I'm interested in data that is just very often owned by companies, so if I actually want to develop in the area that I think is most interesting for me, then I want to go there. Also this idea of not just putting the PhD on the shelf and that's it – I think that is very important.

24:04

Alexey

**So, you already had some connection to the industry and you did not want to lose this connection, but you still wanted to do a PhD. Then you figured out that you can actually combine both – you can stay in academia and work on industry projects.**

24:20

Polina

Exactly. And I think both can actually learn a lot from each other. I still think I would not have done it differently. I would have done the industrial PhD if I was asked again.

# How common the industrial PhD setup is and how to get into one

24:38

Alexey

**How common is this setup? For example, I know that – again, at TU Berlin, where I studied a long time ago – I don't remember that the group where I was had such a direct connection with industry. Usually, the company usually gives some money to the group for them to work on something, and then they give some data. Then they're like, “Okay, bye. Now figure this out and then come back to us in five years,” something like that. That was my feeling of how this thing worked. My question is, how typical is what you had, in general?**

25:24

Polina

I think it depends very much on the industry, and very much on the subject of the PhD. I am a social scientist by education (by training) so it's very not common for social science, or rather unusual. I think in other disciplines, it's probably more common. Somehow I'm thinking about chemistry, but do not quote me on that. I think there, it might be more common to have a collaboration. But yeah, if any of the listeners are interested in finding the PhDs, you can go (at least in Germany) for the big companies. SAP, obviously, has this setup sometimes. Also, I think Siemens Bosch is offering this. I think the automotive industries in Germany also have this industrial PhD setup, certainly, at the moment, in the area of machine learning. If you're interested, just do a bit of research on their job seeking websites.

What I also think you can look at is university websites. Sometimes they advertise such cooperation programs, or industry-related PhDs. I think the Technical University in Munich definitely has a page that offers that. But again, it is always a question of needing to find a professor that supports this collaboration. And in a company, it also must fit some project. It's not exactly common for companies to just hire scientists to run around their own and do wild things. But if it fits together, then... There are actually more opportunities than I thought there were. As you can see, I'm asked this quite often, [chuckles] so I already did some research like, “Okay, what can I recommend? Are there any companies that are doing that?”

27:41

Alexey

**So if I wanted to check what possible projects I can do with SAP, what kind of Google query would you say I need to use? “SAP machine learning and industry PhD,” something like that?**

27:54

Polina

Yeah, I basically used “PhD,” or “doctoral student,” or “doctorates” in the search, because I think that's how the positions are frequently advertised. But, again, each team can define it in different ways. I think a query with a PhD student or a PhD research position should work.

28:22

Alexey

**Do they require you to speak German? Or is English sufficient?**

28:27

Polina

This so much depends on the company and, again, also the team. I think there is no 'Yes' or 'No,' over here.

28:36

Alexey

**So it depends, right? [Polina agrees] You just find a position and they probably say if you need to speak German or not. Okay.**

28:45

Polina

My fingers are quickly Googling, I think Daimler required German for some positions. But... yeah, it's very, very dependent on the specific team.

28:57

Alexey

**Was your dissertation in English or German?**

28:59

Polina

It was in English.

29:01

Alexey

**English. Because I know that in some countries, you cannot write in English – your dissertation has to be in the language of the country, university, whatever. But I know that in Germany, that's not the case. In Germany, you can publish either in English or in German. I don't know about other languages. Maybe you can do it in Latin, or? [chuckles]**

29:22

Polina

In Latin, I'm not sure. Because in what language would you defend if you are writing in Latin? Not sure, but... I think now that the research is basically very international, a lot of universities actually expect you to publish in English. So I think that's what motivates this idea that you can actually submit an article-based dissertation in English as well if that's exactly the expectation for publication.

# Organizational trust theory

29:52

Alexey

**I also wanted to talk about the content of your dissertation. At the beginning, when introducing you, I said that you were developing a framework for churn prediction that used organizational trust theory and explainable and interpretable ML. So what is organizational trust theory?**

30:16

Polina

[chuckles] I think this will need an episode of the podcast of its own because you can actually look at trust from so many angles. Why I say “organizational trust theory” is because very often, when people say “trust,” they mean interpersonal trust – basically, “I trust you as a person.” But in organizational trust, it's different agents – basically, organizations interacting with each other. All of them have people inside. You can go to this personal trust level, but that's not what I did. That's why I say organizational trust as a term. What I used was the “ability, benevolence, and integrity” framework. I called it ABI. Basically, there are different layers of trust and there are these more technical trust abilities. So in the context of software that's, “I believe that this software will work.”

Let's take an example of Microsoft Office, because I think everybody is familiar with that a little bit. The technical trust is, “I do know that this software allows me to do what I need, like writing or PowerPoint [presentations], or emails (or something like that).” Then, in a long-term relationship between two companies or, between a company and the person, there are these more relationship-based elements – benevolence and integrity. This is where it goes into a more interesting (for me) direction. With integrity, it goes more into this, “I know that they will be there for me if something goes wrong.” direction. And with benevolence, there is this long-term support delivery, basically. Integrity happens after you sign the contract, where you don't really know what's going on, but then somebody talks to you and they assure you that it's going to go right.

Benevolence is more like the actual support that you see over time. Why it was interesting for me is – for the relationship between two companies – before that, you could just like... I'm, again, switching more to people, because I think it's more understandable with people. Imagine you bought a Microsoft CD – I still had CDs when I was in school. Basically, you have this CD forever, and that's basically what you know is working. Until you have a completely incompatible system, it is going to work. But let's say you buy it now – you buy it now and it works for some time. Then you have a question, or there's a feature that you bought it because of, and it's not delivered. You open support tickets, and you go to the community website, you ask questions, and they maybe don't get answered. So there are a whole lot of differences for the relationship in the subscription context compared to this, “I have this one CD and nothing goes wrong until the system gets so updated [that it doesn't work].”

33:56

Alexey

**It might go wrong because there's no way to get support, right? There was no way [to get support]. Or was it not difficult?**

34:00

Polina

I mean, there was definitely a way to get support. But I think it was not as prominent in the relationship. With the subscriptions, I think after I started researching it, I realized that it's actually kind of growing everywhere – anything that has a subscription, you can kind of trust it to go for a long time, but if it accidentally breaks or something happens, then you actually will go to the company and ask for help. This is why it's just so much more relevant nowadays.

# How price relates to trust

34:36

Alexey

**And what about the price? One of the big drivers for me to change... Let's say I use some product, some internet company, and then they decide to increase the price. Then I think “Okay, what are the alternatives that might be cheaper?” Then I do research, I find a cheaper alternative and then I switch. Does price have anything to do with trust? Or maybe it's a separate thing?**

35:06

Polina

This is actually not something that I personally researched, but I think... Also looking at just the consumers, again, outside of the companies' relationship. Because [with companies] there are contracts, and there is a lot more governing this. With regular end users – let's say I have Spotify and they increase the price – I think this is definitely a mechanism that goes to this ability, again, to kind of the level before trust. This is a place where it's actually also harder sometimes to justify the relevance of this relationship, because it can be broken because of some other things. Let's say, I use Spotify and then at some point, they just don't have the songs that I like anymore. This is, again, this ground reason for why use this and not exactly the...

36:07

Alexey

**Could it be ability? You expect the software to be able to give you the songs you like, but then all of a sudden, they don't have the songs you like. [Polina agrees] They kind of violate your trust to be able to provide you with the songs.**

36:23

Polina

I think in one of the papers that I read, ability was always the most important and the most significant element of trust. Basically, how it goes is – ability is mostly consistent over the relationship. You can imagine if Spotify doesn't have the songs that I like anymore, then, even if they're a nice provider, even if they support me so much, I would probably not continue using the service, (hypothetically – they have a lot of songs that I like). [chuckles]

37:04

Alexey

**Then maybe the example with price could be related to integrity? You kind of expect that the price is a certain number, but then all of a sudden they change it. Maybe I misunderstood the framework, but could it be related to integrity or is it more like something else?**

37:22

Polina

I think it always depends on how your relationship was set up. What I mentioned before with the contracts – you actually signed a contract and there is an agreement between you and the company that there are specific regulations about the price change. So it's actually hard to answer this one without knowing what the agreement looks like. I'm unfortunately not the expert on contracts in this [inaudible].

37:51

Alexey

**How many people actually read these agreements before starting using services?**

37:57

Polina

I try to, but I always ask myself, “Well, what should be in the agreement for me to not use the service?” And then when I realize that it's not that much, I just skip reading it. But I actually try to be conscious about that, because there are many things that can be in the agreement.

# How trust relates to explainability

38:19

Alexey

**And then how is it related to explainable AI? I guess you use this framework of organizational trust to somehow create features, maybe, for your model or somehow guide your project. [Polina agrees] But then there's another component, which is this Explainable AI. How are these two connected?**

38:44

Polina

I think this is also maybe one of the arguments, or one of the ways, how I turned to research more towards the explainable AI direction. On the one hand, I think in social sciences, it's very common that you have the features (or variables – however you'd like to call them) and then you try to understand how they are connected to the outcome variable, in a way. In industry, it's very often about just modeling and being accurate, but not about building a theory and trying to show how it works in detail. When I was starting, I was actually overwhelmed a little bit by the research on churn that exists. I tried a lot of neural networks and was actually disappointed a little bit, because on tabular data, they were not exactly performing perfectly (to maybe understate this). They were not performing at all.

Then I turned to the more black box models – random forest, XGBoost, some of the things that are more classic – and then I started understanding that, actually, you have the feature importance, of course, but going into this understanding of how the features actually contribute to the outcome and how your model works, this is not exactly a standard element of data science (or not a standard element of the black box models). What I actually loved is that I realized that sometimes the stakeholders can feel the same way, so there was also a demand that the model that I'm building should be more than just a score. Because the score doesn't really always tell you the story.

So I realized that, on the one hand, it's something that I have, in me, trying to explain with a model, but also knowing how it works. But then also, on the other hand, that this accuracy-driven (or curve-driven, whatever metric you have – because for me, it was the classifications of these two that come into place) is sometimes not enough for the end users. That's also what I realized during my PhD project. With a social science background, I'm able to communicate and this explanation actually helps to communicate models in many ways. This is how I ended up researching, in 2019, everything around SHAP values, LIME, and all the things that seemed super groundbreaking at that point – and now everybody uses those. So I think it was actually very exciting for me to understand the entire data science community growing into the direction that feels so natural for me. So that was just very, very inspiring, and a very good moment. And then of course, that's what I used in the dissertation as well.

# The importance of actionability

41:54

Alexey

**In practical terms – for you, it was about discovering the connection between... Maybe you have different groups of features: features related to ability, features related to benevolence, features related to integrity. And then you use tools like SHAP values and LIME to determine how exactly these features connect to the outcome. Then if a stakeholder asks you, “Hey, why is the score 0.9 for this user?” You say “Because, they think that our ability is not really good.” Right? Was it something like that?**

42:27

Polina

Yeah. I think there was also a post on LinkedIn, where I showed a more public display of what came out of my PhD. So that's the level where I can stay on this public thing. For the end users, I basically show several components of how the outcome is modeled, and it's actually very important – and that's also something that I learned over the years of trying to explain models – is that sometimes it also has to be actionable. So integrity, per se, is, for example, an interesting thing to discuss or to research, but telling the end user “integrity is the reason for this and that,” is not exactly actionable. They don't know what integrity is. They don't have any options to connect with this.

So I also realized that actionability is a very important element during this. If you look, for example, at the upcoming, first World xAI Conference (1st World Conference on eXplainable Artificial Intelligence) coming up in July – I'm very excited about going there. There is definitely a group of sessions on actionable explainable machine learning, so you can also see that actionability is a very prominent trend in the research, currently.

# Explainability vs interpretability vs actionability

44:03

Alexey

**What is that? What does “actionable” mean? Did you say “actionable, interpretable AI (actionable explainable AI)”? What does it mean?**

44:14

Polina

Yeah. I'll maybe comment on explainability and interpretability. “Interpretable” is basically, for me, more of a technical term, meaning, “My model is logistic regression, I can fully see through that.” There was another podcast, I think, a couple months ago on DataTalks.Club about explainable machine learning, where you mentioned the glass box model – so that's that. And “explainable” is more user-facing. Are you able to explain your models?

I have a metaphor for this. Somehow I got obsessed with cats this year with cat GIFs and so on. There are many GIFs with cats getting out of boxes and they're curious. [chuckles] For me “explainable AI” is all about these curious cats and to think about “To whom are you explaining? Who would be jumping out of the box to ask you more questions.” So that's kind of the explainability, where interpretability is obviously an element, or an attribute, of the model. “Actionable” is more like, if you have already mentioned this cat in a box – can the cat do something with what you explained? Or is it just a curiosity?

45:37

Alexey

**That's “actionable” machine learning. And then we have “interpretable” and “explainable”. So they're kind of different characteristics of a model – the same model can be interpretable, it can be explainable, and the model can be actionable.**

45:51

Polina

Yeah. For “actionable,” the way it is used, for example, within the conference that I mentioned, it's actionable/explainable. So you not only explain, but you also give some insights into what actions the end user can take based on the model.

46:10

Alexey

**Which is quite important, right? If our model gives us a churn score of 0.9. For us, it's like, “Okay, what do we do now? Does this mean we need to send them a promotion to try to keep our users, or we need to do something else?” So this is a part of actionability plus explainability?**

46:32

Polina

Yeah. It's also something that's not relevant in all cases. I think explainable machine learning, I'm always arguing in favor of that, because it helps you as an end user. It also helps you as a data scientist – it has many positive sides to it. But actionable, I think, is very often maybe an attribute of actually something where the end users want to take action based on that. So maybe it's more thinking of decision-making.

47:03

Alexey

**Then you mentioned this term “glass box model”. I guess this is the opposite of the black box model.**

47:10

Polina

Exactly.

47:11

Alexey

**We have things like a random forest, which could be considered a black box model. And then a glass box model would be logistic regression. Right?**

47:21

Polina

Exactly, yeah.

# Complex glass box models

47:22

Alexey

**And then would random forest plus SHAP values be a glass box model or black box?**

47:31

Polina

I think maybe I will mention a couple of models that are more complex glass box models before that. What I'm mostly excited about in the research now are generalized additive models and neural network based models that are basically additive. Neural additive models, or neural basis models, are very exciting examples of those. I have been, in the past weeks and months, looking at how they compare to random forest and SHAP and I think what I don't know yet is, “What is the baseline?”

Because basically, in the end, we have a way a neural network would model something and we have a way which random forest with SHAP would model something, but what is the ground truth? And very often, we don't know. With random forest and SHAP, for example, the SHAP values only try to approximate – they never really tell you, “This is 100% how the model works in every situation.” So you still don't get this kind of real see-through glass box feeling of it, but it's still more than just a black box. So I do see that it's definitely not a glass box model, but maybe like...

48:58

Alexey

**Interpretable?**

49:00

Polina

I think it's explainable. It's not interpretable, per se, because it's not exactly 100% that the model itself that gives you the outcome, for the interpretability.

49:14

Alexey

**So for interpretability, it has to be a glass box model, like with logistic regression model?**

49:20

Polina

For me, yeah. Yes, I think there are different ways to put it. Basically, I think every person in the field maybe has a slightly different definition of interpretability. I think that's...

49:37

Alexey

**But linear models are usually interpretable.**

49:39

Polina

Exactly.

49:40

Alexey

**Like decision trees, probably. Right?**

49:42

Polina

Yeah. I think with a random forest, I think you could actually go into all of the trees and learn it all. But sometimes your brain just doesn't have the capacity to do that. So that's where interpretability breaks, I think. There's also a trick with explainability – there's recent research that I mentioned, that I like a lot – with explainability, you would say, “I have a random forest forest and SHAP, and for each feature, I know what the contribution to the outcome is, so that's explainable.” But in fact, when you add a person to this, sometimes it's completely not explainable, because maybe your features are named in a way that nobody can read. Or maybe they are just so not understandable that no human can know what this means. I think this is more commonly applied to computer vision or text where it's like, “This pixel is gray, so this is why it's a cat.” This is not exactly explainable in terms of how we would put it for humans.

50:47

Alexey

**But then for neural networks, I think, there are these techniques that show activation regions. They show something where you have a picture and it highlights the area around the ears, highlights the area around the nose – around the areas of the picture that shows that, “This is a cat. These are the areas where the neural network is activated. That's why we think this is a cat.”[Polina agrees] And this would be an example of explainable machine learning, right? An explainable model.**

51:20

Polina

Very often... I'm not an expert in computer vision, I'm more of a tabular data person. But for me, when these kinds of activation maps are displayed, this is definitely an element that makes it explainable because, of course, there is much more in the background that is calculated to just display that. But this display helps to communicate it to humans.

51:48

Alexey

**I just want to summarize what you said, to make sure I understood. So an interpretable machine learning model should be a glass box model – it should be something like logistic regression, linear regression, generalized additive mode – the sort of model where you can look at the coefficients that the model learned, and then you can kind of make sense of this. And then an explainable ML model can be a black box model, but there could be a method that helps us understand the output and then we can explain why there was a certain prediction made. Then an actionable ML model would be a model where there is a score and we know what to do – what kind of action, what kind of decision to make – based on the score. Right?**

# Does the explainability of a model follow explainability?

52:37

Polina

Yeah, exactly.

52:39

Alexey

**There is a question from Satyajeet. The question is, “Is an interpretable ML model necessarily explainable as well? Does it follow from being interpretable that the model is also explainable?”**

52:57

Polina

For me, it does not. If you're interested, there is a very nice paper – maybe we can link it in the notes for the episode later on. There is a paper about explainable feature spaces – it's from a research group at MIT – and they are looking into different... Basically different “curious cats,” if you put it in my words – “to whom do you explain your model?” It actually has a very nice visualization of how different explanations may matter for different groups. If you're talking to data scientists, for them, an interpretable model is probably explainable, because your end user cares about individual features and they know what the features are, and they are also very technically skilled to actually also understand everything that the coefficients mean.

If you're looking more in the area of ethics, for example, you have very different backgrounds there. For some people, they actually want very different explanations for your model than this kind of strictly technical outcome of the interpretable model. Even if it's logistic regression, you can get very confusing features in and therefore very confusing outcomes. Then the closer to the end users, or to decision-makers, you get, the more explainable your feature space must be. So it's not enough to just label your features “1, 2, 3, 4, 5” and then put it to a decision maker and tell them “Well, five is one, therefore, it's 0.9.” This is technically interpretable – not at all explainable and not at all understandable.

54:44

Alexey

**Okay, so what is explainable for a data scientist is not necessarily explainable for the marketing person. Right?**

54:52

Polina

Exactly, yeah. [cross-talk]

54:55

Alexey

**So for a data scientist, maybe it's true – from interpretability, explainability follows – but for the rest of the world, maybe it's not. [chuckles]**

55:08

Polina

Yeah. I think my logic is always like, “Think about who you're trying to explain to.” And your explainability is always based on your audience.

# What explainable AI bring to customers and end users

55:23

Alexey

**Then there is a question about trust, but I think it's a different sort of trust, not the one we talked about (organizational trust) but maybe it is. You will probably tell us which one it is. “Do you think that explainable AI models can bring trust among customers or different stakeholders?”**

55:38

Polina

Yeah, so this is a research direction that's very common now, I think – also in computational social science, in AI research – that focuses on how people interact with machine learning. It's not the trust that I focused on. In the neural basis model and the sparse polynomial additive model, they actually test how comfortable the end users feel when they see the explanations. I think that's something that is getting more and more prominent now. You also see it in the SHAP papers, that people test, “Can people really understand it?” Researchers want to know, “Does it really get nice feedback from the humans interpreting the models?” I think explainability also really helps to maybe demystify this machine learning phenomenon a little bit, where people might think that it's just like, “Press a button and then you have something and you will never know because the machine is smarter than you.”

I think the machine is not smarter than you and, very often, humans and end users have information that the machine doesn't have. So it's actually helping to build trust, in my opinion. But also what it helps to build is this power of people knowing what the machine learned, and then adding something on top of this to get the best outcome. It's very interesting because for large language models, for example – also, I'm not an expert – but for large language models now there is an increasing demand of “Where does the information come from? Is it something that model learned? Can you point to a document where this comes from?” So it gives you a more grounded interaction between the model and the end user.

57:43

Alexey

**Because you want to know if the model just hallucinated and came up with this out of nowhere, or if there is actually a document where this is described, right?**

57:51

Polina

Yeah. With tabular data and simpler black box models, they do not hallucinate in this way. [chuckles] But with large language models, it just gets more prominent that you really want to make sure that it's not a hallucination.

58:12

Alexey

**Do you have a couple of more minutes?**

58:15

Polina

I do, yeah. I'm here.

# Can all trust be turned into KPI?

58:18

Alexey

**Because it's a very interesting question and I see that we are running out of time. I want to ask this question from Antonis. Antonis is asking, “Is there a way to track organizational trust? Is there any KPI or metric related to that?”**

**I was thinking, when you were describing that, “Okay, we have ability, benevolence, integrity, and then we also have this framework from interpretable, explainable machine learning.” So we can kind of link all the predictions because we understand that, “Okay, these features are related to ability and we see that more and more users churn because of that.” It would make a great metric that people from top management would really understand. Do you see this happening in practice?**

59:10

Polina

I think what I must say here is that trust is incredibly difficult to measure. What you have is basically a lot of proxy variables – a lot of variables Alexey, as you said, associated with (related to) trust, but you can never say that something 100% captures that. Because it's so hard to catch, [chuckles] I think it's impossible to make it a KPI because... You can have measurement errors – you never know what's really happening between the people.

Because organizations, as I said, are actually people, so there are a lot of things that you would not want to track. There's also compliance, GDPR – a lot of laws and it's just not ethical to track this on a level where we would try to observe it. So I think, overall, I would say it's impossible to really make it a very well-measured KPI, and when it's not well-measured, then it probably shouldn't be KPI. That's my point on that.

60:29

Alexey

**Are there good proxies that will at least give you some indication that, “Okay, we're losing customers because our integrity is not good?**

60:40

Polina

I think this will very much depend on the company. Each company has very different perspectives on who the customers are. I talked about Microsoft Office customers and Spotify customers – of course, you cannot measure it in the same way, for example, for these two products. So it's very different. One of the learnings from my PhD was that there is also a lot of research in marketing and relationship – studies between companies that show that it's important to show that, yes, it is in fact true – there is a role that trust plays in the relationship. But I think it's so hard to measure and it's so specific that...

I like that it is a research project for me, but maybe not exactly a good KPI for companies. Then we also talked about the fact that ability to making a great product is actually more important than... I think we have the response or the answer to that. Building great products would be a priority anyway.

62:01

Alexey

**I guess it's a good idea for another research project, right? Because I can imagine that this could be useful, or at least I think that now. Who knows what the reality is? But I can imagine that, for executives, it could be useful to see, “Okay, we're losing customers because of that thing. Let's see how we can improve that thing.” But as you said, if you focus on ability, maybe the other things will fall into place.**

62:28

Polina

Yeah. I think there are many people researching churn in many different ways. There are a lot of ways of how to look at that.

62:41

Alexey

**Okay. I think we should be wrapping up. Thanks a lot, Polina, for joining us today, for sharing your experience with us, for telling us about your experience doing a PhD, and your work. And thanks, everyone, for joining us today too, and watching us, asking questions. Have a great week, everyone!**

63:02

Polina

Thank you so much!