1:22

Alexey

**This week, we'll talk about causality. We have a special guest today, Alexander. Alexander is a machine learning researcher, educator, and consultant. He has worked with many companies across Europe, in the United States of America, Israel, where he designed and built large-scale machine learning systems. He is also known as the author of Causal Inference and Discovery in Python – this is the topic of today's interview. Welcome to the show, Aleksander.**

1:52

Aleksander

Welcome, Alexey. Thank you for having me and thank you for the invitation.

1:56

Alexey

**You're more than welcome. As usual, the questions for today's interview are prepared by Johanna Bayer. Thanks, Johanna, for your help. Yeah, let's start.**

# Aleksander's background

2:06

Alexey

**Before we go into our main topic of causality. Let's start with your background. Can you tell us about your career journey so far?**

2:15

Aleksander

Sure. I started my journey with computers when I was a kid – I was like, 5, 6, 7 years old – I was doing a little bit of programming, because my father was a programmer back then. Then I had a very, very long break. After studying, doing my second degree, which was psychology – social psychology, experimental psychology with neuroscience and so on – I fell in love with statistics. Looking into and Googling about statistics, because I was very interested in this topic, and I was also Googling about what is going on in computer science at this stage. I learned about Python – this led me to machine learning, and that was the start of my journey into this rabbit hole. One of my first machine learning projects (real machine learning projects) was a scientific one. We worked on predicting – on finding prediagnostic predictors of dyslexia in children. I was very excited about this project.

Actually now, after many, many years, there's a tool being developed that will help people diagnose the risk of dyslexia in very young children, which is very important because early diagnosis can help those children start specialized training that can help them overcome the difficulties in their adult life. From there, I started working with an international consulting company called Lingaro. That was a place where I started developing many more complex architectures for global clients, working with NLP and other models. That was the beginning. Then I worked for other companies. Then I moved to Tel Aviv and I worked with a cybersecurity company. And then recently, I finished writing my book and then moved to doing consulting for companies and focusing on my educator and causal ambassador – causal global advocacy work.

4:45

Alexey

**Do you remember as a kid what exactly piqued your interest in programming? I have a kid – he's seven years old – and I tried to show him what you can do with an algorithm. We bought a robot. It's possible to program a robot – there's like this visual interface where you can tell the robot something like, “First, take three steps forward, then rotate.” You can create loops there so it can rotate 10 times or something like that. And he was not impressed at all after I showed him that. [laughs] So I'm wondering how to show him that programming is cool. Do you remember how it happened to you?**

5:26

Aleksander

I don't know how relevant this will be for you and for him (or for her). [chuckles] For me, you know... I didn't have a robot back then. I just had a monochrome display PC (186 or 286 or something like this) with GW Basic. I wrote myself a piano – I could play melodies using computer keyboards. That was my main achievement back then. [chuckles]

5:54

Alexey

**That's amazing. I have no idea what these terms that you mentioned mean, because I got a computer pretty late myself. But anyway, thanks for sharing that. I guess not every kid should be immediately impressed when they see an algorithm – a robot that could be programmed.**

6:12

Aleksander

Perhaps. You know, everyone is different, so... [chuckles]

# Aleksander as a Causal Ambassador

6:15

Alexey

**Yeah, right. Now, you focus on education. And did you call yourself a causality ambassador?**

6:26

Aleksander

Yeah, that's one way to think about what I'm doing.

6:30

Alexey

**So what does it mean? What do you do as a causality ambassador?**

6:36

Aleksander

I do a couple of things. I do my best to democratize the set of methods and the style of thinking – I think both things are very important here. I have a feeling that causality in general is a tool, or a set of tools, that can be very, very helpful for individuals and businesses alike. This is just something that, for one reason or another, we haven't learned in our curricula. Many people were not lucky enough to just get these ideas passed on to them from the teachers, from the environment, and so on, and so on. I believe that everyone deserves to understand how it works and to have a chance to apply this to their work.

7:31

Alexey

**You said your goal is to democratize this style of thinking. So what exactly is this style of thinking? What is causality? How is it different from the usual style of thinking?**

7:46

Aleksander

Well, from the usual style of thinking – I don't know. It probably depends where you grew up and in what circumstances, what context, and so on. But thinking about the data science community, most people are going into a journey that is focused around traditional statistical machine learning, which means that we look at associations. Seeing associations is great – it gives us great opportunities in many different contexts. But sometimes, it also comes with a set of risks, which means that we can see an association and this association might be because of another variable that we do not observe. And for whatever reason, we can think about this association as a valid tool to make a conclusion that is either implicitly or explicitly causal. In other words...

8:52

Alexey

**Do you have an example, maybe? It's a bit abstract.**

8:55

Aleksander

Yeah, sure. Let me make it more concrete. So I was just going back from my lunch and I was watching YouTube Shorts. You know what that is?

9:06

Alexey

**Yeah, they're addictive. [chuckles]**

9:09

Aleksander

Yeah, kind of like Instagram – short movies. And there was one where there was a guy speaking and trying to convince the audience that there is a correlation between the color of your skin and the likelihood that you will commit violent crimes. He was citing different types of statistics for this. So this is an example of associated thinking. Maybe we say something like, “Hey, there are people with certain properties – physical, mental, or psychological properties – and there are some outcomes. And we think that these outcomes are more frequent for one group versus another.” Then if we just speak using this associational language, and we talk about those associations, “Here's one, here's another one, here's another one.” It's very easy to make people start thinking that this property of this group of people is linked to this outcome. Sometimes it might be something completely different. So maybe people with this certain property – physical or psychological – are also for whatever, let's say, historical reasons in a group of people who have much lower income.

Maybe they have less parenting skills on average within a certain geography or location. And this might be also related to the fact that there is more violent crime in this particular neighborhood or for a group of people that accidentally also have another characteristic. But if we don't look at this third variable, we might start implicitly thinking that those two properties that we are observing are related. We also usually say “related,” but what we think is “related causally” – so there is something that is causing the other thing. From a purely associative point of view, or a correlational point of view, we often cannot distinguish which type of situation we are in. We might see an association with temperature and I don't know... there's an example in my book – it's a very simple example, perhaps an intuitive one. Drownings are correlated with ice cream sales. Right?

11:50

Alexey

**I can see where you're going with this. Yeah. [chuckles]**

11:53

Aleksander

Somebody who is a scientist might hypothesize that, “Hey, there is sugar in ice cream. Maybe people who eat ice cream become a little bit slower to react.” And so on and so on. Sometimes after we eat, we feel a little bit like, “Hey, maybe I'll have a nap,” especially if it's carbohydrate-rich food. Somebody could make a hypothesis like this and invest a large budget in exploring this hypothesis – try to falsify it. While there is a common cause and it's temperature. When it's warmer, people are more likely to buy ice cream, but they are also more likely to go and swim. And if more people are swimming, more people are also, unfortunately, drowning (usually).

# Using causality to make decisions

12:41

Alexey

**I guess for some applications, we don't need to think about this causality – we just see a correlation and we train our logistic regression. I think maybe a good example is the famous “B feature” in the Boston dataset. It's also related to the color of skin as in the example you mentioned. I think it's the proportion of color of people in the neighborhood in Boston. Then if you train a model, maybe this model is accurate, to some extent, because it uses this problematic B feature. For some purposes, we will not think about whether it's correct to use this feature or not. But maybe it improves the performance of the model. Right?**

13:28

Aleksander

What you're saying is a very interesting point because, well, statistics is just enough for predictive tasks. Now, the problem starts when we want to make a decision. Right? If you imagine... There was a very famous case some time ago of a company called Zillow. They were actually doing what you described – they were trying to predict prices of real estate. Then what they also did based on those predictions was deciding if they wanted to buy a real estate and then flip it (which means renovate it and sell for a better price, or maybe even without renovation – I don't remember what their business model was) or not. This turned out to be a very good business for them for a very long time. As long as the distribution of all the variables in the background was the same.

Machine learning models are IID machines, which means that they assume that we have the same identically independent and identically distributed dataset in training and in the test (or in the real world) and this is not always the case. Causal models (depending on the type of a causal model) might address this if you have a lot of information – if you have a rich causal representation, you can address a situation like this. So if Zillow had a very rich causal model, they would not fall because of the fall of the market.

15:14

Alexey

**What happened to them? I heard they...**

15:17

Aleksander

They went bankrupt, essentially.

15:19

Alexey

**Because of their machine learning models?**

15:21

Aleksander

They bought a lot of real estate and then the prices went down and they were not able to take it.

15:31

Alexey

**So the models were predicting that the prices would go up, but they did not. Right?**

15:36

Aleksander

Yes. As far as I remember, that was the case. And they made a decision based on this prediction. When we're making a decision... Let me give you another example. When we're making a decision – usually, not always... Sometimes we might have a decision threshold, like in credit risk, somebody says something like, “Hey, if the probability of default is higher than X, we're just not giving money to those people.” This might be good enough, depending on what you really want to achieve in the long run – this might be really good enough. But we can think about another scenario with, let's say, marketing or churn. In marketing, we are interested in targeting people who will respond favorably (which means they will purchase). If we target them, this will increase the likelihood of purchasing. Because targeting every person is spending a little bit of our marketing budget. But there are different people, right? Some people might react...

Some people might just not care if we target them and they will buy anyway. Some other people might not buy, regardless if we target them or if we don't target them. And there will also be some people that will buy only if we target them because maybe they just feel a little bit special when we send them this discount, or we give them this personalized email. But there's also the fourth group of people who will buy from you as long as you do not target them. And if you target them, they get angry and they say, “Hey, it's too much marketing – too much of this bullshit. I don't want this. Goodbye.” And they just stop working with you. For this problem, predicting the probability of the outcome is not enough. Because we might predict that there is something like a 60% probability that this person will convert (they will buy from us) but now we don't know if they will buy if we target them or if we don't target them – or maybe their probability will actually drop if we target them. For this, we need to model something that is called “counterfactuals”. We need to model how they behave under the campaign and under no campaign. Then, by comparing these two outcomes, we can make a decision.

# Counterfactuals and and Judea Pearl

18:15

Alexey

**Counterfactual? What does it mean, exactly, in broader terms? It's a complex word.**

18:22

Aleksander

Yeah. “Factual” means something that actually happens in the world. And the “counterfactual” means something that is not happening in the world, so we change something. In Pearlian language – Pearlian comes from Judea Pearl, who is a computer scientist, who is the Godfather or even a father of modern causality (or graph-based causality) – in these terms, this means that we perform in a minimal intervention, as they call it, in the world. For instance, you have a red shirt today, a red t-shirt, maybe. I'm picking certain examples when you ask me and we could ask the question, “Would I pick different examples if you wore a blue shirt? Would that prime me somehow differently?” Now, if we have a very rich causal model, we could answer this question. This is not always very simple, but at least theoretically, it's possible.

19:30

Alexey

**But we don't know what would happen if I wore a blue t-shirt, right? That's why it's contrafactual. We don't know what would happen “if”.**

19:39

Aleksander

Yes. We will never observe it, moreover. We will never observe you in the same situation – you and me in the same situation – and you wearing a blue t-shirt instead of a red t-shirt.

19:51

Alexey

**So the other example you gave earlier was... Let's say we targeted somebody and we saw how they reacted to the advertisement, but we don't know how they would have reacted had we not target them, right?**

20:07

Aleksander

Yes.

20:09

Alexey

**Another example could be, let's say we have a recommender system and we show certain items. But there are other items we don't show. Maybe if we showed other items, they would have clicked, right? But they did not because we did not show them – and we have no idea what would have happened if we did that thing?**

20:30

Aleksander

Yes, exactly. That's an amazing example. Recommender systems – it's the same, yeah. It's the same structure as you noticed.

# Meta-learners vs classical ML models

20:44

Alexey

**So I guess our typical “classical” models like logistic regression, decision trees, XGBoost, whatever, classical neural networks – they do not really cover these cases, right? We don't know. In the example of targeting somebody, all we know is how people reacted to a campaign. We don't know how people who weren't targeted would have reacted. So what kind of models do we need to use to model this specific case?**

21:22

Aleksander

That's a great question. You are correct. Out of the box, supervised models do not have the capabilities to reason causally and there are many different types of causal models. But the one that I think is relatively the easiest to to grasp, and that also, behind the scenes, uses traditional machine learning models, is a family of models called meta-learners. The name is maybe a little bit unfortunate, because we also have meta-learners in traditional non-causal machine learning.

22:04

Alexey

**Like Ensemble learners? Is it the same thing?**

22:07

Aleksander

No. No, I think it's slightly different. But anyway, causal meta-learners are called meta-learners for a very particular reason – because they take regular machine learning models and they use them to produce those counterfactual worlds. Of course, they're estimated counterfactual worlds. Probably the easiest example of a meta-learner is a very simple meta-learner called T-learner. T-learner stands for two-learner because it uses two machine learning models. It uses one machine learning model to learn the response function for under no treatment.

Let's assume that the treatment is binary, so we do something or we don't do it. And the second model is used to learn the response function, which means mapping from the treatment and maybe some features to the outcome. It learns the response function and the treatment. So we have one model that learns response function under no treatment and another one under the treatment.

23:27

Alexey

**I just want to make sure I understand. We have two models – for the first model, let's say we're talking about this campaign example, when we targeted people with an advertisement. We have this pool of people who we target – our audience. We send them some sort of campaign and we collect the data. We know who opened the email, who ended up clicking it and we know who did not do this.**

**We have this dataset with the target variable. But then we also have other people to whom we did not send the campaign. We can observe what they do on the platform. We know that we did not send to this pool of people but they still may buy the thing we're advertising. So we just take all the other people and see who actually bought this thing in the end. And then we have two models.**

24:24

Aleksander

Yeah, so we take these two groups of people – one that we sent the campaign to and the other one that didn't receive the campaign – and we train one model on one group, another model on another group. Now, you also said about clicking emails and so on, so there is compliance, which means that if somebody clicked on something, etc. – it makes the thing a little bit more complex. Let's put it aside for now and let's just focus on those two models. So we take those two models, and then for any new observation, we make predictions using both models. It only makes sense if we also have some features that describe our population.

For each individual, we make a prediction using the “treatment” model and the “non-treatment” model. We take the outcomes from both models and we subtract the outcome from the non-treated model, from the treated model, and this gives us a quantity that is called a “conditional average treatment effect”. This outcome can be interpreted as a conditional average treatment effect only under certain circumstances, which means that the original data that we trained the model on has to be unconfounded (which means that there is no causal bias in this data) and this might be accomplished in two ways – either by randomizing the treatment in the training data, which means that we basically perform an experiment or...

26:16

Alexey

**In the same way as an A/B test, right?**

26:17

Aleksander

Yes, yes. It will be like a well-designed A/B test. Or the second option is to perform causal feature selection, which might be a little bit more difficult because we need to observe all the variables that can have impact on the treatment and the outcome at the same time and we need to exclude certain other variables that might have certain structural relation to other variables in the model. Basically, there are these two ways to do it.

26:48

Alexey

**What do we do with the results? We subtract one from the other, we get some quantity so I guess there could be three possibilities, like negative, zero, and positive. Right? [Aleksander agrees] What do we do in each of these cases?**

27:05

Aleksander

Well, it all depends on the setting. If we just say that the outcome is binary – they either buy or they do not buy. Negative, as I understand it, will be if somebody would have bought unless we'd targeted him. If you want to make an optimal decision from your budget allocation point of view, you should only treat people who would buy if targeted – that will be an optimal decision for you.

27:41

Alexey

**So only if it's positive, right?**

27:43

Aleksander

Only if it's positive and if it would be positive and if it would be negative otherwise.

# Average treatment effect

27:52

Alexey

**Okay. Can you explain again? [chuckles]**

27:53

Aleksander

[chuckles] I think we had different understanding, so let me expand on this. We should only target people who are positive under the treatment model and negative or zero in the non-treatment model. If we apply this ATE (average treatment effect) formula, that would be one minus zero, which means one – their outcome is one.

28:20

Alexey

**Okay. So the treatment model predicts that this person would buy, the non-treatment model predicts that this person would not buy – in this case, we go ahead and target. [Aleksander agrees] In other cases, we do not.**

28:35

Aleksander

Yes, in other cases we do not, because people who would buy under no treatment and treatment – it doesn't make sense to target them because it makes no difference to them based on our estimation, of course. We don't target people who don't buy anyway because it seems that it doesn't matter to them as well. And of course, we don't want to target people who are buying under no treatment and stop buying under treatments, because this is just generating a loss on both ends for us – losing a client and actually paying to lose a client. It's like paying to lose a client.

# Reducing causal bias, the super efficient estimator, and model uplifting

29:17

Alexey

**It's like the worst possible thing. [Aleksander agrees] You spent money but you also lost the client. I imagine it can introduce some problems. Let's say we take this model, deploy it, apply to the entire population (to all our customers) and start using it. We continue collecting data. The data we collect might be... We applied the model and now we introduce some bias by starting to apply this model. Should we maybe always do some sort of randomized test trial when we deploy this model? Or is it okay to go ahead and apply to everyone?**

30:03

Aleksander

This is a great question. The short answer is that if you use a model like T-learner, you might have certain problems. For instance, you can show that those simple meta-learners will have a little bit of estimation bias, which is different from causal bias. We assume that causally we are okay – our data was either randomized or structurally, the variables were chosen in a causally meaningful way. Then we still might have some estimation bias from those models. There are different models (other models) like double machine learning, for instance, that are trying to remove this estimation bias from those models.

Just last week, I think Lasso Lars published another paper where he introduced triple machine learning, which uses another piece of statistical information, let's say, to debias the model. He achieved something that is called a super efficient estimator, which means that it converges to the true value much faster with the sample size than a traditional estimator. So this is one thing. But I think what is more important is that sometimes it might be difficult for us to also get rid of this causal bias.

In this sense, if we are not sure that we were able to get rid of the causal bias, it will definitely be good practice – before the ultimate deployment – to deploy this model to have part of your customer base, if that's possible. This is something that I usually recommend to my clients – that we deploy a model to a part of the customer base and we always compare it to the baseline (whatever our baseline is). The baseline might be just a simple machine learning predictive model.

32:31

Alexey

**In this case, “compare” means we compare using some sort of business metric, like what the revenue that these two groups brought is, right?**

32:40

Aleksander

Yeah, we basically evaluate the policy. We can think about a causal model like this – this is often called “uplift modeling”. “Uplift” because we change whatever metric goes up when we use this causal modeling technique. So yeah, this is evaluating based on whatever metric matters to us. This might be revenue, this might be churn, this might be anything that would be relevant.

# Metrics for evaluating a causal model vs a traditional ML model

33:14

Alexey

**There is a question from Taras. Taras is asking, “How do we estimate the quality of a causal model if the metrics that we use are the same as for plain regression (traditional ML models}? Or are the metrics different?”**

33:33

Aleksander

Let's unpack this question. There are a couple of levels of evaluation of causal models. So the first one is regarding causal unbiasedness. Is there a causal bias in our dataset? Here, “traditional machine learning metrics,” or “traditional machine learning evaluation approaches,” like cross-validation, for instance, are not really useful. Why? Because the observational distribution (associational distribution) we can get the same associational distribution from different interventional distributions or different counterfactual distributions. This means that we can have different data-generating processes that end up giving us exactly the same observational distribution. So this is not very useful and we need other stuff to make sure that it works. One of the ideas that we can use is actually what we just discussed – deploying the model and looking at how it works.

Another way is using so-called “refutation tests”. So refutation tests try to falsify the causal structure within the model, which means these tests usually are changing something in the data, for instance, and they check if the causal coefficient that we are finding is also changing or not. This is a scientific method – Paparian scientific methodology. We're trying to change something in the world and we say, “Hey, if this model will react to this in a certain way, it means that it's almost certainly wrong.” Those tests cannot confirm that the model is correct, but they can falsify the hypothesis that the model is correct. Then we have statistical estimates. The question from Taras was, if I remember correctly, “If we use the same set of metrics for both models...” Yeah, we can do [that].

If we want to evaluate the policy, which means – if we want to evaluate whether we make better decisions based on a causal model versus on a non-causal model, then we definitely should use the same metric. Because if we use different metrics, then we are not comparing apples to apples. There's also a third dimension, which is the quality of the estimator itself. Assuming that the cover part is okay – we don't have any causal bias in the model – we might be also interested in (we should be interested in) seeing what the quality of estimation of statistical parameters within the causal structure is. And here, things like cross-validation and all those traditional metrics can be helpful because now we assume that we split the dataset into the training and test part.

We assume that they are IID. And we just want to see how well our estimators are estimating model parameters in the model based on the performance on the test set while the model was trained on the train set. In my book, you will find multiple examples of these procedures.

37:37

Alexey

**Yeah, I was going to ask about your book. Because to me, it sounded quite abstract. In general, metrics is such a topic that, for me personally, without examples and illustrations and actually going and trying to implement these things, play with them – they are just too abstract. What you're saying is that if somebody felt lost during this description, or wants to learn more about that, they should check out your book. Right? There, you describe in more detail all these things that you just talked about?**

38:12

Aleksander

Yes, definitely. In the book, we actually go almost from scratch. We start by talking about basic fundamental causal concepts and then we move gradually, step by step, towards machine learning methods, heterogeneous treatment effect estimation (which is another name for uplift modeling, let's say, plus or minus – this terminology is maybe not always consistent). Then we also talk about another topic, which is called “causal discovery,” when we are trying to discover causal structure within our dataset from observational data, or observational and interventional data.

# Is the added complexity of a causal model worth implementing?

38:54

Alexey

**From what I understood, these causal models are pretty useful and we should use them when possible (when needed) but they introduce an extra layer of complexity. Right? Right now, let's say you have a traditional model – you have just one model – you deploy it, you use it, and it seems to be working fine. But then if you start thinking about causality and causal models, then in the simplest case, you at least have two models.**

**It becomes two times more complex – your system becomes two times more complex. Is it always worth it to introduce this complexity? Or maybe there are cases when we shouldn't worry about causality yet and postpone this to some later point?**

39:43

Aleksander

Great question. Great question and a very loaded question. So starting from, “Is it worth it?” As to this question, it depends on what you are trying to achieve. If you're only interested in predicting something, and you say, “Hey, this is an IID case and I just want to predict if this will be more than five or less than five.” Then there is no need for causal models. Because causal models...

40:17

Alexey

**Maybe Zillow was thinking in the exact same way, right? “We're just predicting. We're not interested in...”**

40:24

Aleksander

Yeah, but they were making decisions, right? They were actually making counterfactual bets on reality based on a single model prediction.

40:32

Alexey

**But don't we always...? In most cases, we have a machine learning model to make a decision – to act on this decision. “Should we give money or should we not give money to a prospective client?” “Should we target somebody or should we not target?” In most cases, for classification, we want to make a decision. Right? “Should we put this email in spam or not spam?” “Should we write a recommender system? Should we display it or should we not display it?” In most of these cases there is a decision.**

41:14

Aleksander

Yeah. Always, when there is a decision and you also have some treatment that is under your control (which means that you can change something in the world) there is a potential of benefit for you in using causal models. But you asked me if it's worth it. I'm smiling because just two weeks ago, I got a message from my colleague, and he told me, “Hey, in my company, I just started analyzing the machine learning model that the whole marketing in the company is based on and I just discovered that for three years, we just have losses on our marketing.” This is making decisions based on the machine learning model. And it works. It's easy, because it's just one model – maybe. Probably they have more, but it's relatively easy. And everybody's happy.

Then somebody comes and they do the math and it seems that the marketing is actually throwing money away. And then he started analyzing this. He sent me a screenshot of his like, “Hey, this is my causal model and how it works. What do you think about it?” Every time you have this kind of a problem, I think it's worth it to go into causal models. Now, there is a psychological block, I suppose, in some people because they think, “Hey, we have some status quo. It works. I don't know, maybe even how it works, but it seems it's okay. We are alive. We are moving forward. So maybe, let's not touch it.” But then it depends – it depends, again, on what your goals are. What are your long-term goals? If you really want to maximize your gains and minimize losses in the long run, maybe it's worth it to just stop for a while and say, “Okay, let's see how it works. Let's see how much the investment today is and then what we can expect in the future?”

43:25

Alexey

**So we should think. Right? [chuckles] I think one of the things you mentioned previously is – when we deploy a causal model, typically there is a baseline. It's always a good idea to compare this causal model to the baseline. This is the data-driven way to see if adding one extra layer of complexity is actually worth it. Right?**

43:52

Aleksander

Definitely. I think even if you feel internally convinced that your data is causally unbiased, I would always recommend this. Because sometimes we just cannot think about something. There's maybe a little thing we missed. I think it's always an incremental work, really, to make things better and so on. But I think this is the same in life and in business.

# Utilizing LLMs in causal models (text as outcome)

44:26

Alexey

**There is one quite hot topic these days – these LLMs. Everyone is talking about LLMs – natural language models. In our podcasts, we were actually pretty late to the party. But recently, we had two podcast interviews that were about our LLMs. So better late than never. I guess LLMs are kind of hot because of ChatGPT. At least this is when I noticed them. Before, when it was just GPT-3, it was like, “Okay, so what?” But when I saw ChatGPT, it completely changed my perception of what these models could do. You recently gave a talk (I don't know how recently, but you did at some point give a talk) about causality and NLP and you tested LLMs with causal questions. Can you tell us more about this talk? Can you summarize it for us?**

45:21

Aleksander

Yeah, sure. Since this talk, a lot has changed in the research community and also, in the LLM space something has changed so... I will give you a summary of the talk and then also a short summary of where we are today. In the talk, we discuss the idea of combining natural language processing with causality, in particular, large language models with causality. There are a couple of ways that you can think about the intersection of those two areas. One is about using language models as elements of a causal system, perhaps as some kind of a feature extractor – in the role of feature extractors.

46:17

Alexey

**To me it's a bit abstract.**

46:19

Aleksander

Yeah. Okay, so let's think about a more concrete example. Maybe we have a situation where we have a program that aims to help people write more clearly. There's maybe a one-week or two-week workshop – people are sitting, writing, learning how to write more clearly, and so on and so on.

46:50

Alexey

**Without LLMs – just a workshop.**

46:53

Aleksander

Yeah. Just people.

46:54

Alexey

**There's a teacher. They learn and then, as the outcome of this workshop, they walk out knowing how to create a better copy – create a better article.**

47:03

Aleksander

Yes. Now, we might be interested in evaluating if this workshop worked. So if they are writing more clearly, really. If you want to do it at scale, it's challenging to engage (to hire) many people that will do the evaluation for us because they will need to read many pages of text, and so on, and so on. So we could potentially use an LLM (large language model) here and ask it for the clarity score for those essays. Then we could say – if we randomized the treatment, where some people got to the workshop and others did not based on a random assignment – we could basically compute the average treatment effect. So that's one way. This scenario is sometimes called “text as an outcome,” because the text is the outcome of some experiment. It also may be “text as a mediator,” which means that we have some treatment, then there's text produced and caused by this treatment (or some aspect of this text is caused by this treatment) and then we have some other outcome.

48:22

Alexey

**So if I understood correctly – there are two groups of people. One is the treatment group – people who went through the workshop. The other group is people who did not go through the workshop. Then each person in these two groups produces some text and then for each of the texts, we ask an LLM, “Hey, what's the clarity score?” And then we just compare using some sort of T-test or whatever, to see whether the workshop was actually helpful. Right?**

48:58

Aleksander

Yes, exactly. That would be an example of the scenario that is called “text as an outcome,” because the text is the outcome where we expect some change (some difference).

49:10

Alexey

**Okay. So what's wrong with that?**

49:14

Aleksander

Sorry?

49:15

Alexey

**What's wrong with that? Sounds like a good approach. [chuckles]**

49:17

Aleksander

Yeah, this is a good approach. Now we are talking about a scenario where we are using LLMs as an element – as an decoder or encoder within a system. We know that the system is causal. We know the causal structure of the system, and then the LLM is just one of the elements within the system that helps us turn this multi-dimensional entity that a text is, into some numerical summary. We can also use it in another way – maybe “text as a confounder,” which means that we have some treatment and we have some outcome and some aspect of the text is affecting both the treatment and the outcome. This is interesting because sometimes it's difficult to actually say, “Hey, this is one outcome. There's just one thing in the text that is influencing something. How do we extract this information?” Maybe its style. It's very difficult to quantify style and large language models can be helpful with this. Let me give you another example with text as treatment. So maybe we have a copy.

# Text as treatment and style extraction

50:45

Alexey

**Do you mean “text as confounder” or “text as treatment”?**

50:47

Aleksander

It will be text as treatment. I think it will be...

50:49

Alexey

**It's the same as previously, right?**

50:51

Aleksander

No, previously it was text as outcome and now it's text as treatment. So maybe we have a marketing copy and we have a bunch of people receiving this copy. And there's another version of this copy and another bunch of people are receiving this copy. Then we want to compare if they bought or they subscribed for DataTalk.Club. And maybe the copies are different. They have the same semantics – they're talking about the same stuff, but they just have a different style. Maybe just one is just like more...

51:31

Alexey

**I've just thought of an example. Usually, we have a newsletter for DataTalks.Club that has a sponsored slot – a sponsored block. Usually our sponsors give us some text but then what we do is look at this text and say, “Hey, we don't think this way of speaking will appeal to our community. Let's rewrite it right.” Usually, the marketing department gives us the copy and we rewrite it slightly so it doesn't have these words that marketing people use, because they're a turn-off for engineers.**

**So in this case, we rework the copy, and then we include it in the newsletter. But maybe what we should do is take the original one, take the reworked one, and test which one is better. Right? Because right now, it's just our gut feeling that the engineers (our community members) like the reworked version more than the original one. But it's just some gut feeling. We did not actually evaluate it.**

52:32

Aleksander

Yeah. That's a great scenario for an A/B test. Yeah.

52:37

Alexey

**But in this case, it's people who do this. But if we're talking about an LLM, it could be like, “Okay, there is a copy that a sponsor gives us and then there is an LLM that rewrites this.” Right?**

52:49

Aleksander

There could be an element that rewrites this. There could also be, if you take many different copies that are written in style A and some in style B, the LLM would be able to maybe extract the style property. Because the embeddings in the embedding space could be encoded in a certain way that could be useful.

53:15

Alexey

**Style can be something like “marketing language” or “engineering language,” right?**

53:22

Aleksander

For instance, yeah.

53:23

Alexey

**Formal, informal, right?**

53:24

Aleksander

Formal, informal – yeah.

53:26

Alexey

**Instead of people going through this and saying, “Okay, maybe this text was written by data scientists, this text was written by a marketing person.” Instead of a person going through this, we can ask them an LLM to say what kind of style it is. Right?**

53:43

Aleksander

Yeah, we could do this. We could also use it just to classify the style. Yes. So that's one way of thinking about this. We can have this text as treatment, as outcome, as a confounder, as a covariate, and so on. We could also imagine a situation – just to give you one more example – where we're interested in let's say whether gender predicts the popularity of your posts on social media. Gender is unobserved, but you have some description and we can assume that gender influences the style of your description. But it's very...

54:34

Alexey

**What do we mean by gender in this case? If text was written by a male or a female?**

54:38

Aleksander

Yeah. Male or female, whatever person identifies as. We might be interested in a hypothesis like this. You can observe a phenomenon like this in scientific citations as well. For for instance, it seems that from the observational point of view, female researchers are just getting less citations than male researchers. And it seems that this effect is stronger when those female researchers can be identified easily as female researchers – maybe there is a full name in the abstract and this full name suggests to another person that this person is female. In the Western culture, if somebody is called Stephanie, for instance, most people would assume that this person is female.

So we might be interested if gender influences the number of citations or the popularity of a post on LinkedIn, but it might be the case that the gender is unobserved, but there is a style of text for instance that is influenced by gender. To what extent is this realistic? Probably in certain circumstances more and in certain circumstances less, but this is just an example. Then, this text style – or the way that gender manifests itself in the text – is very hard to capture. Now, if we use an LLM – and in particular, we this fine-tune this model on this core task – then we can assume that it will learn the important characteristics of how gender relates to the style of text, even if gender is unobserved and then we can do causal reasoning on this system. An LLM in this situation is just like a very fancy feature extractor. So it extracts from the text anything that was related to the gender.

56:57

Alexey

**So in this case, there's a variable that we do not observe and typically what we would do without an LLM would be to build a model for extracting style, right? We would collect training examples, we would label them, we'd train a model with an LLM (just take GPT-4 or whatever). With the instructions, we can use it to extract these things. This is what we do. We extract the variable that we do not observe.**

57:30

Aleksander

Yes, we could try an approach like this. The problem here is also that the gender is unobserved. Then it might be complicated to label this in an automated manner.

57:45

Alexey

**So how do we know if the verdict of the LLM is correct?**

57:54

Aleksander

If we have causal structure, then we can do a smart architecture that will...

58:01

Alexey

**Ah, I see.**

58:02

Aleksander

Yeah. I don't want to go into too much detail because I think that this is very abstract without visualizations and other stuff.

58:10

Alexey

**But does the talk talk about that?**

58:14

Aleksander

Yes. This example is in the talk. And we discuss an architecture that is called “CausalBert”. You can find the talk on YouTube. It was a talk given on PyData Berlin 2023.

58:33

Alexey

**Okay.**

58:34

Aleksander

Yeah. And if you decide to watch this talk –I want to give you one more pro tip. In the library that we use, there was a bug in the code and I changed the implementation right before the presentation. So if you want the code that works properly, you can go to my book's repository and look for “CausalBert” and there's an implementation that doesn't have this bug.

59:04

Alexey

**Oh, we should be wrapping up. But I'm wondering if you have a few more minutes. Because there is another interesting question and maybe you can try to answer this question.**

59:14

Aleksander

Let me quickly check. I know that I have a meeting, but... We can stay for 10 minutes more.

# The viability of A/B tests in causal models

59:33

Alexey

**10 minutes. Okay. Well, this question, depending on how deep you want to answer – because the answer could take another hour. This is a question from Akil. “Can we use causal ML when we cannot use A/B experiments? And if yes, what kind of methods can be leveraged?” The first part of this question is “Are there cases when we can't use an A/B test?” I imagine that there are. In this case, “How exactly do we approach this situation?”**

59:56

Aleksander

Great question. Yes, there are cases where using A/B tests might be difficult for ethical, financial, or whatever technical reasons. If we can use causal machine learning in those cases, the short answer is yes. A longer answer is – it really depends to what extent you are able to fulfill causal assumptions. We didn't discuss this too much, but in order to make a causal model causally unbiased, we need to fulfill certain assumptions regarding which variables are observed. If we have some variables that are unobserved, we might use certain methods for something that is called “partial identification,” or we can perform things like causal sensitivity analysis and so on. And we might still get useful information out of those models, even if we cannot get a precise point estimate or precise confidence intervals. So the short answer is yes. The longer answer is, “in certain cases, when you cannot observe certain variables, this might be a little bit more difficult.”

In certain cases, it might also be impossible. But I think a great power of causal thinking, if you understand how those graphical structures relate to the problem of estimation and so on, is that you can very clearly state your problem and understand to what extent this problem is solvable. If you just drop a machine learning algorithm on your problem, you will get some answer. But if you don't analyze this problem in advance, you might not really know or understand fully what this answer is and to what extent it's useful for decision making.

Thinking causally gives a lot of clarity in this regard, especially – in particular, I'm talking about graphical models, or structural causal models that Judea Pearl proposed – I think that's a great tool for clarity. So even for people who are not planning to use causal inference or causal reasoning in their business problems directly today, the idea of understanding this is something that can give them long-term and very, very pronounced benefits.

# Graphical structures and nonparametric identification

62:38

Alexey

**You mentioned one thing here in your answer – you mentioned graphical structures. In the previous part, when we talked about LLMs, you mentioned an architecture. I guess in both cases, you mean a way of designing a model in such a way that it's clear which thing causes which, and then you have this causality chain or whatever. We probably don't have time to go into this but your book, I assume, covers this in more detail. Right?**

63:09

Aleksander

Yes, my book definitely covers this and your intuition is absolutely correct. Graphical model encodes the causal structure between variables. Let me take a step back and let's very briefly discuss the Pearlian definition of causality. The basic Pearlian definition of causality is, “A causes B, if B 'listens' to A.” “Listens to” means that if we change something in A, we expect to also see a change in B. We could have a deeper discussion on this, but I'm sure this is not for this format.

63:50

Alexey

**One of the examples you gave was about temperature, ice cream sales, and the number of people who drowned, right? [Aleksander agrees] Here if we change the temperature, then both these variables will change as well.**

64:03

Aleksander

That's what we would expect, yeah. On average. Of course, there are people who are drowning in winter and eating ice cream in winter, but in a statistical sense, we would expect that it's reasonable that they also go down. And then we express it in a graphical form, saying, “Hey, this is temperature and there's an arrow to ice cream sales and there's an arrow to the number of drownings.” And the interesting fact is that those graphical structures have certain properties, where we can identify causal structures without even looking at the data. So if we know the core structure itself, but we don't know anything about the data, we can already say something about the system.

This is sometimes called nonparametric identification. Based on a structure like this, we can say, “We actually don't need to observe this very costly variable and this one as well. It's sufficient that we just observe these three variables. And if we have these three variables, we can build a causally unbiased model.” And this is a great tool because some organizations that treat the data stuff very seriously tend to invest a lot of money in observing as much stuff as possible. And sometimes, some of those variables might not be very helpful in answering their most pressing questions. Yeah,

65:47

Alexey

**Okay. Thank you. I guess the to-go reference for everything we discussed today would be your book. Then, also the talk that you gave at Berlin PyData, which I missed. I was at the conference.**

66:01

Aleksander

Oh. So we were very close.

66:04

Alexey

**We were close. Maybe next year, we will meet.**

# Aleksander's resource recommendations

66:07

Alexey

**Can you recommend any other resources [besides your book] for people who want to learn more about the topic? I guess one of the things you mentioned was Judea Pearl's book?**

66:20

Aleksander

Yeah. If you're just starting, it's called the Book of Why. That's a great book for a starter. It's a great book if you're just starting and then if you want to go into more practical stuff, especially in Python, there's my book called Causal Inference and Discovery in Python. It also goes almost from scratch. I wrote it for people who have like 3–5 years of experience in machine learning and they want to learn about causality. But the Book of Why will also give you a lot of very, very nice motivation, beautiful examples from the history of science, how non-causal thinking failed, and how thanks to global thinking people were able to solve problems.

67:28

Alexey

**Okay, thank you very much. Thanks for staying a bit longer with us and answering this very interesting question from Akil. Thanks, Aleksander, for being with us today. And thanks, everyone, for joining us today – listening in and asking your questions.**

67:43

Aleksander

Thank you so much. Thank you for having me. It was a pleasure to have a conversation with you, Alexey.

67:49

Alexey

**Well, let's hope we meet next time at the next PyData Berlin or maybe some other conference.**

67:56

Aleksander

Yeah.

68:00

Alexey

**Yeah. Well, have a great week.**