

I: We are doing this interview in the research project [PRJ] and it's an [FND] which goes over the time of one year. Basically, in one year we are trying to see what's out there in terms of transparency, explainability and interpretability, maybe also understandability. Because after review of literature it's seemed like everything is very fuzzy in terms of what's what in which context and terms seem to overlap so we are trying to figure out how this is all connected. That's a very short description. Maybe after the interview of course you can ask me any questions but before (?) keep it short. You signed a consent form and filled out the questionnaire. During the interview please do not be afraid if I ask you to explain terms that are familiar to us but it's for our understanding and our understanding of others so we can get a better understanding of the term. The first thing is that we want to know a little bit who you are, what you do and in which you are an expert.

B: okay. who am I? I'm a scientific researcher at the [ORG] at the university of [ORG] and I've been a researcher since [JJJJ] What I research so for it's two main things. Apply (?) artificial intelligence to multiple fields and explainable artificial intelligence, XAI

I: can you describe in your own words what AI is for you?

B: AI is quite broad, it's not limited to machine learning. I would say that AI is the field in which we try to create synthetically which means it's going to be an algorithm that perform tasks that we human consider intelligent.

I: what experience do you have with AI so far? I'm speaking generally, broadly

B: I have professional experience in this case professional experience means that I work in the industry, in companies that implement AI for real world use in the (?) industry, in banks, and shortly in manufacturing. so that's for the professional experience and broadly three years implementing this kind of systems in the real world. that also means that I've got experience in research projects of applied AI with trains, cars, most of transportation, quality control, but mostly I've dealt with machine learning which is a small subbranch of narrow(?) AI.

I: besides the professional experience

B: so like my day to day every time use? I use AI for investing, but I'm relatively new to it, like two years or three years. so far so good and AI is a nice set of tools for it. I use it for tedious small tasks. Count my Pennies and stuff like that. just out of joy or laziness. I've used it for minor tasks that are repetitive like monitoring web pages to see if nothing changes or if something changes or not, for various random reasons to know when there's going to be a new call for something for papers or something like that. if I have just a big corpus of data I try to also use AI to analyze it out of laziness. obtain the relevant parts. but that's just personal views.

I: so it's more out of laziness than out of interest if you use it personally?

B: if I use it personally I do have some interest in it. I like it a lot and I try to learn stuff. but usually when I learn stuff I'm not directly applying it. it's after I learnt it that I end up applying it in one way or another and what forces me let's say to part from the step in which I'm learning to the step in which I'm actually implementing (?) is laziness. That's the perfect word.

I: That's interesting you mentioned that because that our thinking yesterday about we are working with the AI now and that's actually do something where do I need it and then I had this here all in German and then I realised you now normally I would've you know it was late at night yesterday it was like 10:30 and I still hadn't translated this and then I realised I still have to translate this and I realised I don't need to do this anymore. there are computer programmes that can do that for me and then I just put that form into DeepL like not even just the words but the whole document and it just you know british english, american english, whatever I wanted and that was the same. that was laziness. or lack of time.

B: But it has you saved also the lack of time or you are trying to work productive and then that also make it so even if you do not intend to use AI (?) simple translate. Google translate, I use it in my daily life. you can also see it in your emails. there is more (?) trying to (?) emails in which are dates which are social media that's also AI that you are using although you are not (?) it too much. when you are writing your cellphone then you have a small corrector that is also AI that you have there. you are more (?) than you realize. when your cellphone is trying to connect to the communication towers there is a small algorithm which is an AI algorithm in your phone that it's trying to decide when to switch towers and when not so it doesn't interrupt the signal for too long. then like there's the intended use of AI in which I'm actively saying that okay I'm going to search a (?) tool to solve this task and there is the unintended use of AI which is like (?)

I: yeah that's interesting. where do you think though is what areas of AI or the use of AI makes more sense? like to you? which type of areas or domains do you think AI makes the most sense to be used or be utilised?

B: I could say in which type of cases. the first case it's if you are performing a standardised task that a certain degree of complexity which means that like basic code would not be able to solve it but if you learn a little bit from the (?) data you can solve it and that can come in a lot of different domains like in the industry you go for let's say quality control or actually controlling a robot you can also use AI there. if you go into big companies that have different process of accountance than you have (?) scanners that scan just pieces of paper which are bills and then they learn to understand each one of the bills the structure and then transform that to a single format that they can use. in orthodontics when people try to design teeth for implants and stuff like that of even like surgery and bones so that you have a surface which is more let's say friendly with the body and has less rejection you also certain types of AI you know to insert basically prothesis (?) so there is actually a lot of fields. but basically it's either to optimize

something or to perform a repetitive task or learn more about something which means that's in a more explorative sense that you have a data that maybe you as a human are able (?) through the data but given the volume, and it can be like either because of the rows or columns, it is hard for humans to deal with large quantities of data so you try to do it you try to use certain types of AI to learn something from the data and then you study what it learnt from the data and you learn from it by proxy.

I: (?) it's a specific domain so it's the type of task that you want accomplished

B: yes

I: I haven't thought about that but yeah. last kind of provocative or more philosophical is do you use AI or do you work with AI?

B: do I use or work with AI

I: yeah

B: so far I use AI

I: why you say so far?

B: in order to work with AI I would have to somehow guarantee that we are working towards the same goal and that's sort of a corporative situation and aligning the goal of whichever AI you are using with the goal of your perform? class is actually quite hard so it's more like you literally using the AI or technique like please do this although you don't actually know the goal and it has not been a corporative scenario so far.

I: do you think that in the future it could be?

B: yeah, definitely, and not even in the future, like even right now in some restrictive areas you can work together with AI. let's say there's chess competitions in which it's not like humans vs humans but like human plus an AI vs human plus an AI. so that there you are working with it and not only using it.

I: do you think also in the production context?

B: sure, like collaborative robots

I: Now I'm kind of interested in what your experience is with AI in your every day work and that means what works well, what fails, what frustrates you or what have you heard from colleagues that's difficult or challenging.

B: so right now I work with (?) specific field of AI research which is narrow AI which is how to perform a single task extremely well and from there let's say some common tasks that I encounter on a daily basis (?) regression, clustering and then all of them

somehow you are trying to optimize something right? you have an AI, an machine learning algorithm which is basically a bunch of linear algebra (?) and then you try to somehow put in math terms in equations what's important to you and what do you want it to learn. let's go for a simple case, classification. then you have to communicate somehow or to state what do you want to classify and for that you either like create labels and that's fine or the labels that you create can be of different granularity and that can be simple labels like binary labels like this has it this doesn't or like pixel level labels and then you have to properly draw which parts or areas you are looking for and that's the way you state the problem. the way you state the problem is the way you state what's important and what's not and how it should learn and I think that's the most important issue on AI how to state problems. because no matter if it's something simple as a classification or if it's something a bit more complex like (?) optimization or (?) learning is the way you are defining how good you are at the task. and that's also going to be the way or the definition that the AI is going to use, to optimize (?). the reason for why this is important is because in general you have some inputs and outputs, that can be formatted in different ways depending on the task, but you have to have that somewhere. you have to first somehow guarantee that there is some type of information link inbetween the input and the output. it doesn't have to be causal but it can be like some sort of implicit correlation somehow. if you don't have that you are not going to perform a task and the more complex the relation the harder it is going to be to solve the task. that's why sometimes we just take the data and modify to a different domain to make it easier, to learn something from it. let's say when you are dealing with bayesian sensors than you better use not the wrong signal but you use a fast Fourier transform to see the histogram of the frequencies because it's easier to learn from there than (?) other things. so I think problem formulation is one of the biggest issues that my colleagues (?) also during my work because depending on how you formulate a problem you can learn easier or not.

I: that's from your point of view in your work, what about from a usage point of view, as in a user of AI. I mean you are more like let's say researcher developing, do you have any experience or issues that when you are using AI were frustrating or that you have heard from friends or family or colleagues?

B: So I'm going to keep my point. it's a lot about problem formulation. If you do not formulate correctly the problem, then the AI can learn to solve it, just it was not a problem that you wanted and then that you have some reward hacking. that has happened to me and then I'm trying something and just focus on something different that I knew I had in the data. and then that's complicated and you surely don't know and then this is the second thing some sort of interpretability for models, because at some point I was working in a freelancer job with a bank and then they wanted to try to predict some seasonal effects on something and wanted to create a small model that predicted somehow or estimated the risk of certain movements that they had there. once I did that and that person took it to their boss the question was like why, but not why we did it, but why it was making the predictions. and it's weird because in that part in the how to understand the models and how much we trust them, we apply higher standards to trust AI than we do with humans. the reason why we say this is because even in the industry

I have encountered in that same case there was a different guy that was doing a similar task and when anyone asked him, why did you make that decision, he was like years of experience. and an AI can't compete with that. and can't defense itself. so you can train a model, you can give measures of uncertainty, you can give a lot of stuff, it does not explain itself, sometimes it's kind of hard to get the trust of people.

I: that's wat I was actually looking at in my master's thesis because one of the things that's heavily impacting trust is like the experience that the agent has. and I was looking at this, but I'll show this to you later, I was telling the participant of the study, ok this AI has this much experience on doing this task and it actually had a big impact on the trust and confidence there was given to that.

B: But it's not only that. in this case, in a more pragmatic sense, there's a lot of places in the industry where the field of AI has evolved like super fast and than there's this (?) not enough good practices out there, but there's a lot of people that will use but not understand what they are using. and that also means, that safety measures for implementing AI are not well understood. something as simple as hey when you are using AI just try to check for domain shift or concept drifts which means that the data that you are processing or the task that you are solving it's similar to what it was trained on. because that shift, that's the real world. and in many cases people don't know that and than initially the AI works fine and than it doesn't and than they say like that doesn't work for anything. like we better not use it. and that's not correct, that's also a good utilization of the tools that you have, which right now a lot of people are not (?).

I: did you have any experiences with you said like this was a mind blowing moment for you? where you said like this is something that you can do

B: I had multiple of them with AI. The first one was back when I was in the end of my first year of university, so we are going back to 2008 and at that point in time I was studying control theory, like you have a signal and than you have some sort of equator(?) that works over a system and than you try to control a variable. and the professor was teaching like traditional methods for that and at some point I thought to myself like would it work if I would force it and than I didn't know at the time what was doing. I did a small algorithm to optimize something over somethings so I was learning a small model over something else and I was learning it just purely brute force like a stochastic (?). at the point in time I didn't know it was that. and it worked like out of the blue I didn't have to compute anything else just learn from the data and I like wow. than I learned a little bit more about that and what the hell I was doing and that was like wow, it can be done this way, that's super amazing. you don't have to trust a specific value you just can learn it from data based on a set of measures or matrix and some restrictions that we put on it. that was my first aha-moment. because that was quite powerful for me. if you can state something that is an optimization problem you can learn it.

I: so it's the learning part that is most mind blowing?

B: it was the fact that it worked in a problem of that complexity, let's say that smoothly. it showed me the power of how. the second aha-moment that I had, was when I was studying in France. there I was doing some modelling for (?) some like complex parts that go into airplanes for my master thesis and the professor that I had in that point in time said like why don't you try to use like bayesian analysis to know the energy that you are consuming in the machine and I like do you think that is going to work and he was like you may as well try and he was giving me a (?) and I was like sure and I try to train a small algorithm that predicted one from the other one and it worked. and at that point I was a bit baffled, like why would that work and then that was my second aha-moment. my third one was I was implementing something in the agrar industry, something that was meant to learn how healthy animals were, in This case chickens. I had a raw signal, I did a new (?) about the company at the time, I didn't know much about the whole topic and then I just learnt from the data, I created a model etc. and a structure (?) to give me way more than what I expected, like behaviour pattern of the animals, when were they stressed, when were they not, and that was something I was not looking for. that were just patterns.

I: how can you measure that?

B: In this case, there was just a small scale where the animals just jump in and out and the data that you got was the weight in the scale and depending on that you could estimate the weights, you could estimate how active the animals are, you can estimate how they are moving through the day

I: from a single data point?

B: from a single sensor and then you have like 600 to 3000 data points a day

I: okay because they are constantly jumping and it's the same animal?

B: but that's the trick. you don't only check the animals, you also check the raw signal. and then if they are pushing each other on the side like that will also generate signals and you can learn from that. I had no idea, I just analysing with my model, explored it and I found like okay what I was looking for, which was the main signal and some weird behaviours, weird patterns. and then I went to this company and said like hey your study is done but, and I didn't like it's not a bug it's a feature I didn't Knew that, I said but I have this issues and I showed them he cases and they were like wow, that's amazing and they explained it to me. It was like you know that happens because whatever they wake up early, they wake up like thirsty and then go to drink that's why the weight is going up this time and then this is the hour of most activity, that's what (?) like that, so actually I get more information from the real world (?) because the algorithm actually learn more than what you expected.

I: so that's what you were saying were the second case you were talking about that we can learn something from AI actually.

B: yeah

I: that fits actually with my next question and I would like you to describe the process AI goes through from the input to the output and how you would explain this to a lay person

B: a lay person? oh my. okay, let's (?) this a lot. so there's something really really important for AI and for a lay person. first you should be able to measure stuff, you should be able to measure the data, we are going to have as an input or what you can observe and what you wish to predict, right or what you can actuate(?) over. that's really important. this can be I mean as a human body what you can measure is your eye sight, your I don't know touch and then the way you can actuate it's moving your head and making decisions. so now you want to arrive from what you can measure to what you can actuate. first thing first is you have to be able to measure how good you are performing a task. and this is trivial for humans. like we know it, that's it. but it's not trivial for AI. so we have to explain it to it, to an equation, it's going to be a loss function. it's basically going to tell you if you did something right or wrong. once you have what you can measure, what you can actuate for or your output, and a way to measure what's good and what's bad, look at data. that contains your inputs and your outputs. in general, AI it's going to be a bunch of mathematical models stucked together that will take your input data and start to change it a little bit. it will start to find interesting patterns that are useful for the task that you are trying to solve. and then you aggregate that slowly until it makes the decision and it's going to be able to learn this based on the loss function that I gave it so based on how well it its performing a task and the inputs that you give it. a simple example: you have a surgeon that is trying to detect a melanoma so have this spots on the skin and the guy is trying to detect if it's a normal one or a bad one, you know what's your decision. if it's a good one or a bad one and you know what's the data that you have which is going to be the image of this small thingy. when a doctor is training he is going to look a lot of spots on skin, some which are good, some which are bad and he is going to start understanding a little bit better that for example the crispy shape of the border is important and he is going to use that as one of the basis for checking spots on skin in the future. it's similar with AI. with examples and the loss function it will start to learn slowly small mechanisms to perform the task that you gave it. and that's in general AI, I suppose.

I: related to that, why do you think it is so difficult for users to understand AI? as in non-experts.

B: to understand what it does or why does it?

I: both.

B: to understand what it does I think it's because a lot of applications that we use AI for we are quite imaginative, like we are quite creative. and even though 90 % of the AI that we have right now it's basically optimization processes, we are optimizing something. but in a lot of cases we stretch a little bit the definition of the task in order to like to do a

problem formulation, like alphafold, that's folding proteins, that's a complex task. and the way it works it's complex even for the guys that study that. because the AI experts have to stretch a little bit the definition of the task to let it learn stuff. it's similar with your cell phone. in a lot of cases, people do not realize the data that is available and how to properly formulate something as learnable task, as people do not understand how to formulate something that is a learnable task. they do not understand how a lot of tasks can be done by an AI and what it is doing. basically, part from the first lack of understanding of what's a data that's available in the world or in the devices that they have and how could some data relate to some task that they want to do. and as long as this first misunderstanding cannot be better explained, it's kind of complicated. once that is explained, (?) that has to be explained is that if there is any relationship in between some data and some task that you want to perform, you can use basic historical data to learn it, but you can learn it to an extend of information available to data. but I think as long as the first part and the problem formulations are not understood, it's kind of complicated For people to understand what AI does, because what AI does, is quite super simple right now, like you're (?) a function, you're literally approximating a function, all of the AIs, all of the weird ways of using it right now, is the same. you define something, you have a metric, you have an optimizer, you have a model that is learning that. and then you just stretch it in different ways.

I: it's difficult for people to understand that

B: but that's okay, because they will all be like okay than what about a car, self-driving cars, like what's the data, which are the functions and then if they don't understand okay, the data can be like you have camera that have images, you have sensors in the car, like in the engine and then the function is going to be, and of course, the more complex a task the more complex the problem definition, it becomes like a weird mental gymnastic that they are not used to doing and in order to do that, and they have to start to build it slowly, it's as how do you explain someone that does not know to code what code is. it's a set of instructions, but then they can't jump from a set of instructions to okay how does Google work. you have to understand small bits of it in order to construct over it. and as long as the small bits are not there, it's kind of hard to continue forward.

I: maybe we come back to the reason of why they can't understand or an explanation like a little bit later, because I have a part about this also. The next part is about the characteristics or requirements that AI must have. so, the question that I formulated yesterday is, what characteristics does AI need to bring to the table or what requirements must AI fulfil for good cooperation?

B: for good cooperation?

I: yeah or that you can

B: entrusted?

I: let's say use it well. we shifting a little bit more to how can I interact with AI and not from a technical point of view.

B: so there's going to be a couple of things. first, something that we don't use a lot right now, but that we could. a measure of uncertainty, basically how sure it is of what it's doing.

I: can I just ask you call it uncertainty, I've heard a lot before like accuracy. can you explain the difference that you see? because uncertainty is something that has a negative connotation, in my mind at least. so can you explain why you use that word?

B: yes. you have to guys, you task them with investing your model and the first one, he's like super sure about himself, he just (?) say like I know what I'm doing, don't worry about it. you trust him somehow for whatever human connection. and then he goes, he is as good as he is bad, so he barely wins 60% of the investment that he does, right, that's accuracy, like how well you perform a task. and he believes in that, that's good. you have the other guy. he says, yeah you know I'm doing the best that I can and he usually, he is (?) as he is good, he wins 80% of the investment that he does. but he also understands when he doesn't know if the case is good or bad. this means that he goes to the market and in comparison with the other guy that didn't have an uncertainty measure, so he didn't know that he didn't know, he just invested, every time. this guy looks at the market and it says like okay here I know how to make the decision, like in this case I know it and because of that, he makes the decision and in another moment he says like okay this is behaving weirdly, I actually don't know what's happening and then he says in this moment (?) we are going to invest because I don't know what happens. so the two are tied together, if you can somehow know how certain you are of a decision, you can tell me trust me on this or actually have no idea what I'm doing right now. and this is the basic of human relationships and the basic of our understanding from cooperation. if something it's super sure of what it's doing, but then you start to see that it fails but he was super sure, then trust starts to deteriorate. but if in contrast you have another case, in which there is a model that performs quite well or it can even be bad, but it tells you when it doesn't know what it's doing, then you like okay I know that this case he was not trying to deal with it, then it's going to fail less. and it's not going to fail less, because it has higher low accuracy, it's going to fail less, because he knows when the case is different than that he was trained on. and this on the sureness of the decision. and this is important for humans, because in some cases you ask an expert, hey what's this and the expert will say like seriously I don't know then trust my word here and it's okay, you want to still hear the opinion, but you can also ponder it in a different way.

I: I've also heard a lot when I ask this question about performance, how does performance fit in this for you?

B: so, a performance metric, it's a numerical value that tells you how good you are in a task. one performance metric can be accuracy, it tells you like how good you are performing in a task and That's the basic of how good you fit it to the data that you have

I: you mentioned there are few things, so, uncertainty, are there other things?

B: important for dealing with AI of course uncertainty, that's super important, because you want to know when you don't know or when the case is different than expected. and some sort of interpretability. either when using it or when designing it, depending how much you trust the designer. because you also want to somehow ensure that it's making the decision for the correct reasons.

I: what do you mean by interpretability exactly?

B: in this case I'm differently going to fail this question, but basically depending of the context you can have either the developer or the user should be able to see a little bit more of the data on how the AI is making a decision, but this data should be in a way in which it reflects its decision making process and it's understandable for humans. and then you can interpret what it did.

I: how does that differ from explainability?

B: when you have something which is interpretable, it's just making its best effort to be comprehensible. when you have something which is explainable it means that it's making the direct effort to explain you something. let's say in my mind, and this I can have the concept wrong, when you deal with interpretability you are trying to make it so that's like that you have representations that are somehow understandable by humans, but you actually don't care too much about humans, you just care about your model and about obtaining different representations from inside of the model that are useful. when you go into explainability, you also dealing with the mental models of the humans. that's a little bit more complicated, because then you also have to be careful with which priors the humans have and how well they understand different representations of the data that you are using to explain stuff.

I: we've mentioned uncertainty, interpretability, talks a little bit about performance, accuracy

B: then there is robustness, which should be a measure of how well it will perform in different cases or how less it will fail if different conditions change. I mean that tied to safety and security. somehow it can guarantee some of the things there you end up having to trust, like the humans and that trust in the thing or not, but I would say that for the moment. oh, accountability.

I: so, who is responsible for the decision?

B: who is responsible for failure or what is responsible for failure. I mean come on you don't usually account for when things go well. you just do (?). but in the case you want to actually want to use it, you should be able to have some accountability, but not only like in person, like in people, but also in the decision making process, like hey you failed,

why or what was the cause of the failure, and that's also part of accountability. because one thing is like ok who are we going to sue and another case it's okay how are we going to deal with this, like why did it happen and how do we make it so it doesn't happen again.

I: so out of all of these, do you think there are dependend on the field of application, use case or is it over(?)

B: so there's multiple things, robustness, so safety and uncertainty, those are mostly going to be mathematical concepts of some sort. and than that goes a little bit more in theory, but if you go into interpretability, that's going to be a little bit more tricky because that depends a lot on the problem that you are trying to solve. what I mean here it's let's say that you have one type of AI that is driving the car and than how do you interpret what it's doing, that's going to be trickier because than you have to be able to understand what's the process behind this model and to get some representations that are meaningful for its decision making process. but than you go to something different, a surgeon that is taking an xray to perform a surgery but than he has to detect some spots if it has a disease or not, than the way you explain that is going to be different then the self driving car. so when you are going to interpretability, explainability that depends a lot on the context.

I: so they will change in the nature but they will not drop out of the importance phase, they are all important

B: they are all important, except this ones have to be careful with are the priors of the context.

I: let's talk about transparency, what do you feel is transparency in AI or use of AI and is your system, that you are currently working with, transparent?

B: okay, transparency. I mean that's a term that I always find a little bit ambiguous(?) because that something is transparent does not mean that it is interpretable. so what the hell is transparency, that you have an AI and that you can somehow see what is happening inside of it, that you have at least insights of what is happening inside of it. that's okay, I mean a lot of methods are transparent, you just get the matrix in between some operations and than (?) that's transparent, the problem is that it's not interpretable and if it's not interpretable you can't explain it. so it's not explainable either way.

I: it's a little bit of a chain, like transparent

B: it's a little bit of a chain

I: interpretable, explainable

B: let's have two things, one which is like a normal neuronal? network and another thing which is let's go to partical physics and mean flow models. in neuronal networks you

can have in the middle you take an activation map, just a matrix of what's happening inside of it, so you actually get a (?) resolved that you can somehow see. that means that it's somehow transparent, like you can get middle steps that are tied to the decision making process, of the model. if you go to mean flow models, than that's more complex. it either converge or not. and than the whole how did we arrive to the convergence is not transparent. the process in itself, you cannot audit it in any way.

I: so that means that transparency for you is that you can look into it or take a snapshot

B: yes. what's happening inside.

I: that doesn't mean that will necessary help you to understand it.

B: exactly

I: okay. so, do you think there's more to it? to that term?

B: to that term transparency?

I: yeah, because for what you say now it's like okay, you can look into the model but it's

B: I mean there is a couple of things to the term transparency there, if you have something that is transparent that means that the decision making process of the algorithm in itself is a structure in a way. and that sole structure that you can say like okay you know we have this model which has this structure and than we are taking snapshots here here and here. that alone is useful. because than somehow you are structuring the process in someone else's mind and even is the results there the matrix you are not understanding, you know that it's a coherent what's this structure process

I: so what information do you need, which information needs to be made transparent to you so you can understand, explain, interpret, what kind of information? because you can make a lot of information transparent, generally. what is the information that needs to be made transparent for you to be able to do these things?

B: that's quite case specific. so if you go to neuronal networks, it's super simple. activations. this means you have a lot of small functions tucked together and than you want to know what's the result after a certain function. that's named latence pace or activation map. if you can get that, that's more transparent. that tells you something about how it's making the decisions. that's for most neuronal networks. if you go somewhere different I don't know a random forest, a different machine learning model, that's the AI. than (?) a little bit different, you have activations, you have something different, either the decisions that are made in each one of the branches or the (?) importance of each one of the decisions. there is a metric that tells you how important that decision is, so what has to be made transparent it highly depends on how the method or the model works.

I: but than you say it's dependend on the how and not the what, it depends on the process or model and not the use case

B: yes. so something simple. you are estimating something. let's say, you measuring the heart rate of someone. and than you want to know if that's okay or like that there is the chance that he is going to have a heart attack. that is a classical problem, that can be solved in many ways. one of them is neuronal networks. and than if you make it, that a neuronal network that helps you with that, that says like It's fine or he is going to have a heart attack. you take the signal and than you going to see after a couple of layers of the neuronal network an activation map. that somehow is going to tell you a little bit of how the decision making process is working. but you can also use it with something named a (?) filter or a partical filter and that's completely different, saying in a partical filter you are doing like a swarm of predictions, 2000 or something like that. and than what you want to make transparent there it's what happened inbetween predictions. and than you have just a swarm of like a distribution of small predictions and that's what you want to make transparent. because that's the thing which is directly tied to the decision making process of that model. like you have either the distribution of particals or the distribution of corrections and that's what tells you what's what. similar in a (?) filter, you go for how different the model of the system is from the prediction itself and that's what you want to make transparent.

I: do you think it's always needed?

B: no

I: no? give me an example.

B: sure. there's cases in which the metric for which know if it's need or not it's how critical the application is. for applications that are critical, you definitely want it to be transparent, you even want it to be interpretable. because that's somehow you are guaranteeing that it's working correctly, but if it's something which is not critical

I: what's critical? I mean, for me in my mind of course critical is something life and death threatening or something that has to do with injury of course. you think that's all or

B: that's definitely a tricky term. no critical is going to be that the decision that is made by the algorithm has a high impact. and this can be either on danger of human lives or on money or it can be in many other cases and there's

I: which is also very subjective

B: exactly, it's very subjective. let's say if I can have a conversational agent that just chat (?) with AI, so with this big model that is just telling you like hey let's play dungeons and dragons

I: that exists?

B: that exists

I: that's crazy

B: I know and it's even good. it's impressive. in that case, I don't care how it works.

I: you don't care?

B: I don't care, because for me like that's not critical. if it misinterprets my name once or twice like that's okay, I don't give enough value to it. but if you have the same model that is reading law suits and it's telling you if it's critical or not than that's different. you want to know the decision making process because that will have an impact. so there's two things. it can be critical for us as a society and than that's a decision that we have to make in that's actually not even a hard one. or it can be critical for you subjectively or personally and than it's up to the user itself.

I: and than to define what's critical or not, I mean for me it's more of an ethical and philosophical question agin, for as in a society.

B: in the society, sure. that's an ethical and philosophical question. like if you are going to make a decision like if you are in love, that's critical. if there's human lives like that can be affected, that's critical. if you have a monetary incentive on it so manufacaturing, that's critical. but than we get to the tricky parts. let's say the youtube recommendation algorithm, how critical is that? initially people were like yeah meh, but than you get that the way it (?) or the wayit chose content, make it so that more inflammatory conent gets shown more

I: Facebook news feed for example

B: for example and than you have issues as a society that initially was saying like you know we don't actually require to know how those algorithms work because they are silly until we say like secondary consequences or not intened consequences and we say like maybe the AI was more critical than we expected. maybe we want to take a look at it. so yeah that's something that us as a society it's tricky.

I: because first of all, when we decided to find some use cases and we decided to put them into high to low critically and than of course, first example was spotify recommender and everyone said like yeah that's very low critically, who cares. we don't need transparency for that. but in the end it's the same as with youtube algorithm or people depend on that

B: plus if you have something like a spotify recommender, for you it's meh but than you go to one of these companies that actually produces music and than for them that's critical and they would try to do their best of the effort to better understand that algorithm. it's the same for youtube now that there's a lot of youtubers that say like yeah

I'm going to do that because like the algorithm will censor me, that's critical for them

I: what I mean is also critical for the company let's say spotify, because if the algorithm is not recommending what people want

B: that depends. is it critical for them if there's no other competitor, like in this case there are, but if there's no competitor and there's no choice than their business model does not depend on it. it's like the guy that creates more filters for apps, so like you have a (?) face that overlaps or something like that, how does it work, do they care, no, they care that it mostly were (?) users are happy

I: we've talked about it a little bit, but let's come back to it. the main questing is how does transparency help you to understand, explain and interpret the process and decisions that AI makes. I mean there would be a lot of things that we already discussed but if there is something else you feel you need to add

B: interpretability, it's just a prerequisite if you want explain something. that means that someone must be able to interpret that in a way. that tells you that well, you probably don't want to give them matrices, probably want to give them something a representation which they can understand, that they can make sense of, than that's interpretable. because if let's say if the data that I show is in binary it's I mean it's probably not as interpretable

I: would it be transparent probably?

B: It will be transparent, definately, but not interpretable. than you go for explainability. you want to be able to explain something, so if you get it to be interpretable that's good, that's a first step. but than you want to be able to explain the representation that you have or that those representations are in themselves an explanation. for this you have to add a little bit more to the representation that you are using, dependeing on the case. because if you go for explainability, now you not only have to be careful that the representation is easier to interpret, but also that it takes into account the priors of the context

I: that's when priors come, because that would be my next question, into play when you need to explain something, you need to know what the person knows as in what kind of knowledge the person has

B: yes, exactly. because my algorithm can be Interpretable, I mean I have images and than I show you like you know we made the decisions because of this (?) of the image and something simple. metal casting, so you have a metal piece that has some pinholes, small holes in it and than you are detecting defects, you make an algorithm and part of the interpretability methods that you put there it's like you have the image and you just highlight the small spots there. someone that does not have the context will be like yeah that looks like the other part of the image that's fine. and tha you can interpret what it's doing, you cannot explain what it's doing, beause you have to be

careful of the priors of the person. and understandability is a promise that you never make

I: okay did I mention it, understandability?

B: no

I: but you came up with this?

B: but it's because people have asked me many many times like hey okay how do we it understandable and then like understandability is a promise that also involves the human that he will understand it so there's a clear possibility or high possibility that you are going to be able to create a mental model in the human involved regarding the whole decision process and that you are not only promising something about your method, about the process itself, but also about a person that is trying to make sense of it.

I: so, understandability comes also after explainability?

B: you have explainability and then you promise I can explain it to you. that you understand it is not my issue. that's explainability. but I promise I can explain, but you cannot understand it for them. when you say like this is understandable, then that's trickier because like

I: you can make it explainable but you cannot promise to make it understandable

B: exactly. and this is also regarding not only the priors of the person but also the priors of the priors of the person regarding the problem itself.

I: regarding this, we talked about this a little bit earlier also. let's say we need we want to make AI more transparent for people, let's say in a broad sense. what content needs to be there in training courses?

B: there are some things that we have measured and that's the basis of hey what's data, where is data and how it relate to each other. that's really important. you have to learn how to write before you can actually like write a book. so dealing with data is extremely important. then some basic understanding of what's approximation or what are functions. that you can (?) a lot. everyone does functional approximation in high school, not even high school, in school. it's that you have to make it approachable to people and say like you know like there's something that's approximation when you have some data here some data here and then you try to find the relationship inbetween those. once you have like better understanding of what's data, where does it come from and what's an approximation function you can start to compose and on top of that saying like you know there are typical problems that we know how to solve and then you go into like some examples of AI and then you start also to compose a little bit more. you say like you know a lot of AI is neuronal networks that basically are a lot of functions, paste it

together, that try to approximate the relationship in between data.

I: so it's a process with first taking the baby steps like with everything that you learn and then coming to the bigger things. there's not something you can tell people and they go like ah, that makes sense, I now get it.

B: it is definitely taking the baby steps. I compare it a lot with programming. what I mean with programming, it's initially like you go to programming 101. They say like we have variables. and then they tell you like one thing it's number, another thing it's a string and other thing is something else. and then at some point, you stop there. and then someone tells you like you know I have to programme to make a server to show this webpage and you are like that's blue, that's not a variable. that's a colour. so you have to start to build on top of that slowly. and say like you know like you have variables and then you can mix them together to create something else that is meaningful and then you start to build from it. similar here. if you don't understand like that a letter is also data, that A has a code behind. then you will never understand like hey can you read an article and transform that into mathematical terms to make it into the model

I: so it's also a question of mapping what is in the real world and what is in the computer

B: but that's the part of the data, that's part of understanding the data. that part of understanding you know when I'm doing this like I'm pushing a little bit there's physical (?), there's the force there, the pressure and there exist some devices and sensors that can translate that to numerical values, that later you can use. and there's a lot of people that don't know that. and then when someone says like you know the car is driving itself, they are like oh this is black magic.

I: let's come to the last part, which is trust. I'm going to ask a difficult question, but just go with what you are feeling. under what condition do you trust AI?

B: do I have to trust AI oh right that's a good decision. if I don't have any more data, I trust an AI if it's certain, it performs well and either itself chose somehow its decision making process or someone, a third party, can guarantee that the decision making process is correct. which will be by proxy, but that's a lazy explanation.

I: so you have an AI and that means you need some measures of uncertainty performance and you need a guarantee of the decision making process

B: if I'm going to put something valuable in the hands of the AI that's a lot, but the uncertainty measure is basically a way to ensure you like it was training similar cases, that it can go as slow as that. because uncertainty, basically it tied to it knows what its doing and that if you are going to tell it to a human like there's multiple ways of even understanding that in case you don't have the numerical value. the performance, there's two ways of looking at that. either the company that is providing the AI somehow promises that it's going to work whatever percentage of times or there has been test somewhere and they say like it works in this many cases, which for example you can see

that in medical applications. that they usually say like you know in this test there's 5% chance that (?) a false negative etc.

I: if you can explain it is it gonna help you to trust it more?

B: if I can explain it or someone can explain it to me it is going to help me to trust it more, definitely

I: because you know what's going on behind the scenes?

B: but that's a tricky assumption. because I have priors that allow me to understand that which is not the case for everyone else. so the interpretability part I personally understand what's happening behind, so that's useful information for me, which may not be the case for a lot of people.

I: what about understanding? so it can be explained to you, would if you understand it?

B: I mean, sure. but you usually don't need the interpretability or understandability part unless a) you are indirectly interested on it or b) something bad happens. no one complains if everything is good. and then if everything is good, like you trust the performance that they told you, that's implicitly you usually trust it or the studies that's fine. and uncertainty you don't care too much while everything is fine. when something starts to fail or when the risk is higher you want to know how uncertain your decision is. and when something fails it's there where you want to know. that's the chronological order of the steps, that I've seen in many cases. it's just that for liability if I'm in a company or if I have a company and I want to provide an AI service, it's in our best interest to explain it and make it interpretable at least for us, so that we know how it's working for accountability.

I: which is however not the case if a task is already critical and has not been used before. the system has not been used before and

B: on the contrary, the system has not been used before and you are actually creating it, you want to make it interpretable, to know why your stuff works, if it was not used before.

I: yeah exactly. and then that's what you need and you cannot wait until it fails. so where do you think is transparency in all of this

B: transparency is the bare minimum the minimum requisite so that there's a traceability to how the decisions are being made. and this is a little bit tricky, because you usually required transparency, interpretability, explainability, uncertain measure and a performance metric because that's the alternative to something else that we don't have in a lot of cases and that's a mathematical guarantee. let me explain to you. in control theory there's something which is a mathematical guarantee. what that means is the system works, we've proven it mathematically. and it's going to work unless we deviate I

don't know two times from what the system was designed to do and than that's a guarantee. that thing will work, like unless the physics are super weird. and that small mathematical guarantee can translate to unless the force is ten times higher than whatever it's going to work. but that requires like in the mathematical sense a lot of issues that you can not comply with in a lot of AI (?) systems. and than in AI systems, as you cannot mathematical guarantees, you usually have either proxies for understanding or statistical guarantees. and than the statistical guarantees are going to be the uncertainty and the performance.

I: how do you see AI in the future?

B: that's a nice question. it is going to improve, definitely. but there's two things that are going to improve. one, which is on the side of the AI and one which is on our side. right now. we have amazing techniques, but they are narrow techniques. and there's one weird thing that we've not done yet and that is why let's say you take Bob. Bob was raised in a village in I don't know, Madagascar, and than comes here and tries to talk with you. why is it, that you know more about computer than him?

I: as in the computer in general?

B: yes, the computer in general. you know that it is steel, some sort of metal, you know that it has somehow electronics inside that's the bare minimum. you know that it has a display. why do you think that you know that that he probably doesn't?

I: I as in coming from here know more about a computer than he does

B: yes

I: I'm not sure if that's the case anymore. Is it?

B: if there is a case anymore?

I: like if that's the case

B: sure, sure, I mean, this is just a random example, but sure. you didn't see anyone put the electronics there, hell you didn't see anyone like look at the electronics inside, someone told you. yes, you saw that information somewhere. the information was transmitted. so one thing that I think AI will do in the future, it's we still need a way, a generic way, to encapsulate information. so that multiple types of AI can consume it and create it. which is important, which is really really important, especially for generalization and for general AIs to be able to represent informational concepts in a better way. this I think is going to change. there's a lot of people working on that, I think that's one of the biggest issues right now. that's from (?) and having that

I: you're saying that this is a step from moving away from narrow intelligence that is only capable of one task or one type of area of tasks.

B: I mean, maybe. that's definitely a requirement. yes. the only one I don't think so, I think there's more things, but that's definitely a requirement for that, but second, that's also a requirement for being more efficient with data. if you compare a neuronal network with a physical model, the physical model is going to do amazing and the neuronal AI will try its best. it will not beat the physical model. there reason is there's a lot of priors in the physical model. the priors is like you know in the world there is the exist of gravity in this way tec. all of those are priors. that the network doesn't have. but if it has a possibility to search for different types of priors somewhere that's going to be amazing. so it will help us to be more data efficient somehow. and that is the famous common sense. common sense is just a bunch of priors that you have and from the human side I hope we will be able to learn a bit more how to interact with it, but the real world is not that interconnected yet. so there's a lot of things, a lot of systems that we have not digitalized completely. and that's literally like a chain that is not allowing us to improve, even more in large systems. let's say the trains in Europe could be managed by an AI. that's a simple case. the problem is that the systems are not that well digitalized yet.

I: and you feel that's something that we need to do or that's something that you wish would happen?

B: I think that would be something valuable, but there's people that are especially interested in not letting it happen. it's something valuable because it's to be more sufficient with the resources that we have. and I think that is really really useful. I mean, if we as humanity as a society are able to use our resources more efficiently, we can then direct them towards stuff which may have more impact. and AI is going to allow us to use our resources more efficiently. the issue that this has a big shock. especially society right now. but that shock is unavoidable. that's an ethical question. but it happens. technology evolves and so is our use of technology. I think we somehow have to be more interconnected with like digital systems and then AI can be used in a more large scale system. that's from our side. and from the AI side there should be a way to use or to extract priors from data and represent them somehow.

I: what do you think about the interview or which aspect of the discussion were most important or most interesting for you?

B: i like the interview. it was an interesting experience. it caught me a little bit out of guard. the let's say when you mentioned transparency, interpretability and explainability. because as you may know those are terms that have been defined multiple times and it's a bit fuzzy here and fuzzy there and no one really let's say agrees with stuff.

I: so it caught you of guard because you realized that's an issue?

B: no, I already knew that it was an issue.