A1  
 Exactly, very cool that you are here. Exactly. Yes, we have already done a brief introduction. Now a bit more specific. So, what is your role and what are your experiences working with AI or ML systems?

I8  
 Exactly, I work at XXXXXXXXXXX and centrally manage the AI platform there. Exactly, we have been working on this for a little over two years now. At the end of last year, we completed the first version of the project, encountered a few challenges, and are now going back into review and refactoring the whole platform. The goal of the platform I oversee is to enable machine learning development in a centralized way, so that we can create synergies between departments, meaning we can say that we have a core competency in machine learning and it then serves many business units, either with external service providers or internally. Exactly, this has changed a bit now. We have now added generative AI, which is why we are also doing more refactoring, because we want to maintain or create the connection between machine learning and generative AI. Many products are offered and now the task is to hold all of that together and support the projects for machine learning, which has been basically my responsibility.

A1  
 Okay, cool. Yes, Gen AI is something we have already noticed a lot in the first interviews, that it is kind of shaking up the AI field from behind. When we planned the study, ChatGPT was already on the market, but it really came out last year and more and more companies are now using it. So, not meant critically, but it is always exciting. I think if we had conducted this study two years ago, it would have been completely different. Back then, AI would have been just image classification or something like that as an obvious task. What are your concrete, daily processes with it, as far as you can communicate and are allowed to? What is your daily business regarding this platform?

I8  
 Exactly, the daily business is that we have different projects, now I can report best from that. We have a project in the power plant sector, where we want to use predictive maintenance, working with an external service provider. I work there in an advisory role, ensuring that we at XXXXXXXXXX also understand what is happening, so that we do not end up with a black box and can later transfer it ourselves. A lot of it is about our data, as we are largely in the critical infrastructure sector, which is being expanded, so that energy providers are increasingly considered critical infrastructure. This means we cannot hand over data freely. So we have to ensure that service providers can work safely in our systems without seeing too much. My task is to bring understanding of machine learning and later also help onboard service providers into our systems. Then we have a project in the mobility sector, predicting bus occupancy, where I analyze and visualize the data operationally. Now it is about using ML models to predict how full a bus stop will be, and more generally, the platform itself, meaning setting up the infrastructure, creating the framework, enabling machine learning operations, so that we can also use them for language models. So, it is framework consulting and a bit operational, although the operational part is currently handled by student assistants and I am less involved in that.

A1  
 Okay. Super interesting. So it is very project-dependent. And it sounds like there are many interfaces, some very different from each other, that you try to cover in your work. For context, I also work in an industrial transfer project at XXXX, they build water pumps, maybe you have heard of them. They are also in XXXXXXXX and I am trying now, in the third year, to integrate AI into their company along with others. We have a strategic partnership. We jump from project to project every few months, sometimes doing lots of data and business understanding, exploration, and prototyping. It is very different depending on which people you work with, and what you do concretely varies a lot. They are also building an AI and data platform, and it is very interesting. There is a lot of change, especially with systems that have existed for a long time, trying to integrate AI into them.

I8  
 Yes, that is very exciting. There is no single answer. Every manufacturer now also offers their own solutions. This shows that language models at the end are just an interface. They are much simpler than ML systems. I was hired last year and the plan was to eventually build a central data platform and take over the topics. Now with AI we are creating a separate department with four people. This brings a completely different financial background. In my previous role at a machinery manufacturer, you could say that certain algorithms had reached a certain economic efficiency, which somewhat sold the system. But that is just a fraction of the investment going into systems now with language models.

A1  
 Absolutely. There is a lot of change right now, and it is exciting to observe. What are the biggest challenges in these daily processes? Of course, it depends on the project you are currently active in, but can you break it down a bit? What are the biggest hurdles you encounter on a daily basis?

I8  
 Exactly, so I would say that currently a big hurdle is that we, for example, are city networks. We have many departments and we have different service providers. So we have relatively many people who actually have to work on the platform somehow, and the challenge is to separate them systemically so that they don’t really know anything about each other, but to open central components at the right points so that they can create synergies again. So, to restrict access in the right places, but also to be open enough in the right places so that there is an argument for a central platform. It gets interesting when you also consider permissions at the storage level, documentation for projects, or dealing with service providers. We are currently in the process of gaining experience. Data integration is always a topic. Compared to machine learning, I would almost say that in terms of time, not effort, it’s faster to have a model up and running and evaluated than to have the systems properly integrated and the output placed in the right location, because that requires so many agreements. There are different software providers, the interfaces don’t exist yet, and a lot of cleanup is required in the systems because things haven’t been thought through that far. Some systems are ten years old, so usually you modernize them first before integrating. The biggest challenge is really these different work areas, workgroups, and integration into the systems.

A1  
 These are all very technical things, so of course it makes sense that these are the everyday challenges, and definitely data and business understanding, that is always the main effort. In every industrial collaboration I have worked in, it was the same. Reaching the point where you can actually train a model is already about 75 percent of the way there. I think it’s a bit different when you imagine monster projects like LLMs or image classifiers, because then we’re talking about weeks of training. That’s not something you just train and evaluate quickly; the process itself is very extensive. But with industrial data, where we have, I don’t know, fewer than 100,000 samples and maybe fewer than 50 or 60 features, training is not the main part, it’s getting to the point where the dataset is curated or a live connection exists. Moving away from technical systems and infrastructure topics, how is communication about AI with you? I can imagine that at interfaces there are very different people with different professional backgrounds. They may be more or less interested or even critical of AI. Is that also a challenge for you?

I8  
 Exactly, yes, there are actually different points. At the beginning of the year, we said that an AI department would be established centrally. This was also because so many departments, we are about 1400 people with countless products, work relatively independently and only follow limited central guidelines. They naturally have their own budgets and ideas. With the Gen AI topic, everyone wanted to do their own thing and it wasn’t moving fast enough. Eventually, the CEO decided to create a central point, and all AI projects in the language model or ML area, which already existed, are now managed centrally and go into the company from that central point, because there was initially so much interest from different departments to do their own thing. This was to prevent uncontrolled growth. The message was basically: cut, we will approach you when we have something. The second part about fears is a bit difficult. The voices that dominate tend to say it’s good. I think it’s usually the 20 percent who are the loudest. The skeptical people don’t currently have a channel, I would say. There are enough who are cautious because the tasks will change significantly with AI. At the end, we see the works council as a kind of channel, which currently speaks very positively about AI. I think their role should also include raising critical points, giving people a voice. If all the managers only say everything is great, it becomes difficult.

A1  
 Yes, a certain technological skepticism is healthy. But it’s interesting that in your company this criticism doesn’t seem to get much attention, and everyone is rather hyped about it, saying, yes, “let’s go” chat models and such. If you ask people in the general population, the results would definitely be different. Many are afraid that AI will take everything away and break everything. Maybe not entirely wrong. Do you have any ideas where we, as machine learning scientists working on fundamental research and developing this technology to make it usable, could improve? Something that would make our work and the transfer of knowledge much easier?

I8  
 Yes, in machine learning, these are like labeling approaches, deciding which model to initially use. That’s knowledge transfer, which is exciting because it tells you: I have a problem, this is roughly the scope of the data, and this model has worked well. That saves effort on our side. Or currently, I’m a fan of AutoML, where some of the work is done for you, or having libraries that collect models for specific problem types that can be directly used, which I find exciting. In Gen AI, explainability is key: reviewing what’s behind the proprietary system. Since these models are on the market, and we won’t host large language models ourselves because it’s too expensive, adding more scientific explainability into procurement systems is very interesting. For a company, it’s also valuable to see how it actually works.

A1  
 Exactly, so explainability is definitely also a huge topic in research and it happens on different levels, whether it is explaining the model in general, how it works. Of course, an explanation helps, like, these are the neurons and you calculate them together, in the end it’s all some huge matrix multiplication, and a result comes out. That’s not really an explanation that is useful to a human, especially if they don’t have a technical background. So a lot happens in trying to make this black-box model a bit more transparent, and opening it up I would actually exclude a bit here, because that’s not really what we do with the labels at the core. Sure, with the labels it is, yes, that’s why we are doing the study, of course. But I actually have to ask again, would this be something useful for you in your everyday life and processes, or do you rather feel it’s a good idea but you’re not the target group, because I’m a bit unsure about that right now.

I8  
 As it’s developing right now, I’m no longer the target group. So it would rather be kinda the people who actually implement it operationally. Right now it has more to do with the framework, and I am relatively active in that at the moment. So the target group would rather be those people, saying here, you have the label, these are good models, there is a catalog, kind of with the models that work well. I can imagine that working well, but I would also say that I’m not the target group for that.

A1  
 But you’re saying there is a target group and a need for it, and you would maybe manage such projects, and would you want reporting from the engineers who worked on solutions, concretely, and for you it might be helpful to get these labels back as feedback, and then maybe on a higher level you could make decisions based on them.

I8  
 Exactly, if you see it that way, then it would again be about creating synergy effects across projects. In the end, in ML tasks, the service provider often starts from scratch with new data and will have their own experience, and for me, at least this is how I understand it, it’s a kind of central catalog of which models work well and which don’t. That’s actually a good starting point in the end.

A1  
 Yes, cool. That is also exactly what we hope for. So many thanks for the feedback. That is exactly the target group we have. At this point I would actually go into the labels in more detail and share my screen for that. I just need to start the PowerPoint presentation first. I always forget to do that before the conversation because PowerPoint always takes forever. And then I’ll share my screen. There we go. You should be able to see my screen now.

I8  
 Yes, now.

A1  
 Perfect, very good. Without me explaining in detail what you’re seeing here or anything, can you give me a spontaneous reaction to the label? What do you see, what stands out, what are your thoughts?

I8  
 I had read the description before, so kind of like an energy label for washing machines. In terms of the presentation, it still fits quite well for me.

A1  
 Sure, we had already displayed a label on the website beforehand, so it’s not the very first spontaneous reaction, but maybe you can put yourself back in that moment, what was your first thought or...

I8  
 Yes, well presented, looks modern, and I also like the symbols at the bottom. You can see the symbol and the number directly underneath it, and even without reading through, you already get a sense of what they mean, so I have an idea of what you are trying to show. Exactly. ImageNet, overall, there is a good overview.

A1  
 Does ImageNet mean anything to you?

I8  
 Not directly, it’s probably more in the direction of ResNet or something. I don’t really remember ImageNet.

A1  
 Mhm, okay. I can explain that quickly. ImageNet is actually the dataset that was presented in the 2012 challenge. The specific model we are looking at here is a MobileNetV3 Small, but there are also different ResNets that were trained on ImageNet. So the ResNet architecture would practically be a competitor, a counterpart to the model that was labeled here.

I8  
 Okay, that wasn’t clear to me, but I find it even more interesting that you have the dataset behind it. For time series data, there is also a test dataset for that, so it’s kind of more interesting. It wasn’t clear to me before that the dataset is actually listed at the top.

A1  
 Yes, it also depends a bit on the expertise of the people looking at it. Exactly, go ahead and continue explaining. So yes, just as a clarification. It is MobileNetV3 trained and evaluated on ImageNet, and you could do the same for ResNet.

I8  
 Yes, it’s always a bit like that. For me, you have to know the terms, like ImageNet or MobileNet, which is the model. That’s interesting, and then the comparison across the data with the accuracy. Right. The power draw per inference is the energy it takes per execution?

A1  
 Correct, exactly.

I8  
 Okay, and that’s the A100. I’m not really sure, is that a graphics card?

A1  
 Yes, exactly. The A100 is from Nvidia, it’s a graphics card. This is practically a cluster node with eight GPUs and a TensorFlow installation. I usually affectionately call it the environment in which the model was evaluated. Hardware and software are very complex and cannot be summarized in a word or two. For example, I once had the funny moment where I was wondering about MobileNet, and the accuracy was about 20% worse than in the paper. I thought, what’s going on? That can’t be. All other models matched the reported results. Then I updated TensorFlow from 2.3 to 2.5 or so, and suddenly the accuracy matched. This means the pre-trained weights of the model were faulty, or there were implementation errors because the model is provided by TensorFlow via Keras. A model is not exactly the same in speed or resource efficiency; it always depends on the environment. That’s why it is represented here. Refrigerators have it easier; they always draw the same amount of power, no matter where they are. Although, it probably makes a difference if they are in the basement or a 30-degree attic like mine.

I8  
 That would make a difference. What I don’t fully understand about the label is the A, B, C, D, E part. I think it’s just that I don’t have enough experience with it yet. With the numbers at the bottom, you kind of get a feeling because there’s a number. But the A, B, C is harder to interpret.

A1  
 Yes, I can explain briefly. Each metric reported at the bottom is divided into five bins for relative comparison. For the tested models, green means it’s among the best, red among the worst, yellow somewhere in the middle. This is based on all evaluated metrics, actually more than the ones shown at the bottom. They only showed the most important ones, because if you show all ten, nothing is visible on the label. This results in an overall scoring. The interesting part is that the weighting of these metrics can be adapted to the specific use case. You could say, honestly, whether it has 80 or 70 percent accuracy doesn’t matter, but we have a strict constraint that the model must return a result within one second. That changes the scoring a bit. The analogy with refrigerators is different because they don’t have this kind of prioritization. It’s easier to say what is the best model. With machine learning, it always depends on your constraints, use case, and company priorities. If the company says we are green, they might care more about resource efficiency. On management level, they only want to see how much money it earns or costs. But that is hard to quantify, as you probably know.

I8  
 Yes, it’s interesting that now you related it to image processing. But how does it look for a random forest or very simple cases? We don’t always have these big models that make an impact.

A1  
 Yes, for smaller ones, so in principle you can of course issue such a label for a Random Forest as well, I have also done that in a paper. So it always depends on the task you have. For a dataset like ImageNet, a Random Forest simply won’t work, because the task and the dataset have a complexity that requires convolutions. You need some kind of efficient parallel processing of image data, you need specialized things like convolutional layers in the network so that interesting patterns can be recognized on a visual level. A Random Forest that just randomly splits images is simply not suitable for this use case. In that sense, a label would probably not make any sense. You could do it, then it would say something like, I don’t know, maybe 5% accuracy or so and probably a much higher power consumption because it’s not properly parallelizable.

I8  
 Exactly, if you take this standard dataset.

A1  
 Exactly. But if you had a tabular dataset from industry, you could do it just as well, you could just run 30 models from an ensemble through AutoSK Learn or something, profile and benchmark them, and visualize the results the same way. So in principle it’s completely transferable and actually possible through the software I wrote in my dissertation. You can just load a Pandas DataFrame and it will output these labels for you. Exactly. A big factor, which you actually just touched on a bit, is of course that it allows somewhat hard-to-understand papers or abstract implementation details to be abstracted and made more accessible. But in particular, it also enables the comparison of different labels. We can also talk a bit about what stands out to you here.

I8  
 Exactly, it’s relatively nicely presented. You can also immediately see at the bottom that the accuracy is higher, but again at the cost of power consumption and runtime in the end. About robustness, what does that mean?

A1  
 Yes, I explain that, I think, in almost every interview. Accuracy is just given the test data, the accuracy. For robustness, you take the test data, slightly modify it, add noise, distort the images a bit, try to artificially create changes in lighting or exposure to estimate how much the accuracy changes when the data is no longer exactly like in the test dataset. So it’s a way to test the model’s robustness against external influences. This is particularly relevant in autonomous driving, for example, because you want to ensure that if the camera lens or the glass in front of the camera fogs up a bit, the accuracy still holds. And if you only take ImageNet, those are all very clean, orderly images, almost everything. That’s why robustness testing and robustness training are a big topic, people try to do this already during training to make the model more robust from the start. Exactly.

I8  
 It would also be interesting to see how this works for energy labels. You can actually calculate those, there is a certified formula for that. Would your goal then be to have some kind of central formula prescribed, or is it rather meant to be a label for companies or within companies?

A1  
 I think it’s a bit of both. As I said, fundamentally this is basic research, and of course it’s not industrial standard or fully implemented in practice, but that’s partly the point of the study: to explore if there is any demand for it, if it helps people, if it makes sense to do it. Now with the AI Act, companies are required to think a bit when developing AI products and not just freely try to make as much money as possible. Such instances will come into play, and we want to see if an abstract labeling like the one we developed could be helpful and perhaps work like a kind of certificate. And then of course the whole process of how such a label comes about would have to be followed. I mean, it’s described in my paper, you can read it there. Just like you can find information about why a vaccine product comes to the market, for example. There are studies and all the background information, which is actually public. But exactly, it’s a big question mark how to best do something like this. So I can only defer the question.

I8  
 Yes, but just fundamentally, having some kind of baseline to estimate from. Mentioned, of course, exactly now, so I think the restriction to machine learning has become a bit clearer. I think with generative models from others it will somehow be a bit more difficult to explain. But maybe that was also an exclusion criterion for some models, right?

A1  
 The properties we reported below, you couldn’t take them one-to-one, because you don’t talk about accuracy for language models, but of course there are metrics to roughly estimate the model’s quality. A lot is happening in that direction right now, measuring language model quality beyond just asking people empirically if it seems plausible. I think on one hand, the usability of ChatGPT and the fact that it’s so widely used is already a measure of quality, so it indicates that it works pretty well, but of course there are always edge cases and corner cases where it doesn’t work at all. So I would say you can label all that, but the thing is, for example, OpenAI doesn’t release its models publicly. That means you can’t profile them on any hardware yourself. Their paid models are only available remotely with an internet connection and a subscription, so you’re somewhat dependent on OpenAI providing such information, because you can’t measure it yourself. And so far, they haven’t done that. So overall, except for robustness, would you say it’s all fairly understandable what is shown on the label? Sure, maybe not fully transparent in detail, like how the scoring comes about. But if you follow the QR code at the top and read our paper, you could figure that out. Do you feel that there is information missing here, where you would say, there are these metrics I always look at, or that would be really interesting to me and is completely missing?

I8  
 Yes, it’s hard for me right now because generally it gives a good overview. I have to let it sink in if anything else comes to mind, which would then be more in the doing, what I might be missing, like checking more often and comparing, what are the benchmark values, where it should go. Okay, then I’ll think about it more...

A1  
 We touched on this a bit earlier with language models, where you said explanations or the explainability and transparency of the model, if you could somehow measure and communicate that. You brought that up earlier, I think, right?

I8  
 But the label does give a good overview. If you look a bit further, it would be about false positives, like checking the precision-recall curves, seeing what it looks like in that area. Accuracy is still relatively simple. I also don’t know enough about imageNet and how accuracy is measured in the image domain. With that dataset, if I were to use labels, I would look at the dataset, see what’s in it, compare it to my dataset, and then see which model performed best according to my key criteria. A description of the dataset in the background would be nice, like what properties the dataset has. Maybe it would be faster because in the end, I wouldn’t dig into it; I’d check if the description matches, I know my dataset, and I would compare. That would be interesting.

A1  
 That is kind of hidden behind the lower QR code. The lower QR code links back to the paper where the model is presented, and it also states which dataset was used. So you can go as deep as you want. It’s really a trade-off, how abstract you want to stay. The moment you put a wall of text, a textual description, on it, you’re on another level of representation. But that’s not a criticism, it’s just why we are interested in what might still be missing from this label.

I8  
 Yes, exactly, when you think about how to work with the label, if I start with ImageNet, I get a good overview. Then I would maybe start with the first model, pull the library, try it out. In the second stage, I would look closer, and the QR codes show that quite well if you check them. You need to provide an entry point, and I think it’s represented quite well. For certain applications, it might not fit, and you have to adjust the weighting, that’s clear. Hardware might change, but you get a rough direction, so you have a basis for decision-making. Then you move away from relying purely on experience or gut feeling for which model to use. You might otherwise miss a model that consumes less energy and performs better. So it might give a better overview if one model is in one library and another in a different one. If you don’t read every journal or follow every publication, you might not notice it. Especially in industry, you don’t always look left and right.

A1  
 Yes, definitely. Or rather, definitely not. Exactly. No, definitely. Thank you so much for the insights. We have already talked a little bit about this. What about other sources, where would you get information from? So labels are not really on the market yet, but which communication or representation forms, information sources about machine learning models and algorithms do you know?

I8  
 I have worked a lot with IEEE, actually, looked directly for papers there, and otherwise I usually find Medium. I mean, open, yes, open platforms, I find them relatively good there. They don’t have exactly this publication character, but you can find well-prepared information with good examples. That’s basically where I look. Those are basically the sources, IEEE not so much anymore, yes, I usually check Medium, but that’s exactly the place I’m not so active on anymore. But yes, publications are, I think, still the main part in the end.

A1  
 Exactly, we have put together a small list, with publications of course included, and a blog post here from Towards Data Science, which is on Medium as well. There are also Model Cards, which are used on library platforms like Hugging Face, originally a Google idea. Basically, a small mini-manual for a model, very similar to IBM’s factsheets, although those are patented and only provided for IBM products and customers. Not publicly available, so to speak. But anyone can write Model Cards; it’s basically like a readme for the model, the main takeaways. Then there is of course library documentation, and something that not many people know but I find quite practical: Papers with Code. It lists all available papers for models in which they were used, the same for datasets, and it even has small leaderboards and graphs where you can see, okay, this dataset was mined over the last 15 years and the accuracy curve has steadily increased. You can click on each point to see which paper it was from. I find that quite interesting. Here as an example for MobileNetV3, the corresponding paper, also on Arxiv, and the Model Card as it appears on Hugging Face. On Papers with Code links you can see the list of papers where it was introduced and used, as well as how many ratings it has. You also get a direct link to the code, hence the name Papers with Code, so GitHubs are linked as well. Of course, everything is community-driven, so not everything is always entered. Then blog posts, documentation links like in Keras. If you look at it all together, you can already see that there are advantages, disadvantages, and differences in these representation forms. Can you maybe say a bit about that?

I8  
 Yes, first of all, when you look at the overview here and then compare the label on the right and look at all the materials on the left, I think first you need to familiarize yourself with the terminology on the left before you can even extract the content you want from a graph. So it’s always nice to visualize something from a table, and then it’s an easy way to look into it. But it all looks quite deep and complex. You really have to read an article, and you don’t have this one quick glance to get a rough overview of what is there. That’s what strikes me the most, the label on the right and all the content I compiled on the left. You can see there’s a lot of complexity, and you would need a significant understanding for that. For example, thinking of my role, the label alone might be better to quickly say, yes, try these three models from the label, they seem good, you could use them. The documentation on the left isn’t that interesting for that, it goes too deep. You would have to invest several hours just to get into it.

A1  
 Yes, especially, I want to make this clear again, we are of course not saying everything is useless and we just need to label everything and then the problems are solved. Engineers and developers, of course, want information in depth, and ideally there should be a detailed report behind a label explaining how it was created, how it can be reproduced, and all that. That would be ideal. But yes, that’s what we have also noticed whenever we work with companies. For them, when you show machine learning results, it’s pure magic. And we hope that labels can simplify the interface a bit, to exchange information about models without having to say, yes, you have to read this paper and look at this table to see why this is better for this use case. Or compared to other models, it behaved like this or that, you can see that in the table, but it’s not so quick. You really have to dig into the details. And we hope that we can simplify communication, reporting, and the exchange of machine learning results in this niche. Exactly. Coming to the end of the conversation, let’s go back to more general questions. Back to the topic of trustworthiness. Even now, especially in this comparison of models, do you have a feeling, how would you trust these information sources? Which more, which less? Or how could your trust, or the trust of employees, possibly be increased?

I8  
 Yes, I would say that is always very dataset-specific. So, if we really look at data in the predictive maintenance context or something like that, I don’t really have that much comparability and I also don’t have these large models. That’s why I have to take care of explainability myself. Then again, the question is, how is it in the image domain? That’s hard for me to estimate, otherwise I would always say that we actually have to describe and evaluate explainability based on the use case, if you look at it in this machine learning context. The question is with ImageNet, with the big cases it goes back to someone else having to represent explainability. In a random forest you have to take care of it yourself, or if we have controlling data or something, then it’s like the example, it would again be the case that we would do it ourselves, but eventually larger models come along that already take over more and then you can’t really represent explainability within the company itself. So it’s a bit like that, you have to explain a lot yourself, but some explanation also has to come from the documents, papers, etc.

A1  
 Yes, actually the topic of explainability, which I presented a little last year at the conference where I met XXXXXXXXX, is a very complex concept. Many people say that random forests or decision trees, in particular, are inherently interpretable and explainable because you can of course just follow the tree at every decision and see why that decision was made in the end. But that is only conditionally correct. If you have a decision tree of depth 500, I wouldn’t say it is inherently explainable. Sure, if it has a depth of 10, you can still follow why the decision was made. But if it is randomly split across 500 different features, then there is no explainability left. Of course, you can trace each decision back concretely, why each feature was split at which point. But it is still a very difficult question. In the vision domain as well as in NLP, a lot of research is being done on post-hoc explainability. That means you want to generate an explanation for a specific model decision and trace it back, for example, which features caused this decision. And that is of course also a very difficult area. To build this bridge, because the model is so complex and there are thousands or even millions of operations, you can… You can also compute it… Of course it is deterministic and traceable why which layer outputs what value and which node or neuron outputs what value. But it is not humanly interpretable. So explainability is a big issue. I still find it very difficult because… Explainability… You basically have to say that beyond a model size X, the model is no longer inherently explainable. You can somehow try to estimate it based on the number of parameters. But whether the model has one or ten million parameters doesn’t really matter. At that point it is no longer inherently explainable and you need post-hoc explainability methods. A big problem there is that there is no ground truth. You just get… There is no ground truth for explanations. I explain things differently than my colleagues at work. That means even explainability is not… There is no ground truth, no label you want to achieve. This is the right and this is the wrong explanation. Since you don’t have this data for images… Why do I recognize a car in the image, one person says it’s the tires, another says it’s the general shape, another says it’s on a street, so it can only be a car. Since you don’t have this ground truth, it is also very difficult to develop models or methods to produce such explanations. So there are many question marks in research regarding explainability. Yes. How about the authority that ultimately approves or provides it? Do you think that might also play a role in trustworthiness?

I8  
 What do you mean by that?

A1  
 Well, it makes a difference whether someone writes a blog post about MobileNet and says, yes, I used this and it’s like this and that, or whether it’s the developers of MobileNet providing the information. Or whether it is some third-party instance profiling it. In other areas in the industry we have exactly these differences. Do you trust the company that actually wants to sell a product? Or do you trust an independent instance? Is an independent instance really independent?

I8  
 From the work perspective, I would say the source is not that important. It’s somehow more like what works well and what is well explained counts more than who published it. That’s why I also said Medium, because it often explains a practical example well and you can work with it. Papers are of course also interesting to look at. But adapting them to the use case I have myself usually takes a bit more work. Documentation from the manufacturers themselves is always the next step. I take the example and approach from Medium and then look into the documentation. I can’t avoid that. I also don’t know many companies or publishers who present their models in their own documentation compared to other models. At least, I haven’t noticed that.

A1  
 If anything, that would be in the paper.

I8  
 Exactly, that would be kind of the entry point, although it’s not really an entry into one’s own development, at least it doesn’t feel like that

A1  
 Yes, the question comes from the fact that XXXXXXXXX did a study on this last year, on Trust Certificates, so just a large-scale quantitative study. We surveyed 20,000, 30,000, 40,000 people: if an AI product has a Trust Certificate, yes or no, does that make the product more trustworthy or not. And what came out of it was that, in case of doubt, people tend to distrust AI in hypothetical products, whereas in practice it turns out that if the usability and the benefits are there, then people suddenly don’t care at all where the model comes from, how it is certified, or whether it was certified at all, where the data comes from. If it’s just a cool product that somehow makes everyday life fun or takes work off your hands, that is more important. At the same time, of course, there was also feedback that it’s like a certificate and people trust, for example, the Energy Label somewhere and don’t question it, but the question is whether they would, if the manufacturer writes on it that the model is very efficient, or if it’s a third-party instance. And we are a bit wondering where it should best be placed, who should issue such labels? Manufacturers won’t do it because they have nothing to gain, especially since it exposes the weak spots of the models. So ideally there should be a third-party instance. And should this be with TÜV or DIN, or is it a government agency, or an NGO? There’s no clear answer yet, but we are somewhat interested in feedback on this question.

I8  
 I would compare it a bit with the refrigerator label. You kind of have a general trust there. They somehow limited themselves with the Plus Plus and didn’t build it dynamically enough. But initially I assume, yes, this is from a central authority, so that’s good. If you say it like this, an open label, then it’s kind of like with organic products, where every label carries a bit of skepticism as to whether it’s really good. So a central authority, or at least the same format, and somehow already publicly communicated, not even necessarily by the certification body, I think, already gives it a relatively official character, and through this presentation it also looks officially recognized, so people trust it, I think. So I could imagine that the label could already convey trust.

A1  
 Okay, cool. Very good. I think we are actually done with the questions then. XXXXX, do you have any comments, anything unclear, anything where, from your professional background, you think we need to take another look?

A4  
 No, I think we’ve actually covered everything. I can’t think of anything else.

A1  
 Then, XXXX, as long as you don’t say, no, I just had a thought that absolutely has to be included, we could already end the recording.

I8  
 Not right now, maybe something will come up, but I can’t think of anything at the moment.

A1  
 Okay, then we will definitely end the recording.