A1

Great, exactly. And the audio levels are working on my end too, everything looks very good. So then we can start the interview. Right, so your background, you just mentioned you're studying, and you're also working, if I understood that correctly? Or are you doing an internship?

I4

Currently a mandatory internship.

A1

Okay, in which area?

I4

IT security. So it's a monitoring company and at the moment I'm doing threat intelligence.

A1

Okay, all right. Could you describe your experience with AI so far, or in what way you might have already worked with it? If you haven't worked with it, that's completely fine too, but just describe a bit how you’ve come into contact with the topic of AI.

I4

Right, I actually haven’t worked with it so far, and my contact with it is mostly when something interesting comes up that reaches general media. Then of course I see that. So things like that, obviously the well-known ones like ChatGPT or the various models from bigger companies, and also videos or similar content. If something interesting shows up in my feed, I’ll check it out and get a bit of a sense of how it works.

A1

Okay. So that means you haven't really experimented with a library yet. So you haven't actually deployed anything in practice, but rather worked more on the user interface side.

I4

Yes.

A1

I mean, things like ChatGPT and that kind of stuff can be used and tried out even without knowing how to program. Have you done that already, or not even that?

I4

Yes, did that.

A1

Okay. We will probably rephrase a few questions a bit now and then ask you to put yourself in the position of someone else, because otherwise the questions wouldn’t make sense. You are currently studying IT, which means you are automatically learning software development and similar things as part of your studies, and that might be something that could work for you professionally in the future. Now, assuming you were in a position at some point where you are helping to shape AI in some form, using it within a company or co-developing it, can you, based on the experience you already have with AI, somehow imagine where problems or challenges might arise, how such processes might work, and what would be the biggest inhibition level for you in that scenario?

I4

For me probably, as someone working in IT security, depending on what it's used for, the IT security aspects. I’ve actually already seen some interesting cases how AI is used or where, I don’t know, ChatGPT gives away secrets, in quotation marks of course, or things like that, if you just manage to convince it. And that really depends on how it’s applied in a business context. If it's used, I don’t know, as an internal company assistant or something like that, then of course people who aren’t sure what they can or can’t share with it could run into problems. Yes, I think that would probably be the biggest issue.

A1

How is it for you, how accessible have you found information on the topic so far? For example, when you hear about it in the news, do you feel that, with an IT background, it’s quite easy to get into and you can quickly find concrete information on how to apply or use it yourself? Or is it more like a completely obscure black box for you, where you’d say it all just feels like magic?

I4

I do think that it's actually relatively easy to get into all of this, at least that's my impression, because there are fairly high-level interfaces that allow you to work with this kind of thing. And how deep you want to go into it is, I think, pretty individual.

So I believe it's fairly easy to get started. But I also think that a lot of people don’t really understand much about it. For example, how machine learning and that kind of stuff actually works, like how ChatGPT even works. I'm absolutely sure that, I don’t know, 80 percent of people have no idea at all.

A1

I think you can definitely scale that up further. The number of machine learning experts is still quite limited. Sure, it's increasing, but...

I4

Yes, I didn’t even mean it in terms of experts, but just the fact that this thing forms sentences by using tokens and then calculating probabilities. And at that point, I’m already not quite sure myself whether it always picks the most probable one or if it actually selects tokens based on the probabilities, but yes.

A1

I can definitely tell you that the models currently used in the large language and NLP field are actually not based on probabilities. Calculating probabilities is extremely time-consuming. You would first need to construct probability distributions over your data somehow. That is extremely demanding. And the major success of neural networks, which are also used for large language models, is actually due to the fact that they have found a very efficient way to empirically recognize patterns in the data that are not based on probabilities and base distributions or something like that.

 That means you do get something like a probability, you get more or less likely outputs, but these are not mathematically correct probabilities. They do not fulfill the conditions that you would actually mathematically place on probability expectations.

I4

Are these more like confidence values from the AI?

A1

Basically, those are just Softmax outputs. But we're getting a bit too deep into the technical details now. That's not really what this interview is about. But yes, exactly. These internal workings, you notice that again and again at professional conferences, even in conversations with XXXXXXXXX and so on. It's just not easy to really understand all of that. It takes a few years of study. And especially for people without an IT background, even the term algorithm is already kind of a magic box. And then AI is way more complex than just an algorithm. But yes, exactly. You're right in what you said earlier that you can get into it. You can start doing something with it pretty quickly. You can run a few lines of code and use it. But really understanding it thoroughly is not that easy.

I4

Yes, and figuring out what the right model is for my area of application and things like that is of course always a question. And providing proper datasets is also, well, yes.

A1

Yes, exactly. Can you think of anything off the top of your head where you feel like machine learning people could actually do better? Like, if they did this or that, it would make access to and understanding of the technology easier?

I4

I can't think of anything specific right now

A1

That's totally fine. As I said, this is just an open conversation. No exam question. So, as I already hinted earlier, in my research I’ve looked into how to make the development and characteristics of machine learning models and AI models more understandable, especially for people who don’t yet have expertise in the field. Because when you talk to a scientist about it, they’ll say right away: give me the paper or show me the formulas and give me the code, then I’ll understand it. But for people without that background, it’s way too complex. That’s why we, or rather I, have been working on a system that takes a slightly different approach, and for that I’ll go ahead and share my screen now, once PowerPoint decides to cooperate and start the presentation. Alright, perfect. Here we go. You can see my screen, I hope. Perfect. So, this would be the area of AI labeling as one way to make the transfer of this technology easier to understand. Without us saying too much at this point, would you like to give us a spontaneous reaction to the label? What do you see, what are your thoughts about it?

I4

That reminds me of the Nutri-Score. Otherwise, I can already get a rough idea of most of the terms. Power draw is probably understandable, like how much power it uses. Running Time, Inference, I’m not quite sure what that is. It has something to do with the input, maybe a set of inputs or something like that, no idea. I can’t fully picture what Accuracy and Robustness mean. I have a rough idea of what Accuracy could mean, but I imagine it’s hard to quantify. As for Robustness, I’m not sure.

A1

Maybe take another look at the upper section of the label, the part above this score-like area, above this scale.

I4

There are QR codes for information.

A1

Are you familiar with ImageNet?

I4

No.

A1

Okay, then I actually need to briefly explain this at this point, because the MobileNetV3 we labeled here is actually a model for image recognition, more precisely for image classification. ImageNet is one of the largest image datasets. It was introduced in the 2012 challenge and has been mined extensively over the last ten years. It contains 1000 classes, and there are about 1000 images per class. The images have a high resolution, at least back then, around 228x228 pixels, which was significantly larger than other image datasets used in practice back then. This also results in a very large amount of data. ImageNet consists of about 160 GB of image data. As mentioned, with 1000 classes, if you were to randomly guess a class for each image, like saying this is a cat or a dog or whatever, you would end up with an average accuracy of 0.1 percent, because you’re always predicting one out of 1000 classes and the chance of picking the correct one is accordingly low. So that’s kind of the baseline. MobileNetV3 reaches about 63 percent accuracy when properly trained. And inference, which you just asked about, refers to a model's output. So, you send a request to the model and get a classification decision for a single image in return. In practice, though, you wouldn’t send just one image but rather a batch of images, because you can do this quite well in parallel on GPUs. You can get predictions for 32 or 64 images at once, depending on the batch size. So, that also counts as one inference, since it doesn’t matter whether you classify a single image or an entire batch, it’s processed in parallel anyway. Exactly, that’s a bit of the background information. So we're actually not in the NLP field here, but rather in an area of AI that, for many years, was all about image recognition, installing cameras everywhere, recognizing and tracking people, analyzing manufacturing processes in industry, detecting errors, and so on and so on. And now, in the last two years, the general understanding of AI has shifted towards language models, and that has become the most common point of contact with the topic of AI. But of course, smartphone cameras and facial recognition are still very relevant topics.Do you have any immediate reaction to that? Do you have the feeling that something like this is useful or do you think to yourself, I don't know who this is useful for, I don't know? Do you have any thoughts on that?

I4

I can imagine that it's useful as a rough categorization to think about what exactly I want to focus on for the AI I want to develop. So at least I can already weigh things within the four categories below and ask myself what matters most to me, how the priorities should relate to each other. Do I need something that is definitely accurate, but I use it so rarely that energy and time don't really matter to me, things like that.

A1

Yes, exactly, so decision-making processes are also kind of what we’ve identified as our target audience, it's also mentioned on the website. So people who make decisions about which models are most suitable for their specific use case, but who may not have the domain expertise or the time to read ten papers first. So with these labels, you can simplify and make very abstract or extensive information more accessible, and that also enables comparison. Here, for example, is a second model. Could you maybe also briefly summarize this one and share what your immediate reaction to these two labels is, and how you perceive the comparison?

I4

Yes, first of all the overall score, of course. One is graded with an A, the other with a C. And then you can compare them further. One is naturally much more accurate than the other but performs worse in the other categories. That explains the C, but of course you can then also decide for yourself, okay, if accuracy, what is accuracy in German, “Akkuratheit”...

A1

Yes, the “Genauigkeit”.

I4

“Genauigkeit”, yes, accuracy, if that's more important to you, then of course you can weight it that way. Right. So that can already be done relatively easily, and I think for most users, those are already the most important categories. Like, how exactly it works internally is probably less important to most people than how long it takes to give me an output when I put in an input.

A1

Yes, exactly. Especially with regard to your academic background. Do you feel that something meaningful is still missing here, or which aspects do you personally think are missing on the label? Something where you would say, this would really interest me. For example, you might say that you use ChatGPT in your studies, in your free time, at work, or wherever, and there is something you would really like to know but there is no information about it. Does anything come to mind in that direction?

I4

What is robustness for?

A1

Exactly, that's right. We didn't talk about robustness earlier. You have to imagine that robustness always examines whether, when you perturb or modify your input data in some way, the prediction is still accurate or correct. That's practically the difference. So the accuracy that's listed below is simply given by a test set that was not seen during training, so you basically have the classification quality of your model for unseen data. The hope is that this also provides insights into whether data seen in practice, which might differ from this, can still be correctly recognized. And to determine corrupted robustness, you would then alienate the test dataset further, adding noise, shifting the images, distorting them, mixing them together, and so on, to estimate how robust this accuracy really is. And here, for example, with EfficientNet, we see that robustness drops drastically, whereas with MobileNet, the robustness is still relatively high. It's fairly close to the original accuracy.

I4

Oh, so the values belong together?

A1

A little bit, yes.

I4

So the closer one value is to the other, the more it makes sense.

A1

So you can see that EfficientNet is still better in terms of robustness. That means it still correctly recognizes more of the altered images than MobileNet. But in relative comparison to the original accuracy, the difference is clearly worse. This means EfficientNet seems to be a bit less robust, in the sense that it fits the data a bit too closely. It is slightly overfitting in that regard. While it recognizes the validation data, or rather the test data, extremely well, it is actually more susceptible to deviations and changes in those images.

I4

That might actually be an interesting factor, maybe even more interesting than accuracy, depending on the use case, probably. How one differs from the other, because accuracy doesn’t necessarily say much if, well, you don’t actually have… I mean, in practical use you almost never have, or well, I could imagine some use cases, but not really the actual images from the test cases, rather arbitrary images.

A1

Yes, so that's actually also true. Of course, for every model, before using it in a product, you would have to check whether the images that come up in our product have anything to do with it, even remotely. Like, if you think about it, we have some kind of assembly line where objects are being produced, and at some point, a camera takes a picture, and you want to determine based on that whether it was correctly manufactured or not. In that case, the entire ImageNet dataset probably won’t help you at all, because it's not about dogs and cats and airplanes, but about completely different things. ImageNet is more intended as a collection of images of all kinds of things, which becomes interesting, for example, on phones. What you currently see in Google Pixel or Apple phones is that you can take a picture of something with the camera and get information about it, or have it recognized what kind of object it is, that kind of thing. And of course, you're absolutely right, both are interesting qualities. It's really hard to say which one is more important than the other. They kind of depend on each other, and in the end, it's super application-specific how important one or the other is.

I4

Yes. But what you mentioned earlier about how they not only depend on each other, but that the difference between the two is also relevant.

For example, with MobileNet, the robustness is lower but still closer to accuracy, would already be a step further. A step beyond just looking at it and gathering information. But I think it is already a relevant value to determine how robust something is. I'm wondering whether robustness in general says more than just the deviation from accuracy.

A1

Yes, I understand what you mean. That’s definitely something that could be brought up critically. The issue here again is: if you have a different use case than image classification, you end up with completely different metrics. In the language model domain, for example, you wouldn’t even talk about top-one accuracy, because there’s no classification happening. It’s all unsupervised and generative, not about classifying individual values. But to come back to the question I asked earlier: you can’t really think of any other aspects that should be included on such a label? Or is there maybe something that comes to mind where you’d say: actually, I’d be interested in that, and it’s currently completely left out here?

I4

No, not really.

A1

Okay, I would have thought, with IT security in mind and all, that topics like… To what extent the model takes differential privacy into account, or how the data was curated, might also be interesting aspects.

I4

Yes, that’s true. I just haven’t looked into it deeply enough yet to be able to say how it can be quantified or whether it’s even quantifiable.

A1

Yeah, but take ChatGPT for example, a language model like that. You don’t get any of this information when using ChatGPT. Not easily, unless you really go into extensive research, and for some of these things, you actually won’t find any information online at all. Are there any other properties where you’d say: I’d like to know more about that, but the information is just somehow missing?

I4

Maybe for many things, I do wonder how exactly it learns... I mean, that’s probably one of the first things you find out about most systems. But are my user inputs part of the training data for the AI? Like, do I need to worry that if I type personal stuff into ChatGPT, it might come out for someone else with the right question or influence the system in some way?

A1

It’s somewhat similar to what happened with some of the image generation models, where it was shown that with the right prompts, you can almost perfectly reconstruct individual images from the training data. So, for example, if you phrase the prompt correctly with "generate a person with this and that hair length, hair color, these and those features, in this and that setting," then it was actually possible to reverse, without the model having an explicit link, a very, very similar image from the training data where you can clearly recognize which person is shown in the picture. This is kind of going in that direction as well.

 So the question is: can you phrase a prompt like that, or how vulnerable is ChatGPT to someone formulating a prompt and then ending up with a one-to-one quote from some master’s thesis, just because it’s available online? And of course, ChatGPT doesn’t reference it properly but instead…

I4

It doesn't have it as a direct reference, because it doesn't actively look it up.

A1

Right, it doesn’t look things up, but if it gives a plausible answer that is essentially based one-on-one on a data point from the training set, then that is obviously a problem, especially when it comes to privacy.

I4

Yes, that's true. That would be another interesting factor, but I don't know how to quantify it.

A1

Yeah, I actually don't know either. I'm not enough of an LLM expert for that, but I'm sure that metrics for this are being developed, also regarding things like hallucination and so on, because of course this has to be quantified and assessed somehow to understand how problematic it is with these models. Um, right, so that means, apart from that, when it comes to the design of the label, there wouldn't be anything further from your side for now...

I4

Otherwise maybe something else that just occurred to me, an area of application or something. I don't know if something like that would be possible.

A1

Degree of application of the model?

I4

Yes, having it indicated that it is for image recognition, image generation, things like that, I think would make sense.

A1

Yes, that is of course always completely connected to the fact that, well, actually this information is already written there in text form: inference on ImageNet. Then it is clear what is meant, but of course one could also try to represent it more abstractly, for example in the form of a badge or something like that, absolutely. I have actually thought about that before.

I4

The step that would then be taken off the user's hands.

A1

Yeah, and otherwise it's not obvious at first glance either, right?

I4

Yes, which would then have to be checked

A1

Yes, that is actually a very good point, a very valid point, yes. Cool. When you say that you have already acquired some knowledge about AI and are somewhat interested in it, where do you get your communication and information on the topic from? You already mentioned earlier, of course, media and news, when something comes up there, but when you say: Yes, that is quite interesting, now I would like to dig a little deeper, what are the resources you turn to?

I4

Mainly YouTube, so videos on different topics. Then also sometimes where someone implements something with AI. So, mainly that. And otherwise just the feeds.

A1

Mhm, because we actually made a small list there as well. YouTube is actually a good point that we hadn't explicitly considered so far, and I also haven't heard it mentioned in the interviews yet, but it kind of falls under point 4, blog posts, because YouTube is essentially a video blog platform. Yes, on the one hand, we clearly have scientific publications, and fortunately a lot is available on Arxiv, but of course, that comes with the challenge that you first need to really dive into it. It’s not something you just quickly read through. Then there's Huggingface, for example, I don't know if you know the platform. It's a platform for language models, especially popular due to its use with Vision Transformers and NLP models, where you can find things like model cards. Yes. But the original idea actually goes back to Google. Papers with Code… I don't know if you’re familiar with that. It is kind of a mix of the two. You can voluntarily submit results there if you’ve developed a new method or dataset, and then you can present the key takeaways from the paper in things like leaderboards.

Then, of course, there are blog posts, and all the libraries have some kind of documentation. IBM, for example, also offers factsheets for IBM products. To visualize this a bit, we have on the left the publication for MobileNetV3, just the first page of about fifty. On the right, there is a model card, where you can see some measurements on two different devices. Actually no, they are the same devices but different models. One is a MobileNet Small as tier Flight, and the other... as Malis. Right. On Papers with Code, you of course have the paper, but you also get a list of all kinds of papers that have used the model. Sometimes, you also get small graphs showing how it performed. On the right, there's a blog post about MobileNetV3, documentation from Keras, and another IBM factsheet. IBM only releases these factsheets for their own products and models, so there is nothing about MobileNetV3 from IBM because it’s simply not from IBM. Now, if you look at all of this in comparison, where do you see advantages and disadvantages in these visualization and reporting formats?

I4

Yes, mainly in how in-depth they are. I think you can work your way through them from top to bottom, depending on how deeply you want to engage with the topic. So I would rather say that the different materials are for different users. The paper is more for people who want to understand how it works, why it works, what it does, and how the values were obtained. The... what were they called from IBM, the...

A1

Fact Sheets

I4

If I have seen this correctly, this is more high-level information

A1

Yes, exactly, it's what the name says. It's a factsheet. The model cards are also a bit similar.

I4

Right, where you then see less about how the values come about, and instead you are just given the values so you can roughly categorize, classify, and compare them. So I think the advantages and disadvantages depend on how precise or imprecise they are. I would say those are advantages and disadvantages depending on whether you want something to be precise or imprecise.

A1

Yes, one important point is that the labels we showed here are actually generated in a similar form by a software library that I have developed over the past few years. You can simply load a Pandas Data Frame into it, and for each row, you get the model that was used on the dataset along with its properties. The entire scoring process and the generation of the PDF are fully automated based on that. In contrast, I still have the feeling that writing papers, documentation, blog posts, and so on is a much more manual effort. Although this might become easier now with language models, one could also question whether it is actually a good thing that it becomes easier, because if you use ChatGPT to write blog posts, then, well…

I4

Is that then also correct?

A1

You still have the manual effort to proof-check it again. If you want to do it properly. And in my case, I also did a lot in the area of benchmarking, to automatically test models. Right. Yes, I think it's also very important to note that, of course, we do not believe that labels will replace all the other things. It's simply an additional form of representation, which enables comparison and insights at a glance, something that just hasn't been possible so far.

I4

Kind of like the factsheet, just a bit more high-level. Yes. And the score can even take over the comparison between the categories for you.

A1

Right, so it’s also kind of like this: in theory, a company could just decide which values to set for its products. Of course, accuracy is always important, that’s clear. But there are actually companies that say whether it’s 80 or 85 percent doesn’t really matter. However, if the model only uses half the energy, then many users would be very interested in that, because in the end it’s all energy costs that the product incurs in some way. Making these trade-offs more transparent would, I believe, be in our best interest. Otherwise, well, everyone does it. They all want the most accurate model, but cost-efficiency tends to fall behind.

I4

If I want suggestions when writing messages or something like that, I honestly don't care whether it's 80 or 85 percent accurate in suggesting the word, because in the end the user checks it anyway. But if it doesn't have to load for two seconds, only one second instead, or uses significantly less energy on my phone or something, then of course that's more reasonable.

A1

Yes, that is actually super relevant right now, especially in the mobile sector. There’s a real push to make models smaller in that area. That’s actually the whole idea behind MobileNet, by the way. It’s a model that was explicitly designed for smartphones, or at least with a strong focus on deployment on smartphones. But yes, in academia, there is still a heavy focus on accuracy, and people often claim they’ve developed the best new model, which they might have, at least on paper, when you look at the quality metrics. But then their model uses five times more runtime, electricity, and resources. And that’s definitely something we should question. So web-models. That’s actually a big topic in our group at XXXXXXXXXXXXX, where we focus on resource-aware machine learning. That doesn’t mean we’re not allowed to use resources. We can absolutely work on big NLP models and the like, but the key point is transparency, to openly reflect on those trade-offs. That’s something I really advocate for. Right, let’s come back to the topic of trustworthiness. Trust is also a big topic at the XXXXXX XXXXXXXX, and also at the XXXXXXXX XXXXXX XXX XXXXXXXXXXX XX here in XXXXXXXX, and even at XXX. It’s a major focus. Do you have any thoughts on that, based on everything we’ve talked about so far? Trustworthiness?

I4

Yes, so earlier we already talked about what happens to my inputs, whether they are further processed or not and whether they can be output somewhere. That is of course important for trust, or rather, confidentiality is more about secrecy (bad connection) that my inputs are confidential or not- Or at least that there should be awareness about whether inputs are further processed or not, and what exactly happens to them. I do think that’s relevant and should be important to end users of a product. Because, for example, during my internship we were told not to use customer data with ChatGPT, since you are essentially passing it on to a third-party company and you don’t really know what happens to it. But I think this here is a bit of a different case.

A1

Yes, so information about trustworthiness would definitely be something that could somehow be represented on the labels. Although that often doesn't depend on the model itself, right? The model runs on some computer, and by itself, it can't really use the data in a trustworthy or untrustworthy way. In that case, it would be more about the product or service associated with the model, like OpenAI, which is basically a web interface. That’s where this information needs to be made transparent. The model itself doesn’t make the data trustworthy; it doesn’t handle it in a more or less trustworthy manner. For the model, it’s just a bunch of bytes that gets processed and that’s it. What about thinking in terms of who could issue such a label? Because, for context: XXXXXXXXX conducted a study last year on the extent to which trust seals for AI would be helpful. And the overall result of the study was, to summarize roughly, that people… it was a quantitative survey with 40,000 participants, not qualitative, and they were paid to complete the questionnaire online. And the general conclusion was something like: I actually do not trust AI. They created two use cases, one was, I think, early cancer detection, and I can’t recall the other one. They presented the product to one group in a neutral way and asked how trustworthy it seemed. The other group got the same product, but with a trust seal attached. And the average feedback was actually that people did not really care whether there was a trust seal or not. In case of doubt, they would rather not trust it. It was a rather critical stance towards technology, which is understandable, because these were not experts who have to make such decisions. Interestingly, everyone still gets on airplanes. People do trust that technology, because we have many years of experience with it working well. But you can also see it in areas like electric vehicles or autonomous driving. Many people simply do not trust the technology there. We are also wondering how that plays out for them.

I4

I feel like it’s because of the name. That many people imagine more than what AI actually is. So, the average person probably has a more Hollywood-like understanding of AI, thinking it’s something that really thinks and has opinions on certain topics, rather than seeing it as a system with a set of inputs and outputs, where procedures determine what happens. Also, the question of who could issue such a label is always tricky. In the IT security context, there is a similar counterpart with certificates for websites and so on. In that area, end users put an enormous amount of trust into anything related to IT security, which is understandable to some extent, and also not, because not everyone can really dive deep into how encryption works, how the whole certificate authority system works, what a root certificate is, and who we are actually implicitly trusting. So, in the end, it would have to be a company or some kind of authority that is independent from the businesses and where you would also have to ensure that the individuals assessing these things are not being bribed. And I mean, many of these things are relatively well verifiable through peer review the time spent, for instance, can be checked quite easily. Okay, with the specified characteristics, it should really have that runtime or not, and I think accuracy too. Power draw should also be fairly easy to verify… basically, all of the listed values or factors should be quite traceable or verifiable in terms of whether the stated information is accurate. And I do think that something could develop organically over time that people would generally trust. That’s how it works with most things, after all: if enough time passes and it turns out that, okay, for these 80 models the information was accurate, then why wouldn’t it be accurate for the 81st model?And if there’s some kind of authority where many experts say, “Yeah, okay, that’s valid,” then it’s likely to be accepted. Once something like that has been established, I do think trust or rather, confidence in it would make a real difference. I can definitely imagine that having such an authority evaluating and issuing these things would be reassuring. I think that makes a lot of sense.

A1

Okay, cool. So we asked ourselves that question as well. That was also something that came around in the XXXXXXXXXX study. So, people have somehow often said, “I don't know where it comes from, so I don't trust it”. And we generally have such a problem in society with a lack of trust in technology, politics, authorities, where we, even as machine learners, simply can't do anything. So, people have vaccine skepticism, for example, there you have a completely structured system of how to somehow develop vaccine products and bring them to market. Sure, that also has its flaws, that's not perfect either, but it's a pretty well-structured, long-established and also organized one. And also a transparent system somewhere. So, if you want, you can understand relatively well, so you can really find a lot of information about how such a study is carried out and so on.

I4

You can also reproduce it once you've studied enough.

A1

Yes, exactly. The process is actually made quite transparent in that case. Still, people say, "I don't trust it." Simply because they read online that it's not trustworthy. So these kinds of problems are not really ones we can solve. But of course, we're also thinking about how a label could potentially become something where people really say, "Okay, this has substance, I’ll trust it." And as I said, we have things like cars and airplanes, where almost everyone has a basic level of trust, apart from very few exceptions. And in recent years, technology, especially in the IT sector, has developed so rapidly. You can really see that people are willing to throw trust completely overboard, as long as they get cool benefits. Everyone is immediately on board with putting an Alexa in their home, because it's kind of fun. And somehow...

I4

Internet of Things. Very difficult in terms of IT security

A1

So, it's entertaining and practical. But I don't need to tell you that, from a security perspective, it's all pretty wild and actually quite absurd.

I4

That every device, your fridge, your heating system, and even, I don't know, your cat's food bowl, all have computers and are connected to the internet, is problematic.

A1

Yes, exactly. And at the same time, it was said that usability and entertainment outweigh everything, right? People still buy it and are excited about it. And AI is kind of funny right now, because all the companies are saying, "We have to do AI," while many people in the general public are saying, "I don't want anything to do with AI, it's all too creepy for me." So right now, we’re seeing this real break in technology, where I actually don’t have the feeling that the new AI in the Pixel 7 is a selling point. I don’t think people want to buy the device just because it says AI on it.

I4

I also don’t think that... Oh, my camera froze... I also don’t think that it’s a selling point, or a reason not to buy, for most people. Oh, my phone died. Oh, no it didn’t. Why isn’t my camera doing anything?

A1

Don’t worry, everything is fine...

I4

Yes, so I think, for example with the AI in Google Pixel, it is not really important that there is AI involved. And I believe most people do not really care either, because it is one level of abstraction further. You are not talking with the AI, and the AI is not taking pictures itself. You take a picture, and it goes through an AI system. And then it is already so far in the background that most people probably do not care. What matters is the final result.

A1

That was really funny, for example, with Samsung smartphones it was a topic a few months ago that their automatic photo correction started using generative AI for things like sunsets to make them look cooler. And that means, in the end, it's no longer the photo you actually took, but rather a synthesis of your picture and all kinds of images from the internet.

I4

And every photo

A1

Yes, exactly, and then it just looks cooler somehow, but as I said, usability really outweighs critical thinking at that point.  
 Right, yes, but cool. Then we are actually done with our interview now.

I4

Maybe regarding the label: I could imagine that trusting the label is easier than, for example, in the vaccination case we mentioned earlier, because it's easier to verify. I can definitely imagine that it's going in that direction. The barrier of entry is much lower, I think. You can check the runtime relatively easily. You just need the appropriate system and can then check whether it's roughly within that range. In contrast to a representative study on the effects of, I don't know, whatever vaccine, here you can simply run it and check whether it matches. And I suppose that makes it easier to build trust in such a label.

A1

At least for open-source models. One also has to keep in mind that, for example, the GPT variants from OpenAI are not really that open. That means you cannot simply run them on your own hardware to measure them. You basically have to hope that the provided information is plausible. And in fact, in their GPT-4 technical documentation, they explicitly stated that due to competitiveness in the NLP space, they are not disclosing any information about training data, training effort, training architecture, network architecture, hyperparameters for training, and so on. This means there is no reproducibility at all. Not even at the core. Aside from the fact that reproducing the model would require something like 8000 GPUs, which obviously no one has at home. But of course, with smaller models, reproducibility is more feasible. Yes, definitely. Alright. As mentioned, that brings us to the end of the interview. I’ll stop the recording at this point.