A3  
 Let's go. Okay, yes, then the first question, yes, please introduce yourself briefly, what is your background?

I12  
 I have a psychology and human factors background. So I studied psychology in my bachelor's and human factors in my master's and did a PhD in the area of human-machine interaction. There, I dealt with the topic of how people actually understand algorithms. And now, outside the academic world, I work on UX research, meaning the question of how to design software so that it is meaningful and has so-called customer value, that is, meets customer needs. Before that, I did this kind of work in the medical technology field. I was in strategic product management and UX research. And now I work in the field of communication software, telephony software, and I am currently trying to understand what needs, what so-called jobs to be done, our customers have in order to work with this software.

A3  
 Do you want to disclose your company, where you are currently working?

I12  
 Oh, will that also be in the… in the… no? Okay.

A2  
 I think it’s just about us knowing whether it’s a medium-sized company, how big it is, what you do.

I12  
 The company is called XXXXXXX, is in XXXXXXXXXX, and is an owner-managed medium-sized company with almost 300 employees. It is a software-telephony company.

A3  
 Okay, got it. Then you can go a bit more into your role and especially your tasks regarding working with AI and ML systems.

I12  
 So I work in product development, in a product team. Until last week, now I am changing my product team and the focus of my tasks is shifting a little, but it is still in progress. But I am located in product development and work closely with developers, UX designers, and so-called product owners on a specific part of the software. In the past, it was mostly about the experience of the so-called IT admins who use the software. Now it is changing a bit and I am thinking more, but still with the same roles, about which AI features, in terms of which AI feature statistics or insights, our customers need in order to work more productively and efficiently. I think this whole AI topic in our company is closely linked to the concept of productivity and efficiency improvement. And my role is to find out, with the help of which features. And in my work, it is especially about understanding what people do, what people need, from more of a user needs perspective and not really from a solution perspective, to look at which AI features and so-called AI insights, let’s call them, are meaningful for customers. And that really depends on the users.

A3  
 Okay, very interesting. What is your daily work routine? Does a diverse job like yours even have a routine, a flow?

I12  
 Yes. So in software development, we work according to Scrum, which is quite common, and there are clear so-called Scrum events, which in our case are actually meetings, but I think it can also be done differently, so that there are fixed structures. We work in so-called sprints, which are two weeks long for us, and within these two weeks we try to ship a certain goal, a certain customer value, so to speak, that is, to create it and bring it into the world. And that is the usual working mode. Since I am in UX research, it is a little different. I work very closely with the so-called POs, who have product responsibility, meaning they make the decision in which direction my product or this part of my product should develop. And I work very closely with them in so-called sparring, that is, I work on equal footing. I do not just support them, but think together with them about what the next meaningful step is. And this is a hybrid model, between working in Scrum with the developers and UX designers at the operational, very front level of how it is built, and also what is called delivery, and the other technical term is discovery, meaning thinking ahead about possible potentials, which is the close collaboration with these product owners. And yes, question?

A2  
 Yes, exactly, I am just raising my hand. Yes, can you describe where your touchpoints are, where AI comes into play for you, where it affects your daily work?

I12  
 So not how I use it myself, but…

A2  
 Yes, well, yes indeed. Exactly, where you use it and where you, exactly, so about the process, how do you decide on which system and how do you use it and with what background?

I12  
 Okay, so I think you have to distinguish between AI tools that I use myself for my work and the other thing is to what extent we integrate AI tools into our product. Those are two different things and with the first one, it’s mainly that I only work with ChatGPT, but I think I also indicated that in the questionnaire, I hope that’s okay. Okay, and here in our team…

A2  
 You’re definitely not the only one there

I12  
 Exactly, in our company most people are very tech-savvy and early adopters, so they are very curious about technologies and there is a big awareness that this is something very new and that you definitely need to try it out first. We then had special events for it, which we organized ourselves, and repeatedly asked, hey, how could we use this now? What could you use it for? And I think one tip back then from one of the developers, who is also a very enthusiastic ChatGPT fan, was, well, just keep it open all the time and see in your daily life where you could use it, just reflect a bit on it. And that’s what I’m trying to do now, whenever it comes to writing texts or summarizing thoughts, I try to use it, more like, how does my workday go? Can I actually use it here? I think that’s how I’m working with it right now. On the other hand, when it comes to product development, so to what extent we use AI tools, it’s a mix of a strategic initiative, meaning the company knows through market research and observation that this is the next trend and you can’t miss it because otherwise others will do it and you’ll fall behind, so that’s why you jump on it. And this sort of floats around in the company among many people and then there are ideas that pop up like, we could use it for this, other competitors are using it for that, could we do it too? So I think it’s very much looking outward and comparing, and at the same time having many people in the company who immediately think of solutions and how it could be used. And I think the challenge for me in my daily work is to figure out what are really valuable things and not just solutions floating around, and then capturing them, structuring them, seeing which need they can actually satisfy. That’s basically my task. So in that sense, I come into contact with it more conceptually, but also, if I can proactively initiate something, like hey, we have this need, can we solve it with AI? Then I go to my developer and ask, what can we do today? What can we do already? What are you working on? What are you all experimenting with? So it’s a very complex situation, with many actors, no fixed process, many people interacting a lot. And then the trick is to identify the valuable things.

A3  
 And that would be the discovery part of your job.

I12  
 Yes, in the broadest sense, yes.

A3  
 If you can even separate the two so clearly, probably not.

I12  
 Yes, you do distinguish between delivery and discovery, and I think what I just explained, the internal part, goes a bit beyond that. Product discovery would actually only refer to the value for the customers. But it is actually a, it’s a hard-to-grasp concept, but it’s a concept that can also relate to a mindset that applies to the whole company.

A3  
 Okay, all right. You already said that you come into contact with various AI technologies conceptually, but even when it gets to the developers, it becomes a bit more concrete. Could you describe again the concrete points of contact you have with AI technologies and especially what challenges you perceive in using AI technologies, and maybe also in selecting AI technologies?

I12  
 I think we work exclusively with Chat-GPT and I think Gemini, but only because it is cheaper. So, there isn’t much choice. I think that happens with the developers, they choose it technically. I don’t have much to do with that, and how it works for us is that there is some group of people who experiment wildly, but it’s very technical, it has nothing to do with designers, it’s not with POs either, so they do it alone in their little room, experiment, and then say, yes, we can now do summaries, automated summaries with the help of AI, and yes, we have now improved this or that feature or it has become faster, so they do things like that. And I think then there are, so they do it from a technical perspective, and I imagine it like this: they dig from one side, and I dig from the other side and say, well, for whom is this? And for me, I think it’s not that important which technological approaches they have, but it is only important to me whether these summaries really work as well as necessary to be able to place trust in this tool, so that we can really deliver on the value promise of productivity, or is it nonsense? Right? And I think then one digs from one side, does it have value for the customer, and the others dig from the side, yes, is it technically possible, and theoretically someday, if digging is done from the other side, that would be the designers, how can we visualize this well in the interface? And so we all somehow dig our way toward this AI solution.

A3  
 Very interesting. Yes, I think in this digging together, we are also approaching our vision a bit. But first, before I show you our solution approach, does anything spontaneously come to mind on how we could make this work a bit easier? Especially this, I would say, digging toward each other. For example, from the perspective of AI research, what could exist for that, or also more generally?

I12  
 First, we would have to consider, what is actually the problem? So, if we are all digging our trenches and I don’t even know if it is a problem or simply not part of the process, that we all have to align, that is, be clear on what goal we are currently working toward, and have all these people work on the same thing, otherwise in the end something comes out that is unusable, very expensive to develop, too slow, or whatever. So the challenge would be, how can we ensure that we all have the same shared understanding and are all working on the same thing?

A2  
 Yes, a follow-up question on that. Let’s say they now somehow found five to ten products and stuff. What information do you need from them? What is important for them to give you? For example, what the product can do and how well it can do it, or what are the, how do you say, features that are relevant for you in your daily work?

I12  
 So the prerequisites, so to speak?

A2  
 Yes, yes.

I12  
 Technical performance, so how well does it work? In the sense of, does it work fast enough? We are in the telephony area, so these are phone calls. If we stick with the example of summaries, it also shouldn’t take too long for the summary to be created after the call. So performance in the sense of how fast is this solution? But also in terms of quality, in the sense of how well can the technical solution summarize the conversation? If we just stay with this feature for now, there are other ideas as well, but I think this is something you can…

A2  
 You can also feel free to think completely freely. That is of course a nice example, I can relate to that. But you can really, no idea, if you could wish for… also tell us, what would that be? Maybe it’s technically not possible, but what are some, we are interested in your ideas.

I12  
 So what would I need to know about the technical solution in order to be able to work with it? That’s the question, right?

A2  
 Because they could just send you the code and you probably can’t do much with that. That’s why there are probably some features beyond that, which would be more helpful to you in your work than just the source code.

I12  
 And the feature is currently the guy who is explaining it to me right now, what it does. And he just keeps saying, yes, it’s so simple, I’m actually embarrassed that I get paid for it, because I just keep prompting and doing something, I don’t know, and I’m completely overpaid. I had to think for a moment about what I need there. Because I think I’m rarely in a position where I want something from them, but rather the other way around, I get something from them, and I tell them what I want. And then they are supposed to build it, so to speak. So it’s like that. When we are in technical development, it’s like we give requirements and they are supposed to build it, ultimately execute it. That’s why it’s not so easy to define.

A2  
 But it can also be from that perspective. So if you say, it should somehow, you said, summarize texts, conversations.

I12  
 Let’s see. So fast, as data-protection compliant as possible. I don’t know how that could work, but I think it is important to many people.

A2  
 Yes, that’s a good point. We hear that quite often.

I12  
 I can imagine that. Data protection, I actually think from my practical experience and what I did in academia before, this where does it come from, is not so important. So how the model works or this, which sources feed into it, is not, I think, part of my practical everyday reality.

A3  
 Can I start briefly on that? What do you mean by where the model comes from? For me that would also include, for example, Google trained or created this model, or Amazon. Do you mean that, where the model comes from? The developers, basically.

I12  
 If we move away from summaries, for example if I think and ask the question give me the ten most important X, then there is no source listed or anything and then

A2  
 So in the sense of XAI, like an explanation

I12  
 Exactly, in the sense that, I don’t know exactly where you got this information from, but I think it doesn’t really matter in my practical world where it comes from. Rather, it’s more like, a suggestion comes up and I see how I work with it. It’s different from a Google query. But now back to the other topic. I don’t find it that easy to think about it. Or to speculate about it, because I think effectively it’s about performance, that it’s fast, that it works reliably, and that the names in the summaries are spelled correctly, and that this productivity promise can be kept. And this promise arises in the sense that either you are indirectly promising that you don’t have to write this summary yourself anymore. Or something like, in the future, you can always track what you have discussed and so on. In that sense, since that is given, that is my only requirement for the technical solution. And apart from data protection, which is already important to many people, I don’t think I see anything else right now.

A2  
 That fits, so we said at the beginning, there is no right or wrong. And I think what you have said so far fits very well with what we have heard before. But I think we are moving on to the next part.

A3  
 Absolutely. Exactly. So, of course, we also thought about this. I will briefly share my screen and we also considered, yes, how can we even communicate information about AI models and which information about the AI models should be communicated? I will go to this button. I hope you can see the label clearly. Yes, right?

I12  
 Now, yes.

A3  
 Okay, now it’s visible. And without priming you any further for now, I would like to ask, what do you see on this label? What is your spontaneous reaction?

I12  
 It looks like an energy efficiency label or something for buildings. That’s my first association. I don’t even know where that comes from. I think it comes from the shape, that it looks like that, and also from the color scheme, so to speak. Like this light, I don’t know, turquoise or whatever you call the color of the border and so on. And also, in building energy, people probably talk from A to E. And it has this technical touch here, what you see. Otherwise, I feel this is about the visual impression?

A3  
 Anything that comes to your mind, yes.

I12  
 So, visual impression, technical, something with energy, and apparently A is probably very good and somehow economical, and below that I see far too colorful icons for such an energy label. I already don’t like that. And what do I see there? I see a battery, power draw per inference, so the energy consumption, and the battery is green, so it’s probably low. Corrupted robustness, I can’t make sense of that. Also a very interesting percentage with many decimal places. Same with top 1 accuracy and running time per inference. So I assume inference is probably the queries you would send, for example, to ChatGPT. And then how much energy, how accurate, and how fast it is, but robustness is maybe something like consistency, like how often I get similar answers with the same query. And the color of the icons probably indicates good or bad. And why is it A, if two? Because two icons are green and one is red and one is orange, I don’t know.

A2  
 Good question, good question.

A3  
 Very good question. You are not the only one who has asked that. I can try to shed some light on it and briefly explain what you see here. Unless something else spontaneously comes to your mind, a question or something that catches your attention. Yes, I? You mean? Yes. Okay, all right. So, first we see at the top the MobileNetV3 Small. That is the AI model we have labeled here. It is a model, it is an image classifier. That probably tells you something. So it simply says what is in an image, ultimately. That is basically the task it has to solve. And it was trained on the so-called ImageNet dataset. That is a very large sample dataset, freely available on the Internet. And yes, it performs inference. Inference is exactly what you already said. This process, ultimately, in large language models. You ask it something and it gives you something back. That is inference. Yes, and then you already asked yourself correctly, why is this overall score or this big score an A? This score is based on a total of ten different submetrics. Here we see four of them and they all flow into an overall score. This comes from a study done by our colleague XXXXXXX XXXXXXX. He compared a total of 30 or almost 30 different neural networks and measured these metrics for each neural network, and could then establish a ranking. And these individual colors are based on that. So we have a green battery here, and you interpreted that correctly. That means that among these 30, this MobileNetV3 is generally very energy efficient and consumes little energy per inference. Yes, on the other hand, the accuracy is relatively low at 63 percent. But you have to consider that the ImageNet, the ImageNet dataset, sorry, has a total of 1000 classes. That means if you randomly pick a class, you would have one in a thousand accuracy of getting the correct class. So 63 percent, considering that, is not so bad. Yes, robustness, as you said correctly, goes in the direction of reliability. In this case, it simply means that the images MobileNetV3 has to classify sometimes have pixels changed, and some noise, disturbances, are added. And this robustness score tells you at which extent of disturbances the classification results of MobileNet are affected. That means ultimately how robust it is, how reliable it is. Yes, here we see two QR codes. One for the study by XXXXXXX and one for the results. No, no. I think one was for our label website. I don’t even remember. Oops. And here you can see the hardware on which all the tests were run and the framework. Okay. Do you have any impulses from this detailed explanation or any questions?

I12  
 Yes, I have a question. So, you first said that there were somehow 30 icons basically, or 30 criteria, and now there are only four here.

A3  
 Those are 30 different networks, they were compared and there were more metrics, there are 10, and only 4 of them are shown here.

I12  
 My instinct, I think, is the question of who this is made for. Not for me, I think. And I was thinking, okay, who could this be made for? Who in my company would this actually be for? And I would say, it’s more for developers who choose between models. And then I wonder if really these criteria, the four here now, or the, what was it again? Ten, right?

A3  
 Six more, yes.

I12  
 are the relevant ones, because I somehow have the perspective that it’s mostly about costs.

A2  
 Do you mean costs, like how expensive it is to acquire? Like really getting a license?

I12  
 For example, yes. And for example the energy, this Power Draw, whether it really matters or if something else is listed. For example Accuracy or Robustness or Running Time, I see that, because all of that also contributes to speed and quality. And then, I think, there is another perspective, that a request shouldn’t cost too much. Yes. And I don’t know, if I now take someone, a developer, and there will probably be other people involved in choosing, like, do we take this or do we take that, then money, costs will also play a role. I just don’t know, compared to the others, how much, uh

A3  
 Yes, well, in this case, as I said, we took image classifiers, which are all freely available, for example through the TensorFlow framework. You are probably talking about Gemini and Chat-GPT, for which you do pay licensing fees, right? Yes, okay, interesting. Otherwise, we do have the points you mentioned before seeing the label, we could have included them here, right? So Accuracy, how precise it is, the quality, you mentioned that, and here Running Time, the Performance, you just mentioned, how fast it works. Yes, and it’s certainly more technical or aimed at developers. But it’s an interesting question, do you see some kind of communication use for developers with this label, that they might approach you with it, or do you see it as very irrelevant?

I12  
 Nice to know, not a must-have, I think it’s more for technically inclined people who are really interested, but for my work, for my part of product development, it’s not that relevant.

A2  
 Because you don’t actually make the decision, right? So someone else chooses, or you just say this is top or flop, show me something else.

I12  
 Because I think… No, not exactly. Because let’s say I get the label and it just says Gemini. And then there’s another label, maybe B or something, and it says ChatGPT on it, and it’s the exact same criterion. Then it doesn’t tell me whether it delivers the customer value I need. Because I don’t care how much Power Draw it has, for example. I see it most with Running Time, Corrupted Robustness, and Accuracy, but if… I care about the user experience, which means I can’t judge whether very good Running Time and medium good top one Accuracy will result in a “this feels high-quality” feeling for the user or not. I think I cannot judge that with this information, so it wouldn’t be important to me, and I think to do that I would need the technical person who builds it and the UX person who designs it, and then we would test it, like, okay, is a model that is slower by five percent really noticeably worse in perception than one that is a bit faster, and I think only then would it make sense, but I also think I don’t need this label to judge that.

A3  
 That’s still a very good point you bring up. I would like to just talk about the second label we developed here. From our perspective, it’s of course very suitable, maybe very well suited for comparing different networks. In this case, we have MobileNetV3, which I just showed, on the right side, and EfficientNetB4 on the left side. Even though this is very technical, I would like to ask you again, what are your impulses on this if you have a comparison of different AI models.

I12  
 I think the first impulse is that I haven’t really thought yet about what this A to E actually means. It says, well, my first thought is, if I think now about what it could mean, it’s the quality of the model. But even that is probably a difficult concept, what that could be. Or it consists, in your definition, so to speak, of these ten criteria, right? Because it is calculated like that. And then you would have to look at what these criteria actually are. I think it makes sense in a way, I just thought of Nutri-Score and pizza. That I am very influenced by this Nutri-Score in my buying decision. So, I do kind of, ah yes, that is C, that is A, I’ll take the A. But only if it is in the same product category. And I wonder if a similar way of thinking would be appropriate here. So, to use this scale from A to E for comparison. But what limitations come with it if I apply it now? Like with food, the limitation is that it only applies to the same product group. I can’t compare chips with chocolate. And I wonder if a similar way of thinking also applies to AI labels. Whether it makes a difference if I, how did you call it again, that image thing

A3  
 Image classification.

I12  
 Exactly, image classification, that’s exactly what I wanted to say. And in LLMs or so you compare them with each other. And also within LLMs, for example, there are limitations again that make it harder to compare. But I don’t know, that would be a bit wild speculation.

A3  
 Yes, exactly, we currently only have these labels for image classification. That’s basically through ImageNet. I would say, if you’ve worked with machine learning, then you know it as an image classification dataset, and through that it should be identifiable what task is ultimately being performed. But many people have also talked about large language models, actually most of the ones I’ve seen. There are of course benchmarks. But it’s not as developed as it is here. There are numbers you can use to measure quality. But in image classification, which has existed for much longer, the research on how to measure it is of course a bit more advanced.

I12  
 That’s interesting because large language models are much more accessible to many people. That was my first thought. And the other thought is that it shows even more that these labels are rather for technicians, because you assume that you know the datasets, right? Like ImageNet and so on. And for me, that would be irrelevant.

A2  
 Yes, not necessarily. I think ideally it’s really people who haven’t studied computer science and haven’t already come into contact with machine learning, but who, I don’t know, are doing engineering and want to use an AI solution for their product. And they need a certain, for example, exactly, they need image recognition, and they can now, I don’t know, which library do you use, XXXX? GitHub? I don’t remember.

I12  
 Tensor Flow, for example.

A2  
 And then they can make a selection from how many, how many, thousands of models and then pick the suitable one. But I found it very interesting what you said. It was very enlightening, the comparability. That’s something I think hasn’t been much discussed in our conversations yet. The comparability of different categories of systems.

I12  
 I was still wondering if… This thought came up when you said, as a developer you go somewhere and get it. But I think, as far as I know, it’s a lot of experimenting and trying things out. And the question is whether something like that helps when experimenting or if I don’t even look at the label just for the sake of experimenting, but rather say, I’ll just try any… Or maybe yes again. I’ll take one from E and one from B and see how it behaves. So my first impulse was, yes, of course you always want the one with A, but maybe not. Maybe you want a salad, kind of. So sometimes here, sometimes there, because in my software development I want to try things. I want to explore what technical possibilities exist and so on. Maybe that’s another aspect that plays a role for the purpose of these labels.

A3  
 Yes, so basically as a help for pre-selecting models, if I understood that correctly.

I12  
 Yes, although then again the question is, I would also try things out without a label and look for something, and then the question is, what is the difference if I then have a label?

A2  
 Hopefully you can then pre-select more informed.

I12  
 Yes, but I wonder whether that is so to speak inherently part of experimenting itself, that you don’t care. So I also have to think about it, I don’t know.

A3  
 Okay, yes, thank you very much. That is definitely a very good, yes, new aspect, which we hadn’t had before as a use case for this AI label. I would still like to ask, is there anything here that bothers you, that really strikes you negatively, or that you just find very hard to understand?

I12  
 What bothers me… well, I don’t know exactly, what bothers me is that the label… I wondered, is there an expiration date for the label, because ChatGPT and the LLMs, for example, also change. To what extent is the label updated, how fast does that happen, because if I have a label then I have standardized processes behind it, so there is somehow just a certain bureaucratic structure behind it, and I wonder whether the development of these AI models is exactly the opposite in terms of their dynamics, and I think that bothers me, how I can represent this technical bureaucratic perspective on something that is not very bureaucratic, I can’t think of another word to write it now but

A3  
 Almost anarchic at times, the development.

I12  
 Yes, exactly. It emerges first and then it is regulated. Sometimes it happens the other way around, but software development is so dynamic that its speed is too fast to pre-regulate, so to speak, or only to pre-regulate.

A3  
 Very, very good point. I mean, we have this Issued July 24, which is supposed to represent that this test was then carried out, or yes, then this classification was actually decided. Yes, have we thought about how we want to update it? I don’t know, XXXXX, actually that hasn’t really come up, not even in previous interviews.

A2  
 No, I think it came up once that someone wanted to know, yes, when is the label from, but a process for updating it hasn’t really been discussed so far.

A3  
 Yes, but we will come back to that in a moment when it also concerns, yes, who should even issue it? I would like to…

I12  
 Oh, just a quick note on this expiration date. If it is too short, then the information doesn’t matter to me. And if it is too long, well, I think there is a sweet spot between it being long enough that I can trust it is reliable, an interesting piece of information for me, but if the label is updated every month, I wouldn’t even look at it. Then it loses credibility, just as it does if the period is too long. So I think there is a tension there.

A3  
 Good point, yes. Okay. Maybe it would be interesting at this point to ask, do you have anything else that bothers you about the label?

I12  
 I think it all comes down to the question, who is this supposed to be for?

A3  
 Yes, okay, I understand. Then the question from a UX perspective would be, well, what could be done differently with the label, what would you improve? Maybe also with regard to the question, who is it aimed at? Maybe broken down by target group, what would you change here?

I12  
 I’m not a designer, right?

A3  
 Yes, that’s true, yes.

A2  
 But you do have a good design eye. I can tell you that.

I12  
 Yeah, I don’t know. It depends on who you compare yourself to. I think I would approach it differently. So my first thought is that focus is important. I don’t think one kind of label can address everyone, rather different target groups have needs for different information. And the way it’s presented here now, my educated guess is that it’s more for software developers. Probably, I’m not one myself, but maybe they don’t care about how much power draw it has. And if you ask me, I don’t know, how it is for designers, whether that’s important to them. So from my perspective, performance and speed are important, but even there I would need a different kind of information. I don’t need this 59% or 68% robustness, for me this information only makes sense in the real experience, so to speak. In relation to real experience. Um. Because the… The developers get excited, for example, if they can establish the call 0.2 seconds faster. And I think, it doesn’t matter. It doesn’t matter, you get excited, it’s technically very complex. You get very happy, but at the end of the day, two people are talking to each other. And I think a similar logic applies to my role with these labels. In the end, it’s about creating an experience. And for that, yes, in the telephony area I basically just need to know, does the summary work, does the sentiment analysis work as it should, does the topic analysis work as it should, is the topic analysis deep enough in its, um, in its semantics so that customers can make use of it? Or is it too superficial? Those are actually the interesting questions for me and not how accurate everything is. Also, yes.

A3  
 Yes, I definitely understand that from your non-developer perspective, that you have different requirements. Are you familiar with types of communication where this kind of information has already been presented to you? Or is it something you simply wish for that hasn’t yet made it into your everyday work?

I12  
 This is still very new. I would argue that most companies are currently learning how to work with AI tools and how to develop with them and price them, for example. And I think for companies, the question is always, how many euros do we put on top? And in the end, does every company have it, so that we can’t add another euro because it becomes standard, so to speak? I think these are the questions many companies are asking themselves right now. But also internally, how do I make sure my employees actually use these tools? How do I introduce them? How do I avoid falling behind the competition? And that’s why, at least from my experience, a lot is being experimented with. I think a lot actually happens interpersonally, so to speak. More in explaining, more in working on it together than on artifacts like the label.

A3  
 Okay, I understand. This goes a bit more in the technical direction, but we also looked at a few types of communication forms that exist, which are basically alternatives to these labels. I will go through them very quickly. There are usually scientific papers on the individual AI models. In this case, it would again be about MobileNetV3, written by a Google team. Then there are so-called Model Cards from Google, where you can also find lots of benchmarks, as they call them, so ultimately statistics about these AI models, which are summarized here quite well. Then there is Papers with Code. This is a site that links all papers that cite MobileNetV3. And if you scroll down here, you can see which publications in science have already used MobileNet in some way. Then there are blog posts. Here we have an example from Towards Data Science, but Medium.com is also widely read by developers. Then there are technical libraries, really the programming languages. There you can see technical details again, how to ultimately implement a MobileNetV3. This is free at this point. And Fact Sheets, which are basically like these Google Model Cards I just mentioned, where you also have information in text form, broken down into different categories, for example. Bias is quite interesting here. They list how fair the model ultimately is. Issues like potential discrimination that models can produce are also addressed. And again here, how robust is the model? What are inputs? What are outputs? And various metrics. Maybe finally, what is your impression now that you admittedly saw these six different forms of presentation in a quick overview? What do you notice here compared to our label?

I12  
 So I think it is, as I see it as a practitioner, as I would divide the world in my professional environment, there are the developers and there are the so-called product people. People who take care of the technical solution, and I think almost all types of communication relate to these people, but there are also these product people who turn the technical solution into a real product and make it successful. And I count myself among them, I am also a product person, so to speak. And I think in this product world, blog articles and trends and what others are doing play a very important role. That’s why the blog in the sense of, here I am an author with expertise and I have looked at these technical solutions and see the following, kind of doing a form of science communication, plus trends, and what other influential bloggers are saying, is a super important source and is actually used, as opposed to, I don’t know, reading papers. And the label, I would say, is an intermediate solution. It is not detailed enough to reach the level of a paper. But it lacks, so to speak, the... I don’t know, for example, who created this label. How this, how can I interpret the opinion behind it, because it is an evaluation and I cannot classify this opinion. And I think with blogs it’s different, because about trends, if it’s a particularly frequently read blog, there is a community around it, I can interpret it better. For example, if I know, ah yes, look, person XY, I have seen them at conferences, they also write blog articles, they also do workshops. If they say something, it is an opinion based on their experience that I can trust. But an AI label, like in your draft, which looks rather bureaucratic, a kind of TÜV, has to first position itself. Okay, who am I in this world? Am I this filter? Am I a kind of science communicator? And for whom am I actually doing this? Exactly, I think the impression is a bit about exploring the technical solution, the technical possibilities, but also the product people who sit on it and turn this technical solution into a real product.

A3  
 Okay, interesting. Yes, you just asked yourself that as well, which is what I am also curious about. Who do you think should issue such a label? You just mentioned that personification is an advantage with a blog post, so that you have respectable people. Could something analogous be established for these labels? What are your thoughts on that?

I12  
 So do you mean whether it should be institutional, like TÜV, or more like Professor XXXXXX, who kind of makes these labels, or how?

A3  
 Both would somehow be an interesting consideration for sure.

I12  
 A good question, because I think

A3  
 Or it doesn’t have to be an individual.

I12  
 It could also be a group, right? Yeah, yeah.

A3  
 Yes, rather a group, that would be more realistic.

I12  
 Yeah, also funny if it were one person who then had the power over it. I think it depends, so yes, there is this expert argument, I think, ultimately. So I can trust this institution, this person, so to speak. If the name is on it, that is an argument in itself that I should trust it. But at the same time, I also think there are the processes behind it, which are connected to this person or to this institution. So with TÜV it is like this: I have very clear, strict processes that all AI models go through and try to establish maximum comparability. And with one person or a group of people, it is more of an argumentative process, I would say. And now one would have to consider which of these two types would be better suited for AI models, or with which advantages and disadvantages it would come.

A2  
 Yes, nice comparison.

I12  
 And the question now is, for that you would have to understand AI model development better. So is it something that… My first impulse would be to go more in the direction of community people, because it is such a fast development, because it requires a lot of expertise. My guess would be that an institution cannot keep up. But these are all pragmatic considerations, I would say. So… yes.

A3  
 Maybe a third consideration would of course still be a self-commitment and that the model developers themselves take it on instead of a community.

I12  
 Yes, but then I think it is no longer a label. I think that would remove the idea of the label, I would say. Because it is a bit like… I think the companies behind it… I am mainly thinking of companies and things that are marketable. Yes, perfect. Not the scientific ones…

A2  
 No, it is also supposed to be about the discipline.

I12  
 Is there just still a way to use it for self-marketing? Similar to, I don’t know, I try to position myself as an organic product, so I don’t write E-numbers on it, but the nice, understandable, written-out names, and I would see a similar analogy here and rather see a danger in that. And the label, so the purpose of the label is still to establish comparability, and I don’t know if companies would intrinsically want that.

A2  
 Label-washing or something.

I12  
 Yes, maybe.

A3  
 Yes, I mean, with the Nutri-Score it’s also not mandatory yet. That’s also something companies decide themselves. There, it’s the clear rules that can be followed, which then lead to a certain classification. People also somehow trust that, even though it’s the companies themselves…

I12  
 Yes, I think the bigger factor is that it’s very easily accessible information, very visual. It’s very simple in that sense. Green is good, red is bad. There are five levels in between. If I take the middle, I’m not really doing myself good or bad. And the fact that it’s, for example, only comparably limited is probably not even known to many people or not important in that purchasing decision. If you stand in front of the freezer, you don’t think, now I have to start a cognitive process here, you just look at it. It’s visual. Yes, there is green, there is red, I take the green. That’s a different decision situation.

A3  
 The comparison definitely matters there. Yes, okay. But then the final question, speaking very generally, how useful would such a certification be for you, especially regarding your workday and maybe also the developers you work with who could use these labels?

I12  
 Irrelevant. I see different reasons. I think it’s not very important for my own role. Unless, well, the only argument that comes to mind a little bit, because I’ve talked so much about this power draw, is that it gives a kind of environmental awareness, similar to data protection. And companies can position themselves with that. Keyword greenwashing, but it can also be taken seriously. And maybe it’s an interesting piece of information I could pass on to customers and say, hey, we have the greenest AI here. Wouldn’t you pay 50 cents more for that? And I think, I’m repeating myself a bit, these other criteria, which are listed numerically, interest me less. What I care about is what impact it has in real life. So how is 59 compared to 69? And the subjectivity that comes from the application. That’s what matters to me, and the label doesn’t capture that. And on the other hand, if I think about my developer colleagues, I would guess it’s also less important for them. Maybe not irrelevant, but not very important, because they work in a way where they try everything. And I could imagine that in bigger corporations, where maybe things have to be proven, and perhaps there will be regulations someday, it could be useful information to support the existing processes. Maybe there will be a GDPR for AI one day. And then companies will have to prove that they are compliant. So I see these labels more as something bureaucratic and structural, for proving processes, rather than providing information, at least for my role.

A3  
 Okay, thank you very much. XXXXX, do you have any follow-up questions?

A2  
 Wow, probably. Okay. No, I mean, you explained very clearly the advantages and disadvantages and why it’s not relevant for you. I think that’s great. We are also interested in these boundary conditions. So it’s not like it was a waste of time or anything. I found it very insightful.

A3  
 And especially this clash between, yes, we have a bureaucratic label and we have this completely dynamic AI model development process, and everything is always changing. That was a very good point. You really made us think. Okay, good, then I would say, yes. I’ll stop the recording.