I2

I don't know if you, XXXXXXXXX, can record this parallelly. I had that once, too. Depending on the Zoom version, this is possible. Then you can be sure that nothing will be lost.

A3

Yes, so we always have double shots. I even use my phone, because I’m not using a headset today.

I2

The quality is okay. We can do that.

A3

Exactly, so hopefully it's safe. Okay, then I would like to start. Firstly, please introduce yourself. What is your background?

I2

Yes, my name is XXXXXXX. I have an academic background in economics and I'm currently pursuing both my Bachelor's and Master's degrees with a focus on technology and innovation management. Since 2018, I've been involved in two XXXXXXXXXXXXXXXXXXXXX, always working in the field of technology and innovation management. In addition, I worked for over three years as an AI manager, where I focused on artificial intelligence and generative AI, especially from an application-oriented perspective. I also contributed to studies on language model development even before ChatGPT emerged, so transformer models. This has given me a solid understanding of neural networks, deep learning, and reinforcement learning, of course not just from the development side. I'm still actively working in this field. My current research focuses on voice assistant systems and generative AI, as well as decoder systems designed for example for specific applications, such as technical language or medical applications. Beyond that, I’m involved in projects related to big data, data lakes, and value creation through algorithms. So, exploring the future use of AI and potential GenAI use cases.

A3

Okay, very good. Then you've already answered a few questions from my side. So your previous experience with AI is probably not from your studies, but only in practice, now in the work at XXXXXXXXXX?

I2

Yes.

A3

Ok. And are you aiming for a PhD right now or are you satisfied with your master's degree?

I2

No, I'm in my fourth year of PhD in the field of technology and innovation management at XXXXXXXXXX. I do this part-time. I’m a sub-project lead in a large research project, but the main part of my work is in IT consulting and development. I support various industries and sectors in the areas of big data, data advisory, and artificial intelligence, especially where topics like the ones you mentioned are relevant, such as labeling, certification, building trustworthiness, and responsible AI.

A3

Okay, that sounds like quite a varied workday. But could you still describe a typical day-to-day work process? What do you do more regularly, especially when it comes to the development and deployment of AI products?

I2

I don’t really have a typical day. If I describe it over the course of a week, I’d say one day might involve research, project management, preparation of clinical studies, user interactions and testing. Then it could be things like writing applications, or something like BIT preparation, meaning trying to present an offer or proposal to a client. Then there’s actual operational consulting work, process analysis. It might also include internal tasks, such as quality management or preparing for the use of generative AI in administrative software, for example, in something like Conference Tira. So, it’s really very diverse. And then, as mentioned, the project topics are cross-industry. They range from questions like what strategic use or role one wants to assign to artificial intelligence in the field of generative AI, to very concrete things that don't yet have, let’s say, the maturity level to involve any kind of algorithmic processing. In those cases, it's more about data analysis projects. So that's the range. I hope that's clear enough for you.

A3

Okay, yes. What is the part that is closest to the developer perspective? Or do you sometimes work with code or is it more?

I2

So, in most agile projects, we usually have a mix of, let’s say, proxy, program manager, product owner, and development team roles. And I always work as a kind of link between them. I do look into the code from time to time, but I don't have to do any programming myself. Instead, I take more of a triangle perspective: if one side is the application perspective, another is the requirements perspective, and the third is the initial development perspective. Depending on whether things are going well or not, I might have to dive deeper into one side or another. But I’m not the one writing code. I need to trust that what my colleagues are doing is solid. My role is more about following a problem-solving approach: figuring out what makes sense now, whether it’s troubleshooting or defining the next set of requirements and breaking them down into development tasks.

A3

Okay. And what are, maybe also in this constellation, this triangle you just mentioned, what are the biggest challenges you face in your day-to-day work?

I2

So, in projects in general, it's first of all a mix of: are all departments actually communicating with each other, meaning, is business aligned with IT? Are these projects coming from a specific department, or are they overarching projects? I think one major hurdle is the lack of interdisciplinarity. Another is the lack of pragmatism when it comes to assessing the use of technology, and also when it comes to keeping the organizational machinery as small as possible, which, in practice, doesn’t happen often. That means things are often approached in terms of obstacles, limitations, or bureaucratic barriers. And I’m not even necessarily talking about DSGVO, but rather about questions like: Can we even do this? Is this technically possible? Instead of taking an iterative approach and building things in small steps, like baby steps. But usually, everything gets fully thought through first. And I believe that stems from... if I may add a bit of interpretation here… I think this is because we have always worked very strongly with this principle of business process modeling in business informatics and perhaps slightly older colleagues would also like to have these process models. First, the whole process must be worked through. Even though, I would say, technological approaches, especially algorithmic processing or GenAI, could drastically downsize these processes. So there’s a real gap in perspectives here, and in most cases, I feel like the more traditional view ends up dominating. And that, in my opinion, is one of the main hurdles across many areas.

A3

Yeah, I mean, that’s a legitimate issue from the field of business informatics, these different perspectives. I’d like to come back to one point you mentioned earlier. The assessment of technology use. That’s really what it’s about, or at least that’s the problem we’re trying to address a bit with our vision. Could you go into more detail on what exactly the obstacles are that you perceive here?

I2

So, this is very specific for technologies. If you'd like to move on to AI and Gen-AI now, go ahead. Okay, so I think we really need to make a clear distinction between the current AI summer that we've had for the past two years and the time before that. I would say that right now, generative AI is a very hot topic. And I’ve already heard some major investment bankers say that Gen-AI isn’t delivering on what it’s supposed to.

So, massive efficiency gains and, as a result, increased productivity, which hasn’t materialized. But that’s complete nonsense, because the technology has only just recently started being used seriously or even tested in real-world applications.

[3 SECONDS MISSING]

And it's slowly starting to fade again, if I can put it that way. And you can notice that in projects as well. I’ve heard the question a few times, ‘So, what comes after AI? What’s next after AI?’ And I think, on one hand, the question is understandable, because AI has been so polarizing over the last few months, maybe even the past year and a half. But on the other hand, it's also completely absurd, because we’re still relatively far behind in terms of development with this topic. It means, in the past, AI was difficult, just individual projects. People would try things out, and it was very exploratory. Maybe researchers were involved, and they tested things. Kind of like what quantum computing is now. And generative AI, as a representative technology for artificial intelligence as a whole, is already being treated a bit like a stepchild, because the promise of efficiency improvement hasn't quite materialized yet. And also, what I believe is related to your topic of labeling, there is a loss of trust because it has become known that data was used for training that might not even be public, there were no releases, and permission was not asked. The issue of hallucinations is also highly relevant, but for most use cases that aren't in critical areas or critical infrastructure, it can initially be disregarded. And nowadays, there are much better models, like Perplexity for research-based generative AI, which is much better than OpenAI. They are even live and connected to the internet, meaning they can even cite the sources they read out. But regaining that trust, I think, you can already see in projects. What’s also missing is a strategic direction for all of these things. Technologies are often presented from the perspective of a specific field or even the CTO or sometimes from the, let’s say, IT corner. They say, 'We can do this well.' But I think that’s the wrong approach. It should actually go from both directions.

A3

Okay, interesting. Okay, what could be a solution to this, for now also this specific problem that you mentioned last? Can you imagine what might also come from AI research, from computer science research?

I2

So, I think that technology projects, or that we, in general, shouldn't view technology projects as purely IT projects anymore. These are the classic points we’ve been aware of for a few years now, but still don’t apply. This means that they need to be interdisciplinary, they require accompanying change management, they need domain expertise, they need technical expertise, and there also needs to be management commitment. In other words, we need something like technology strategies. And I believe that, at least from my experience, these are only partially available in the broader spectrum, because they are very clearly tied to the respective business strategies. And, let's say, in business model development or business model innovation, so in BMI literature, the approach isn't to quickly move away from the core business area, but rather to see how one can evolve from it. I still think that, let's say, if you're working in a manufacturing sector and you have a product, hardware, or a device that you're selling, you wouldn't start developing a technology strategy to decide whether you're going to become an online marketing company. Instead, you'd look at where you're extremely strong and how you can increase efficiency with the technological tools you'll be getting in the future. Strategically anchoring this and tying the know-how to it, I think, is difficult because there is also a shortage of skilled workers in this area. I'm not talking about younger computer scientists, where there's not just an oversupply, as we can see with companies like Telekom, Salesforce, and SAP, which are releasing a lot, but rather, I mean, people with five, six, or seven years of experience who might have a background in AI, have studied physics or mathematics, and are building something like an innovation lab in these areas, which could then be integrated under a CX department as a staff position. I think that’s a bit missing, and I believe it would be one of the key success factors for what you just mentioned as a problem.

A3

And this staff position, when it comes up with AI innovations and develops AI products, how should those be communicated to all stakeholders? What kind of information would you want to receive about AI products? Yeah.

I2

Well, that really depends on whether it's something intended for internal use, whether it processes personal data, whether we're in a production environment, whether there are things like works council approval requirements, or whether we're in a playground or a sandbox setting. I think the Technology Readiness Levels or this scale, those seven stages, are quite interesting. Maybe something analogous could be applied here. I mean, there's the Bonn Catalogue from Fraunhofer IAIS that outlines aspects of trustworthy AI. You could simply fulfill something like this accordingly. But we also have the EU regulations that need to be addressed, with the traffic light system or those risk-level classifications. All of that, though, is something I’d say is more suited for large tech providers, ones that are focused on obtaining broad certification and work through 300 pages of EU documentation. But if I’m now in a company and I actually want to try things out a bit in a sandbox or small playground setting, then that’s exactly what we talked about at the beginning. You’re faced with a lot of hurdles, a lot of barriers, a lot of bureaucracy. And that’s exactly what I want to get away from. So what’s really needed is something in between, a kind of middle ground, maybe even an intermediate layer. Similar to what you see in the medical field, where they often say: we’re doing use cases, but it’s not for clinical use. And it could work the same way in other areas. And if there were simplifications, like a traffic light system or, quite pragmatically, a questionnaire tool that’s interactive, something like: I have a use case, I want to roll it out internally, I go through 50 questions, get an extract or an evaluation at the end, and then I can say: is it problematic or not? Which committee needs to be involved? I think such tools, I don't know if they already exist on the market, but I can imagine that they would make work easier for many companies in the technology agreement.

A3

Okay, great, you’ve already brought up some important points, and with the traffic light system, you may have already referred to something you saw on the website. I’d now like to share my screen, and hopefully, you can see the logo, which I’ll enlarge for you. This is exactly the traffic light system, or actually, as some have already mentioned, it reminds people more of the Nutri-Score. The design itself is also inspired by the Nutri-Score. And without priming you any further, I’d like to ask you: What is your spontaneous reaction to this label? What do you see here? I can make it bigger.

I2

Can you? Yes, exactly. I need it a bit bigger. So, I'm a bit biased myself because I deal with these labels, for various reasons, but what really stands out to me here, quite openly, and XXXXXXXXX, I'm taking your word that there are no right or wrong answers: two QR codes, I’d say, are too many. Either you reference just one, or you need a sub-description explaining what each one refers to. The codes themselves don't tell me anything. It just says 'issued July 24', so I assume that’s when the label was released. And here I can find more information, okay, not bad, maybe the positioning could be changed. Infer MobileNet3V small, fine, all good. TensorFlow, I can still understand why that’s shown here. Robustness is rather weak, so, yellow. But I wonder: why is this one shown as a filled icon, and the others aren’t? I’d really wish for a more consistent design. Then, Top One Accuracy, and then it says 60%. That feels counterintuitive to me. It should be presented differently somehow. Running time per inference, I think that's initially irrelevant, because it's more of a usability topic. Power draw per inference, if we’re talking about sustainability, I can’t really do anything with that figure. So really, I only see A, which is very good, and with the Nutri-Score, the main criticism I have is: if I compare food systems, let’s say frozen pizzas. Frozen pizzas are usually unhealthy, but one might do a little better than the other, and that’s considered okay. Now, if I think about an AI label in the same way as a Nutri-Score, I wonder: are we in a realm where we’re not doing social scoring, but maybe something like credit financing? If they still use proxy parameters, and I get a very good score just because others are doing even worse, I can’t really judge that properly. That’s why I would have liked a bit more explanation here: what’s the application area, which metrics, why these four metrics, and maybe reconsider the arrangement of the logos. But overall, I think a color scheme is intuitive. One more point I’d like to add, which we’ve already discussed before: there’s also the 'Haltungsform-Score' (how the animals are kept), where the color scheme was disrupted due to lobbying. Meaning: the best score is green, then level 3 is orange, then level 2, which actually sucks, is blue. So basically, if we make labels like this, we need comparability and transparency. And color schemes should be consistent not suddenly having something bad appear in blue. That’s something I find problematic, because I think: if we have voluntary self-commitment for AI labels, it’s going to be difficult. And one final point that came to mind: there’s currently no external validity, no icon, label, or organization that stands behind it. Maybe if I click on it, something shows up, but at first glance, it’s an anonymous label. If you now show me one from TÜV or XXXXXXX or something similar, I could say a few more words about that.

A3

Sure, so we’ll come back to that point again at the very end. Should this really be a self-commitment, or should it perhaps be handled by officials? But I’d like to come back to the label visually for a moment. So, you mentioned it, I don’t know if there was any confusion, Infer, probably tells you, it simply stands for inference. Here we have the name of the AI model. In this case, it’s an image classifier we’re using, which was trained on the ImageNet dataset, I’m not sure if you’re familiar with it.

I2

Ah, okay. Mhm.

A3

Exactly, that’s what it’s meant to be. And yes, this score comes from a paper by XXXXXXX XXXXXXX. It’s based on a comparison between 30 different neural networks, all of which perform image classification. Here, we have the score for MobileNetV3. It’s also based on various measurements for these sub-metrics. There are ten in total, but only four are shown here. As a compound score, meaning the sub-metrics are weighted and combined, they result in this overall score. That’s how it works. Just as some context again. The QR codes link to the paper and more information about the project. And of course, we have the graphic card and TensorFlow framework. With all this information, was there any impulse or thought that came up for you regarding this label that you'd like to share?

I2

Yes, definitely. So, everything is understandable, and I think if I had another five minutes, you'd probably get into the flow of it. I thought that if this is for a specialist audience, I mean, purely developers, then I'd say you could work with it more as it is. But if we’re targeting a general audience or want to make a low threshold offering, which I think is the goal, because if we directly target developers, we wouldn’t need a scoring system at all. Developers could use different metrics. That means: it’s too limited. I need more information. And I see the opportunity to do that with relatively little effort, for example, simply writing up front what the model is, like you said: that’s the model, so Image recognition model double dot GPU Used, Model or something like that. Then remove the icons and instead maybe include five scales, also using the ABCDE format. You’d just apply subscaling and fill them out accordingly. Then you’d have the weightings, two greens, one yellow, one red, and maybe provide a final score. On the right side, you still have a text to name corrupted robustness or running time, with a small label. But then you'd have a weighted scoring and much better traceability. Because right now, I can't tell from the scale: are we 1.5 worse in runtime? Or when is the accuracy, compared to the other 40 models, better? You just explained that to me, but I think 60%... for many applications that’s inferior. For that the image recognition is fine, but those are the kinds of things that immediately stood out to me.

A3

Yes, well, but 60% is clearly already relatively low compared to this MobileNetV3, compared to all other networks, due to the red coloring. Of course, ImageNet has a total of 1,000 classes, which means if you were to classify randomly, you'd have a one-in-a-thousand chance of getting the correct classification, or accuracy. So from that perspective, 63% is still an improvement. But in comparison, and this brings me directly to the next point, we have a second label here for comparison. You can see that, especially when comparing two networks and maybe pursuing different goals, the idea is to recognize the pros and cons. Now I’ve opened the second label, same structure. What stands out to you in comparison? Can you summarize what you see here?

I2

I find that counterintuitive when I look at the two. I can't really tell the color in the bottom right. What is that? Is it a mixed color? It’s either light green or some mix of green and yellow. I can't really see it clearly. I think it's supposed to be the B. Yeah, okay, that's tricky.

A3

Yes, exactly, it's between yellow and green, yes.

I2

And if someone has red-green color blindness, then it's game over at the latest, because then you just have grey on grey. So, it's not really inclusive. I also find it counterintuitive because earlier we were at A, and there we had green twice, so two A's, one C, and one E. And here, if I'm seeing this right, it's two C's, one A, one B. For me, in terms of weighting, that doesn't quite add up, because that would actually mean we're at C, C, A. Yeah, no. So for me, EfficientNetB4 should score higher than MobileNetV3.

A3

Yes, exactly, that's because we have more metrics that aren't shown here. For example, model size is a factor, meaning how large the model is or how many parameters it ultimately has. But also things like Top 5 accuracy or Top 10 accuracy. I’m not sure if you're familiar with that, but it refers to how often the result that the model considers most likely is actually the correct class, or whether the correct class is among the five most likely results. Metrics like these are also considered and have an influence on this compound score, so the overall one. So that's probably where the confusion comes from.

I2

Yes, I also think that's totally legitimate, and I mean, that’s exactly why 'Scan for Further Information' exists. You asked me what my first impulse is when looking at these four, and I would, as a first-time viewer, ask myself why it seems illogical to have hidden information or information that base on a shared knowledge that I can access via the QR code.

A3

Okay, I see. But do you still see, despite the mistakes and areas for improvement, a connection to your work, how this might help you, or at least help the people you work with?

I2

Yes, absolutely. In a use-case evaluation phase, when we decide that we need, for example, domain-specific language recognition or image recognition, then all options are really laid out side by side, and based on that, a selection is made. So, in a use-case phase, in the preparation phase, in the evaluation phase, where we’re discussing how, let’s say, an MVP or a proof-of-concept could take place, that’s where you could compare different models and see how well they perform. Of course, with the idea of using the ones where the metric that matters most to us comes out on top. For example, if we say accuracy and robustness, I’d like to be able to filter based on those and then take the top three and compare them directly. I’d say that would be a huge advantage.

A3

So, definitely an interaction, would you want an interactive label there? Okay, was that just an example, or would those really be the two metrics that are probably most important to you? Especially if we look at power draw and also running time, which is correlated with power draw, when we take another look at those.

I2

That really depends on the use cases we need. So, I’d say, in user interaction scenarios with a broad target audience, robustness can be neglected. What I need there is a very responsive system with okayish accuracy. Of course, if I’m dealing with image recognition in radiology, I want very high accuracy. I would say that 81.27 is unsatisfactory for non-exploratory work areas. I think we need to place much more focus on power draw. Whether a computing system runs on a 30 60 or three 40 80s makes a 20-fold difference, depending on the system you’re using and what else is included. That’s 20 times the energy or wattage being used. And sure, over the course of one year for a single model or system, that’s maybe not a huge amount, maybe 200 euros. But if we’re talking about large upscaling, it becomes a significant energy consumption. And I believe that especially regarding many research areas working on SDG-related topics, it’s extremely important that we start including something like an energy efficiency category in labeling. We should also include something like frugality of system complexity, in other words, how complex the models are written, and how large the models are. That’s something that, in the early stages of technology, is irrelevant. First, the car must drive, how many liters it uses doesn’t matter yet. But at a later point, like where we are now, I think this deserves much more attention. And I would really welcome it if that became possible.

A3

Absolutely, okay. Yes, I totally agree with that. AI already consumes so much electricity today, AI models in particular. That’s why I also think that much more attention should definitely be paid to that. XXXXXXX has already explored this quite extensively in several studies, especially the topic of resource consumption of AI models. Coming back to the labels again: Is there anything here, I mean, you already mentioned it a bit, anything you don’t understand at all or don’t find helpful?

I2

So really just the same things I already mentioned earlier.

A3

Yes. Okay, good. What do you particularly like? What do you like most about it, like, is there something that stood out to you about the AI label?

I2

So, first of all, I like that it’s kind of a card-based system, it gives it a sense of familiarity or recognizability. I thought about how there are these design sprinting approaches where people use card models to discuss and evaluate technologies and areas of application, often working with physical cards. And I can imagine that something like that could be really helpful, for example, if you lay those kinds of things out next to each other or have them presented on the back of the cards. Let’s say we have a use case, image recognition, and we want to recognize bicycles, just as an example. Maybe we’re the local utilities provider and we want to build a new bike lane. So now we have our use case. Then we bring in a few decision-makers who aren’t tech experts, and we simply lay the cards out for them. And then maybe, on one part of the card set, we could show the possible models that could be used, just to make it clear that there isn’t just one model, that there isn’t one single AI or system that can do everything. In fact, there are trade-offs involved, and we may need to experiment. So the question would be: is this more of a data literacy approach, where the goal is to educate and inform people about how such systems work? Or is it meant more as a decision support tool? I think it’s absolutely fair to say it can be both.

A3

Okay. So would you say that the models are definitely comparable with the help of the labels, and especially comparable for non-experts?

I2

Yes, definitely. That should also be the expectation for the label, at least from my personal perspective. But, as you already explained earlier, we still have these sub-dimensions or side-dimensions, and what I personally find missing is a clear weighting. Like here, I think it could simply be that the four aspects listed: Power, Draw, Accuracy, Time, and Robustness, are each weighted at 25%, and then the overall score is calculated from that. If you’re using any other metrics in the background, then that needs to be stated clearly here, or there needs to be a different way of presenting it, like the one shown below, which I mentioned earlier, maybe a table or some other format, because from my point of view, it’s not transparent.

A3

Okay. And can you imagine something like a dropdown button or something like that? That would also go in the direction of interaction, or at least interactivity.

I2

Yes, you can do that, but if you want to make it, let’s say, as a physical product, which I think is a good idea, because workshops also work well in a physical setting, then you simply need to design the card a bit larger or place things a bit further apart, and then see how many parameters you can include that lead to the evaluation. Or use the back side, that’s all fine.

A3

Okay, very good. So, you just mentioned that you could imagine having something like a Nutri-Score for each of these sub-metrics, as I’d call them. Did I understand that correctly? Yes.

I2

Yes, or at least, let’s say, the icon and then a small scale from, let’s say, 0 to 5 or 1 to 5 or 1 percent, it doesn’t have to be a completely new Nutri-Score, but it should show how much has been fulfilled, also in comparison to the others, so that the in-group comparison becomes clearer.

A3

So that would basically be the criticism of the color scale, which is actually supposed to express exactly that.

I2

Yeah, so, the color scale is for me, we have A to E, where E is very bad and A is very good. And if we now have the color scale down there again in this case, that is actually the weighting that results. Now, I see two yellows, one green, one B. So that would place us somewhere between B and C for the left graphic. For the right graphic, we have one yellow, two greens, and one red. So, if I look at it now, we actually have two yellows and one green, so we would also be somewhere between B and C. But then MobileNet gets an A. But if I didn’t have the colors at the bottom, but instead something like a ranking: ‘7th out of 40’, with the weighting N, 0 point N. That would make it much easier for me to understand, or at least more logical. Because the explanation you gave helped me understand it, but I think the visualization should stand on its own, without needing any explanation. And if I’m interested, then I can dig deeper and look up more information. But as it is now, it feels illogical to me. Maybe I’m just thinking about it wrong but that’s definitely what stood out to me right away. I mean, it can’t be rated as A if I’m seeing two greens, one red, and one yellow, right?

A3

You're absolutely not thinking about it the wrong way. Everything you said is totally valid. As I said, there are no wrong answers, it's your impression. And thank you very much for the inspiration and your suggestions for improvement. I'd also like to ask you whether you know of any other forms of communication that serve the same purpose as these labels. And whether you've ever come across something like that in your everyday life, this concept of representing a model in some way.

I2

Besides the Nutri-Score now?

A3

No, I mean specifically about AI models, whether you've ever noticed any forms of communication there.

I2

So what I’m seeing a lot right now is that TÜV and DIN are at least working on certification topics. But what I’m seeing there are very few actual models and mostly just concepts for now. And what I’m seeing much more of are dimensions for Trustworthy or Responsible AI but those don’t really go into labeling, they’re more about evaluation metrics. In the legal profession too. They’re really behind. In medicine, I’ve actually barely come across it at all, because everything there is handled or covered through medical device certification. Extremely little, no. Most of it is, I’d say, technical or even openly accessible as open source. And with AWS or similar providers, to be honest, I haven’t come across it either, no.

A3

Well, we did a bit of research and looked into how this is presented online for developers, and we found six different types of representations. First, of course, there are scientific publications, for example, in the case of MobileNetV3, the original paper from Google, where MobileNetV3 was developed. You can read a lot about it there, and benchmark tests were also conducted for these MobileNets. Then we have things like model cards, in this case from Google. If you click through TensorFlow and look for information about MobileNetV3, you'll find things like benchmarks, details on where it was tested, on which datasets, what the data looks like, how many parameters the model has, and how large it is on the hard disk. Then there’s the nice site ‘Papers with Code’, which links a concept, in this case, an AI model, with the papers that cite it. So, if you browse through it and scroll down, you'd see all the scientific publications that have cited MobileNetV3 in some way. Of course, there are also typical blog posts, you’ve probably come across Medium or Towards Data Science before. And there’s the documentation, in this case, it’s a Python library. There you’ll find all kinds of information, for example, about the specific methods used when calling MobileNet in code. But also metrics, so benchmark tests. And lastly, we have the factsheets. That’s something from IBM for IBM-specific AI models, which provide information about those models based on different categories. For example: bias, how biased might a model be? How fair is the model in the end? Especially when talking about discrimination and similar issues. How robust is the model? That’s something we also had in the label. What do the outputs look like? And so on. So, we’ve put that together now. And if you place the label I just showed you, or am still showing you, next to it, what advantages and disadvantages do you see in comparison? In terms of this way of communicating. You don’t have to go through everything, just mention whatever stands out to you.

I2

To me, this reads like a classic dissemination concept or process, basically how science is conducted and then translated into application. And that’s actually the traditional path: I do research, I have a scientific paper, I have an application case, which then gets published in an accessible trade journal. At the same time, the developers are also researchers. So, after the scientific paper, the content ends up on the company’s website, and finally, there’s a label. That’s how we conduct research as well. First playground, then 'pain of publishing'. And because that takes time, you might quickly publish a few pieces of information somewhere, maybe on a website, then a newspaper. Just to shorten the time-to-market, the time from idea to implementation, to functioning, and to enable faster exchange with other researchers. And I think we need all of these steps. The only difference is the target group. With a scientific paper, 99% of the readers are scientists. When it comes to the descriptions by IBM or Google, I’d say the majority are developers and also scientists. In the bottom right corner, everything you need to know, that’s aimed at developers, users, and maybe some interdisciplinary folks. But still, it’s standard, technically oriented. And then the one on the right is something that abstracts everything further. It’s designed to be more accessible for non-tech audiences, to create traceability, trust, and transparency for what was previously presented. But if I can trace the path back, that’s why you said it links directly to the paper, then that’s really cool. The idea would also be: can you include studies there? Like a 'scan for further information', you land on a dashboard and can choose how deep you want to go. I want the original data, I want to go to the Git repository, or I just want to look at the graphic, maybe take a closer look at [unclear]. Or no, I just want to be redirected to the Google page, or maybe I only want to see the newspaper article. That would be the possibility, because the label is, I’d say, designed to be comprehensive for all kinds of target groups, especially for someone encountering this system, or in this case, the image recognition system, for the first time. So, it’s not really a comparison or contrast, but rather something that complements each other.

A3

Okay, interesting. And again, through some kind of interaction. Okay. Yes, thank you. Then I’d like to come back, finally, to the point you already brought up earlier. So, who should issue such a label, period. What are your thoughts on that?

I2

Yeah, that needs to be tested. I can’t say for sure right now. Institutions that already have a trust advantage in technical fields would do well to keep investing in this area, because it brings market advantages and helps establish their position. I think there will be several players like Google, Microsoft, IBM, maybe even Salesforce and those who might offer their own certifications. On the other side, organizations like TÜV, DIN, and other established companies, also in the German-speaking region, will probably also try to offer certifications, perhaps targeting companies with less, let’s say, development power. So, I don’t think there will be one single label. There will be multiple labels, because the market is too heterogeneous. At the end of the day, there will also be the question: are the label providers still the ones doing things that are state of the art, that are technologically leading? Personally, I don’t know how decision-makers in large or small companies view this. I mean, TÜV is a known name, DIN as well, XXXXXX too, but if I go international, maybe no one knows what XXXXXX even is. And if it says Google on it, great. If it’s a consortium, maybe even better. But really, I can’t say which way this will go. I do think we’ll see a big increase in trust if we have strong partners involved in the labeling or certification of models and systems. It can’t just stand on its own, with new invented company names or proxy names. The logos need to be on there.

A3

Okay, all right. It was very interesting to hear your thoughts on that. Sorry. Okay. Do you generally think certification makes sense? And especially when you apply it to your daily work. Would it be helpful for you?

I2

Well, that’s a very broad question again, because we have to consider what kinds of applications we’re talking about. And when we look at an AI label, we’re looking at models on the one hand, but the application itself isn’t part of that yet. So, for example, if I have an image recognition system, it can be used in either a critical or a non-critical context. That means that technically, I’m just getting an indication that the model itself is very good. But if I’m developing something, I need more than just knowing whether the model performs well or not.  I need to know, maybe a silly comparison, but if we take a car: knowing that it drives fast is good, but as a car manufacturer, I also need to know if the airbags work, if the other systems function properly, if there are seatbelts, and so on. And we need to ask exactly these kinds of questions when using, let’s say, image recognition systems. Do I want to use them in production? What can I track? What kind of data am I getting? In that context, it’s helpful to have an AI label for this area. But ideally, we’d also need, as the EU has at least teased or even clearly stated in the EU AI Act, albeit within a very complex framework, criteria for when something is considered critical and when it isn’t. So, I think having these labels is a good first step. But for me, the next step would be the combination of model quality and application context. What that would look like, I think that would require a whole new workshop with a lot of people involved. But personally, that’s what I see as the most practical approach. That’s why I said at the beginning: something like a questionnaire. I have questions that are asked, and based on the answers, I can choose my labels, or rather, I can choose the models with those labels, and then I move into the application phase. And that might also help make these systems more widely used and improved, potentially leading to a new wave of growth. Because right now, or as I said at the beginning, at least from my point of view, that growth is slowing down, since many people are somewhat frustrated that the systems don’t yet deliver everything that was promised.

A3

And then the last question would be: how would the label need to be adapted to the knowledge background? You briefly mentioned that earlier. But especially when we talk about different stakeholders, different groups, also in your daily work, how do you imagine such customisability... Do you have any idea how that...

I2

Actually, it’s pretty simple. A back side with clearly explained metrics, evaluation metrics or parameters, along with a scale. And then a ‘Scan for Further Information’ with two levels, one written in very simple language, and one more technical if I want to dive deeper. That could also include information from the papers, explaining why certain aspects are important and why the model was rated the way it was. I think a lot of that text can be reused. The introductory explanations for, let’s say, AI beginners, are quite easy to replicate. Then you’d just need to decide whether that content is integrated into a new graphic for each model or not. Maybe it’s enough to keep the layout uniform but provide more detailed information through the ‘Scan for Further Information’, especially regarding the dimensional characteristics. I think that alone would help a lot.

A3

Okay, all right. Thank you very, very much for this input. I’m actually done with my guideline. XXXXXXXXX, did I forget anything, or do you have another question?

A4

No, I think we've got everything so far

A3

Great. That was very interesting. We can stop recording then...