

## Phase 2 – Interview Transcripts

**S1**

00:00:10 Interviewer

So maybe we can start a little bit. I'm curious about how you shift—not like you start to focus more on issues related to AI. I wonder how you start having this kind of shift of focus.

00:00:26 S1

Yeah, you mean what attracted me to the research project on AI?

00:00:32 Interviewer

Yes.

00:00:35

Yeah, so you probably know a little bit of my background. So I did my graduate studies in philosophy, and then my dissertation work is more on structural injustice.

I was mostly interested in having this kind of research project in social, political, and feminist philosophy. One of my dissertation projects tried to use structural injustice to analyze sexual violence. But when I was writing my dissertation, I was trying to think about how other social issues might benefit from this analysis of structural injustice.

That was back in [period of time], and I was, at that time, living in [Location]. Some of my friends are working in tech companies, so we kind of started this conversation—just wanted to know what they are doing.

And there are lots of news media talking about the issue of algorithmic bias. And so I think, like, through the conversation, I realized that, OK, maybe I think that the framework of structural injustice might be helpful.

I think it also helps the engineer because lots of times they feel like people are always saying that they're biased. But they will say there are lots of the practical constraints—how the datasets look, like, this is not... we also wanted to have a dataset that works better.

They have lots of decision-making about, like, they try to track, like, the performance disparity. But there's still this kind of gap. So I think that adopting a structural lens is helpful both for the engineer to have a better understanding of where they are situated at, but also for people who care about social impacts, knowing that there are broader perspectives and many of the different components that contribute to this issue.

So yeah, that's kind of a story that got me into the issue of AI. But once I... the first project I was working on with AI was the paper published in *[Journal name]*.

And then, so starting from there, I realized that there is more I want to think of, and I'm just observing that, like, technology, and especially, like, AI in this era—I think it's playing a much more crucial role in shaping many of other components of social structure. By that, I mean, like, how the resources get distributed, how people understand different norms. We

understand different schemas, like, what counts as, like, a more beautiful, what counts as the morally good one.

So I think technology, like AI as a tool, has, like, keeps shaping how people understand those. So on the one hand, I want to be critical with all this kind of potential negative impacts, but I think, on the other hand, I'm also very interested in exploring or, like, trying to envision what would be some of the more positive impacts that AI might also bring in to shaping the social structure.

Yeah. So that's the overarching theme that I'm seeing and what I'm trying to do.

00:03:42 Interviewer

And I see you have experience in human-centered artificial intelligence. I wonder...

00:03:48 S1

Yeah, [Institution ].

00:03:50 Interviewer

Yeah, I wonder, because I researched about that, and I think, for me at least, human-centered is what designers do. I wonder how it's working there, and what were you working on there?

00:04:03 S1

Yeah, so actually, [Institution ] is a very large institute, and it's kind of just bringing people from different backgrounds together. And honestly, I think when they say, when in [The institution], when people say "human-centered," they intentionally... I don't know how designers understand that, but I think at [the institution], when they use that term, they kind of keep it intentionally vague and intentionally open for broader interpretation. So, they definitely want to say that, yeah, when we are thinking of the design of AI or all this kind of engineering stuff, we should make sure that we know there are... I think they do not necessarily use this term, but I think they want to emphasize the kind of social and technical connections. So, like, human-centered in there... but I think what they mean by human-centered is probably also different. And some people are—and I think that's a philosophical question you can actually ask—like, who are the humans you're talking about?

There are definitely different kinds of values, and how do we prioritize that? Or should we just think of humans? Or like a broader community? Or social? Or should we also extend the similar kind of interest and concerns to include, say, non-human animals or environmental concerns?

So I think there are different layers or different takes that people understand that. But I think the emphasis is more on not thinking of AI as just a design that is isolated, but more on thinking about its potential impact.

Yeah, and then connecting back to your question about what I do there. So, in addition to some of the research projects where I try to address, collaborate, or have this kind of discussion with people, another thing I contribute as a philosopher, as an ethicist, is to serve on their grant committee.

So one of the ways that [The institution] tries to make sure that when engineers or when people propose a new or are just starting to work on some of the new research projects, they want to ensure that people have this kind of ethical lens to consider the potential ethical and social impacts in a very early stage.

There are many grants or seed grants that [The institution] offers, and all the scholars on campus can apply for them. And in addition to some of the typical grant documents, they also need to write, I think, one or two pages of statements that discuss the potential social and political impacts of the project. They need to raise at least some of the potential concerns and then explain how they try to mitigate the potential negatives.

[The institution] recruits people from more interdisciplinary backgrounds of committee members to serve as reviewers for these grants. This committee is called [Name of institution's review board] program or panel. And so, when I was there, I served on the [Name of institution's review board] panel for [a period of time].

Yeah, so I think that's also a very interesting design. I think it's a very interesting institutional design that tries to ensure that this is one of the components people start to think about earlier.

00:07:46 Interviewer

And for the work there, I'm just curious about how you measure—because they submit a two-page document regarding that—but as someone who works on critical theories, how do you measure it? How do you say, this is... this is approved. This is not. What are your criteria, for example?

00:08:11 S1

Yeah, that's a very, very good question.

I think, actually, lots of times, we might disagree. Kind of like... so, for each of the papers, there will usually be a screening by at least two reviewers. And then, the reviewers... it was like independently? So it's kind of like paper review. If both of them say, yeah, this is lower risk, we'll just pass. But if they disagree, then it will go like a triage (??) to a higher level, maybe like a faculty level. So there will come in and then we will have a discussion together. So, it's like a multi-level discussion. But another point in the design of the [THE INSTITUTION 'S REVIEW BOARD] review is that they try to emphasize it's a little bit different from, for example, an IRB. With an IRB, you get the result as a pass or not. [Name of institution's review board] tries to say that, for example, we as reviewers, our role should be more like a coach. So, we are not the final judge because actually it's very difficult to foresee the potential ethical and sociopolitical impacts, right?

But the goal is to bring in more people—especially people like us, who already have some background and some more exposure to the similar kind of concepts—try to point it out and share different perspectives on all the potential ethical concerns that may not have been addressed yet. And then, try to provide more constructive feedback on how they might decide things differently. So, I think the final requirement or result is not just about pass or not pass, but rather, there are some recommendations you should consider. There are some potential concerns you should try to address. And they will need to... Before they receive the funding or the grant, they need to engage in the iterations of discussions. So, I think that's the ultimate goal.

But yeah, as a reviewer, we had some brief training, and we were given documents about a few categories of common types of ethical or sociopolitical concerns. I think that's a helpful first step, and it gives us a framework to consider future ones.

00:10:48 Interviewer

That's quite interesting. I thought it's always a yes or no, but I think this is evolving into trying to improve in a less binary way of doing this.

I think it's quite interesting. Maybe I'll jump to the topic of healthcare. It's a super, super broad question, but what do you see as the main ethical challenges in terms of adopting AI in healthcare systems?

00:11:07 S1

Ah, yeah, that's really very broad, I guess. Like, it depends on what kind of role that...

00:11:14 Interviewer

Analytical tool?

00:11:16 S1

Sorry, say again?

00:11:16 Interviewer

AI as an analytical tool, for example. What I mentioned, using AI in assisting or helping, supporting doctors in their decisions, for example.

00:11:30 S1

Yeah, even like, for example, you mentioned—like medical imaging. Maybe AI suggests or points out some of the problematic areas or tries to categorize.

Are you also thinking of some of the LLM-based models that might try to summarize case studies? Are those instances part of your concern?

00:11:53 Interviewer

That I'm not very sure because, in the end...

00:11:57 S1

More about images, yeah?

00:11:59 Interviewer

In the end, maybe the communication part will also involve LLM. That, I'm not very sure, because right now, it's... this is the engineers' work for now. They're only using it for image recognition, and in the end, to communicate to the doctors. There might be involvement of textual explanations, for example, that will involve large language models, so that I'm not very sure.

00:12:27 S1

Yeah, yeah. I think back to the question, so I think, yeah, mostly thinking of, for example, using AI to help with diagnostics.

And then you say, ethical questions... so yeah, for me, the issue that jumps to me is definitely whether... because I think when we talk about healthcare, the existing situation is that there are long-lasting health disparities among people with different social categories. That's just a very general phenomenon that can be observed in many different places, and I think that is still true. And that matters because healthcare is a very essential part. So, I think what's most dangerous sometimes is when people see how powerful AI is and see the positive sides that AI can be done, and think that once we adopt it, we bring the benefit to everyone, but not realizing there can still be a gap, and not realizing that the gap can be either replicated or exacerbated by the existing disparity.

So, I think that's the kind of... yeah, I'm going to say biases or injustice, is what I think might be the most problematic. And it's added another layer because, right now, AI is sometimes portrayed as a very comprehensive tool. It is trained based on such a great number of data, right? So, it can do the work that can outperform human doctors or practitioners. And that, in a way, might lead people to give more credit to AI—not questioning the results or so. So, when there are similar kinds of biases or problems with the recommendations, or with the AI's suggestions, my worry is that people might not take a more critical perspective as they that might do to a human counter part. That's what I'm probably most concerned about.

00:13:07 Interviewer

In terms of... actually, I'm just disclosing some information about the project. I think actually they only have 50 to 200 scans in their training data. This is a very powerful tool and helps in finding new biomarkers to detect if it's the disease or not. So, I think that's already a very interesting point where the reality is... it's not. It's really not. And I ask about the bias part where I think, for people who are more invested in AI, they know there's limited training data. And for engineers, they say there's data scarcity in healthcare. And they always say if there's more data, it will be more powerful, it will be more accurate. What's your take on that? I'm not very sure that the solution to bias is having more diverse data or not.

00:16:18 S1

I think that can be maybe just part of the... Yeah, I think for different situations, there are definitely potential, like, more distinct parts that need to be addressed.

For example, when you mentioned there are only a few hundred pieces of data, that seems to be one problem. And I would say, yeah, definitely, they will probably need to enlarge the size of the data they include. But I also don't think that just by including more data, that can be solved.

One thing that's very intuitive is, like, what kind of data has been included there, right? Some of those are also constrained by social and economic factors. Researchers need data that is digitized, clean, and curated. Some of the data sources are just skewed, and that's unavoidable. So, I think that is one limitation of just having more data.

In addition to data itself, I think there are also other factors that might contribute to biased AI results. For example, the kind of algorithm that people design. In my paper, I also talked about how, lots of times, when people are training AI systems, they kind of follow existing medical practices as if those are the right answers. That kind of neglect that some of the existing medical practices might contain injustice or biased attitudes. For instance, many disease protocols are based on training populations that tend to be more privileged—white or male subjects. This is still true in much of medical research, where male and white populations are the dominant subjects in many of these experiments. If the training algorithm keeps using that as a reference standard, there will be a tendency to replicate the same biases we already observe. So, I think that's another part of the issue.

And then, with implementation part, there can be also... which parts of the world or which hospitals will be able to implement AI systems? And whether that will expand some kind of digital divide in the medical context.

00:19:39 Interviewer

Very interesting. To be honest, as far as I know, for the engineers, they also need to submit to a board to get approval for the data they are training on and also the data they are testing on. So, there are already, I think, some ethical considerations, at least in the research part.

And I want to ask about, for example, the privacy of the patients whose data is being used for training. Do you have any comments on that?

00:20:16 S1

Yeah. No, I think I will... I will agree with that concern. And I don't think I have enough expertise to provide more helpful insights. I will pass on that.

00:20:31 Interviewer

OK, OK. And...

00:20:32 S1

But I will say I think that is true. And I think I saw some papers that try to propose infrastructure designs—how machine learning can use data but, at the same time, not disclose or intervene in the privacy of the patient.

But I have to admit that I don't know enough of the details, and I don't follow the debate as closely. I don't think, yeah, I'm probably not the best person to answer this question.

00:21:04 Interviewer

And then, in terms of accountability—for example, if the AI system, in the diagnostic process, made a mistake and then the doctor, for example, failed to detect that—who do you think should be held accountable? Or what should be the process?

00:21:23 S1

Yeah. That's a very good question. And I think that's the kind of question that's also very salient. For example, when people used to talk about autonomous vehicle or self-driving cars, right? Like, the AI system is suggesting some things, and then people decide whether to take that or not.

I would tend to think that the responsibility is maybe shared by many different actors. For the medical practitioner part, I think they are definitely still important agents because they are the ones who make the ultimate decision. I think they should definitely be included in the process. By saying they should be held responsible, that probably also suggests that if they are using the system... I don't know the current practice, but I think more training—understanding the system, understanding limitations, or maybe some of the basic infrastructure of the AI systems—should be included in the process. I don't really know whether that's part of the current state.

But, on the other hand, I think there's definitely a part of the responsibility that falls to the different actors who contribute to the design of the algorithms. Just like if there's a vehicle that's always making mechanistic mistakes, we would say the people designing it should bear some responsibility for thinking about how to redesign it. So, I think the same rationale should apply to AI systems used in this context.

00:23:24 Interviewer

And I don't know... just making an analogy with autonomous vehicles, but the testing dummies are always white males. So, I don't know if there could be a similar protocol in healthcare as well, in the testing part. I think you previously mentioned the replication of targeting mostly white males in medical practices, which might also be replicated in development of the algorithms.

I don't know, in your previous project or in the committee you were part of, are there any protocols regarding that?

00:24:10 S1

You mean, by "protocol" you mean, like, trying to reduce that or...?

00:24:16 Interviewer

Or, like, to prevent this kind of thing from happening, in a sense.

00:24:20 S1

Yeah, yeah, yeah. That's a good question.

So, I think there are usually two directions people try to take. One is definitely still thinking about the data sources they are getting, right? But another thing I see in some of the projects I reviewed today is that they try to involve more of the community of people that their product might impact on or might interact with. They start some involvement in the design process, maybe throughout the process or maybe also at the very beginning, to identify what kind of issues really bothers them, really bothers this community of people.

Cuz lots of times, designers might think, “Oh, I have a good tool; I can address this issue.” But that might not necessarily be the issue that the local community encounters the most. So, I think that is something I see people definitely doing—trying to engage with the local community or more diverse groups of either users or people might be impacted like stakeholders during the process.

00:25:57 Interviewer

And I think part of the reason why we’re interviewing you is because we want to involve more stakeholders.

The obstacle that I have right now is that I really want to engage patients, but I think, due to ethical issues, I cannot directly engage them. So that becomes my main concern, because even though our system, if designed, will be used by medical doctors, but the patients will be influenced the most.

That’s what I’m most concerned about. But then also, there are obstacles in implementing research with patients. So, I don’t know if you have... this is just a very practical, very related question that I have. So I don’t know if you have an answer to that, or...

00:26:53 S1

Yeah. So you say that the main obstacle is you cannot directly engage with the patient because of ethical issues?

00:27:02 Interviewer

Yeah.

00:27:02 S1

Maybe, like, interviewing would interfere with their privacy?

00:27:05 Interviewer

Yeah.

00:27:07 S1

I see. Yeah. That’s a very good question. I don’t know exactly, and I don’t know enough because I myself do not conduct this kind of more participatory design or engagement with the local community. So, I don’t know whether there are some more general strategies that people try to do. But, on the other hand, I’m thinking, for example, if not directly—because I guess that’s the kind of ethical challenge you’re facing, right?



You cannot directly ask the doctors to know who their patients are, and so you cannot directly engage with the same patients impacted by the system. But I'm thinking, from a broader design perspective, is it possible... I don't know whether there would be some communities—like patient groups or support groups—that already exist in some social contexts?

Those patients might not necessarily be impacted by this specific AI tool, but maybe there would be easier ways to engage with those groups in general. And people there might also be more willing to share because, if they're in a situation but not currently undergoing some medical procedure, it could be different.

That's one of the ideas that just popped into my mind. I think that's another possible way to try to reach out.

I have an idea that was motivated by... I saw some projects that were conducted at [the Institution] as well. They tried to reach out to, for example, Asian communities or different demographic communities. The goal was, similar to what you are thinking, to improve healthcare quality, in general, for patients. I don't think they were specifically thinking about AI design, but just healthcare.

The situation they were facing was that lots of interview are with the patients who had already been served as subjects of experiments tend to be very skewed populations—who has English as their native language. But in places like [Location] and other states, there are many immigrants who tend not to have that much of the connection already in the local community. Maybe because of linguistic barriers and many things, they lack the channel. They are not recruited in the experiment. So their interests, their perspectives are always not well-represented.

And so, I think there was a time, yeah, when I was at [the institution], that I saw someone else who is a [the institution] researcher. Her lab was doing this kind of project, trying to reach out and recruit subjects who are either Asian, Latinx, or African American. And what she tried to do is she just went to Facebook groups, with a lot of different demographic populations. So she post with this kind of information in a mandarin group. It was a mom group. So that's how I saw the post. And then she was mentioning to me that some of her other colleagues, who maybe Spanish or who know better connections with different demographic communities, just tried to reach out in that way.

So, that's a very long story, but I think that motivates me to think, okay, there are probably some other ways to reach out to the community, who might be similar enough or at least more representative of the patient groups that you're trying to reach out to.

00:28:57 Interviewer

Yeah, but I think it's super interesting. I didn't think of Facebook groups and certain communities where there's a lack of channels to be included in studies.

So, I think that's super interesting for me. And maybe I want to ask about... Since the project right now is that the engineers have developed an algorithm regarding explainability. I don't know—have you encountered explainable AI? And I wonder, what does that mean for you? Do you know it? Do you think it will contribute to building a more ethical AI, for example?

00:32:24 S1

Yeah, I've definitely heard about that.

I think... I also don't know enough about what the current status is. My understanding—I might be wrong—is that they try to make sure AI can explain, like, try to point it out like where it is giving some of the recommendations or categorizations, then try to point it out based on what kind of factors, or based on what kind of data. Or, like, where the data sources for this decision to come up. But I don't know exactly what the current design status is.

I think it can be helpful to some extent, if the issue is that a doctor reading the AI's recommendation disagrees with it and wants to know where, how AI is basing its recommendation and what the evidence is. I think that could be helpful. But I don't know how well-designed the current explainable AI is designed. That's one question mark—I just don't know.

Another thing is that, if one of the issues we're still concerned about is biases or ethical concerns, I'm not quite sure whether explainable AI by itself would be sufficient. Sometimes I worry that, with explainable AI, people might try to think of the medical practitioners engaging with AI as if it's just another peer. But as I mentioned earlier, sometimes people take AI with more credibility. If the guise of credibility or neutrality still persists, I worry about the potential dynamics. If AI makes an argument or explanation for its reasoning but that explanation is biased and still presented in a strong way, whether it will influence practitioners' decisions and then prime, direct them to go in a more problematic direction? That's a broader thing I'm thinking about.

In terms of a research project, I think explainable AI definitely has its value in research projects. I just don't think it can... I just think it needs to be more careful about how the explanations are designed. Also, having some friction or alerts to the practitioners or the users of the system to understand how to take the explanations that the AI system presents.

00:35:59 Interviewer

Because I'm redesigning, I think potentially a UX system. So, I wonder what you mean by having some potential "frictions" in the system that we're designing?

00:36:10 S1

Yeah, you would know better than me.

But yeah, by “friction,” sometimes I see people trying to suggest having some... I don’t know, maybe that’s not such a beautiful design, but having some alerts just hanging on top of the user interface. I think that’s one.

Or, maybe... sometimes AI just has some default settings, like assumes some of the purpose that it should be doing. For example, I’m just thinking of an example... If you use Grammarly, there are default settings about what kind of English you’re using, and it suggests whether you’re writing in a more formal way or so. And that might be the default. You can just type in, and the user might think, “OK, that’s the suggestion I should follow.” Right now, it definitely gives users some options, but it’s a little bit hidden. You can choose to make your sentence more casual, or in a different kind of English writing, but that didn’t show up at the very beginning.

I’m thinking of this kind of thing—how it’s designed as a default. Whether that already encodes some values and whether those values should be made more salient and presented to the user when they begin to engage with the AI.

So, maybe in a medical context, it’s not necessary that AI always recommends something. Instead, maybe the practitioner needs to be more specific about what they’re looking for or what kind of specific task they want AI to do. And maybe by letting the users choose some options before the system does the work, that will prime (?) them to understand that there are different priorities and values embedded in the system.

Yeah, no, that’s just my two random... you will definitely have more ideas about what the design will look better.

00:38:44 Interviewer

And then you mentioned the potential dynamics between the AI system and the doctor. Right now, the engineering solution is to design trust among doctors in the AI system. I wonder, what does trust in AI systems in healthcare mean to you, for example? And also, do you think trust will be helpful?

00:39:14 S1

Yeah, I honestly don’t know.

At first, when, in the very beginning of our conversation, you mentioned that doctors’ lack of trust in the AI system is a barrier... I honestly don’t know whether that’s a bad thing or not. But I also understand that this can be tricky. At least, I think a bit of uncertainty or a more critical perspective on AI would be better than just trusting AI without any doubt. At least, I think having some caution is good.

Let me see what do I think I have something to add...

But I can also see why a total lack of trust—meaning that people completely reject a tool that might have some value to them or don’t engage with it at all—that they could have some miss-out. But, yeah, I think trust is tricky. I think it’s definitely normal,

and it's definitely healthy for people to decide what to trust and to what degree they want to trust something.

So, I don't know if you've found more details about this issue, but I'd be curious to see what factors are, for example what are the perception medical practitioners have of the AI systems presented to them. So, when they say they do not trust AI, what kinds of scenarios are they thinking of? What are the concerns they have? Is it more about the accuracy? Is it more about biases or something else? I would also like to know if they think whether they have interests in using AI tool. And if so, what would be helpful steps to increase the trust they might have for AI systems?

Those are just some random thoughts I have at this moment.

00:41:20 Interviewer

I think it's a very good point because, right now, I'm looking at the project interviews, and I've found that all our related partners working on this are radiologists who work in research settings where they mostly use AI.

So, already, we're working with biased interviews. Although we interviewed, I think, seven medical doctors, most of them are saying very positive things. For that, I also don't have an answer because I think the information I've got is very biased since they are all very prone to AI. Most of their concerns are that they think AI is helpful and that they are willing to adopt it, but they don't want to verify the AI. They don't want to spend more time reading the AI reports. They don't think it's distrust—they just don't find it helpful. That's why they don't use it. So, I think that's part of the answer to your question. And I also want to explore and understand more about what part they don't like. But about this dynamic between AI systems and doctors, I want to speculate a scenario. What do you think is the most healthy or ideal dynamic between these two?

00:44:06 S1

You mean AI and the doctors? Oh, yeah... very hard to say.

So, I think maybe to backtrack a little bit, it's worth asking: in what scenarios do people think bringing AI would be helpful? For example, sometimes people say AI can help assist doctors with some recommendations. But I think, currently, when people are worried about, for example, the accuracy of AI, I don't know how much... but it might still save some time.

Sometimes I would have the question mark of whether it's better to have AI save time or whether it could lead to doctors not doing the work they should be doing by doing some of the research.

I'm thinking them more from the researcher using AI. Right now, I know that many people find GPT helpful for summarizing literature. But for me, I'd be very cautious about using AI for that purpose because, as we know, GPT can provide problematic or inaccurate information. Yeah, it can help me save some time. But you need to be really certain about

the accuracy of the information that you are getting. Then you still need to go through the papers yourself, I don't think AI saves that much time. Rather, it might give you a shortcut: yeah maybe I can get over with that, maybe the accuracy. But I don't think it's a good practice.

That's one of the concerns I have about using AI in some other settings. But I also don't want to deny that many people do report that, for example the radiologists, I think I've heard about things that people find it helpful. So, I think, for myself, that's an issue—I don't quite know where to draw the line.

On the other hand, I think in some other contexts, people suggest that AI might be useful for countries where there's a lack of medical resources or medical practitioners, and suggesting that AI could help doctors speed up or the AI can do a good enough job that they can provide some recommendations for patients, which might be better than the patients not being seen by a doctor at all. So in that case, that's why I asked the questions of what kind of purpose that we want AI try to do? I think there's always a comparison to make. Every tool comes with risks and benefits. I think that's why people are excited about AI because they see the benefits outweighing the risks. But at the same time don't forget that there are still risks that need to be addressed. I think it's a matter of time for them might be able to try to reduce some of the negative impacts if the benefits seem to be very attractive.

00:48:32 Interviewer

I think I'll go a little bit off track. I just want to... I think I remember there's one project that Google Health did in India.

00:48:42 S1

The retina scans? Yeah.

00:48:45 Interviewer

Yes. And I don't think it's a bad thing that they are trying to reduce doctors' pressure or burden and also have more people gain access to this.

But I'm still very concerned about it because it's still concentrated in one powerful company, Google, which has the data processing capabilities and the ability to use the AI. So, I'm still not very sold on the idea that AI could potentially benefit more people in healthcare.

00:49:24 S1

No, I share your concern. I think the power concentration is definitely something to be worried about. With the current infrastructure of deep learning and the larger and larger datasets, the tendency will only be toward more power concentration. That's definitely real.

I also think the power concentration you mentioned could reinforce a broader narrative, like some countries or corporations or some of the populations are the saviors, then other

populations benefiting from whatever like great wars that other people bring into our country.

I would agree with that. And I think that's a real concern, especially when these initiatives are used as a way to build good reputations for big tech companies. I think that's true, yeah.

00:50:36 Interviewer

So, I don't know. It's a huge structurally unjust system, and I don't know. Sometimes I get a little bit lost. What is my role? How do I help with the system? And I feel like there are too many bad things already in the system. I don't think I can change anything, and I'm just putting myself into this...

00:50:57 S1

Yeah.

00:51:02 Interviewer

...this pond of dirty water that I'm not helping. I'm just making myself go through this kind of helplessness.

00:51:06 S1

Yeah. Well, I think no matter where we are, the current social structure is just full of injustices. Whichever role you choose, you will be in it.

I share what you're saying. I share your feeling, but I also think you're already showing that there is still something that can be done—just by engaging with different people and trying to consider different kinds of values in the process.

The overall social structure definitely cannot be changed by a single action, but there are communities working toward this. For example, you know there are many people talking about participatory design or other design approaches for AI. Compared with 10 years ago, there are now more exciting initiatives and more comprehensive perspectives being included.

So, yes, I share your concern, but I would say there still seems to be something that can be done. And that's the structure we are in—this social structure—will inevitably constrain us and present us with limited options. But at the same time, we can try to keep our agency intact or exercise our agency against these constraints. Hopefully, we can reshape those constraints and improve resource distribution for different groups of people. That's the goal.

00:53:08 Interviewer

Yeah. And I think I have some questions that you already answered, but I think I'll still jump back to participatory design. I wonder, in your work, do you engage people with different backgrounds? And how does it work? What is the process?

00:53:26 S1

Yeah. So, my work... in most of the research projects that I do...

I mean, I try to collaborate with others. I'm currently engaged with another philosopher who is more in another subarea. One of the papers that kind of drew me into AI, as I mentioned earlier, is a collaborative project with another AI researcher. That person has more of a technical background.

That's the only... Oh actually, right now, I'm involved in some other research projects with computer science people. So, I think I'm also just getting started in terms of exploring different angles and connections. You know, those kinds of projects can go in different ways. Sometimes I'm driving the project, and sometimes it's them.

But I would say, in acade[Colleague 5] —and I think that's true in many different settings as well— that lots of our agency is constrained by system design. For example, as an assistant professor, I am bound by—the thinking of what kind of work I need to prioritize—the evaluation criteria that people will judge me by. In terms of delivery, it's the paper. I need to think about where to publish—this journal or that one—and how people will evaluate it.

So, I'd say lots of very practical constraints come into play when we're doing collaborative projects. In terms of delivery, a lot of the time we need to align with what grants we want to apply for, which area or discipline the work fits into. But in addition to that, I do think engaging with people from many different disciplines is very helpful—not just for me, but for everyone—to understand the much broader aspects that should be taken into account. Some of these connections don't necessarily lead to collaboration in terms of producing something, but we do have many working groups or communities where people constantly share. At [an institution], we have this humanistic AI working group that brings in people from all kinds of disciplines who are working on AI. More of them are from the humanities and social sciences, but there are some from computer science as well. I think this group serves a similar function to [the institution] at [the university]. It ensures that people at least know each other and have a basic understanding of what's happening in other disciplines, and what are the resources we have to envision the potential future of AI.

00:56:41 Interviewer

I always have concerns about participatory design, where, of course, there are power dynamics. And there's one thing: people don't always get enough credit for participating—for example, the communities...

00:57:08 S1

Exactly.

00:57:09 Interviewer

...and I think, in science research, we engage people from different backgrounds, but I take what they say and embed it in my own values. I think I only value the people whose opinions align with my values, right? I wonder about this kind of balance. I want to give

people credit, but I also want to be sure about my positionality in this project—that I have my own stake in it. So I don't know. Do you have any suggestions? This is just to wrap up, but I think it's also very personally interesting to me.

00:57:55 S1

Yeah, yeah.

No, I think that's a real concern. I think I actually read a paper talking about a similar concern that with participatory design, if it's not done well, there's the risk of it becoming either exploitative of participants or using them, a way of tokenizing them—they are not getting any of the benefits.

I think I've read papers discussing these issues, but I haven't seen many concrete proposals for solutions. There are definitely some, but I don't know enough about them.

I was recently chatting with a friend who works at a university. Her job is to bridge researchers and the local community. She's at [an institution], working with the local community in [a city]. What she observed is that the tendency that researchers often need to publish, get grants, and produce results. So, they engage with the community to do their research.

But for the local community, they spend a lot of time being interviewed and participating in projects. Often, researchers ask similar questions for their own research project, but at the end of the day, the local community doesn't necessarily benefit from all the research done on them.

[an institution] has some kind of institutional center that tries to address these power imbalances and ensure they don't continue as easily as they might otherwise.

I think they also try to shift directions. A lot of the time, researchers go to the community, but not the other way around.

Maybe there should be more engagement with the local community from the beginning to understand what really matters to them. Then, that direction could guide the research project.

Of course, researchers are always constrained by funding and other pressures. But I think raising awareness and trying to foster more grassroots, bottom-up approaches could be helpful.

That said, it's tricky.

## **S2**

00:00:05 Interviewer

I as a designer wanted to know like why did you transit with the background of design and then now why are you designing with XA?

00:00:23 S2



OK, wow. Nice, interesting question. So yes, well, I started—my background in design dates back to more than 10 years ago when I started my studies in product design, actually. Then I moved to communication design, where I explored the topic of data visualization and information design, which is basically my core expertise so far. While experimenting with data visualization and information design, I was fascinated by the role that information design could have in explaining complex things that can't be seen. This was basically the starting point for my research, which is also related to XAI. Information design, as Edward Tufte mentions, is a way of making verbs and actions visible. So maybe we can do the same when dealing with AI. This is the reason why I decided to jump on board this big research project about XAI. This happened back in 2017 or 2018, so OK, many years ago from now, when everything was starting. When I began, I remember that the experience of approaching XAI from a communication design perspective was quite rare. It was very hard to find experiments or even a theoretical framework or background that could help me understand the role of communication design and information design within the XAI realm. So, I don't know if this answers your question or if I should go on. This is basically the reason why I approached XAI.

00:02:18 Interviewer

It's super interesting. And I think back in 2017, it would have been a very niche area, right? I'm quite curious—how did [XAI database] start? Maybe you wanted to build something that could be shared, I don't know ...

00:02:35 S2

Yeah, sure. Great question. Yes, in 2017, I was compiling my master thesis, and my master thesis was about a system able to, which was using AI algorithms for producing some predictive modeling and visualizing these predictive modeling.

Yeah, and my aim with also this master thesis was to try to explain how these algorithms were working. Of course, I failed, but I tried to. Yeah, it was still very complex, but I tried, you know, to do a little bit more research during my PhD.

And actually, when I started my PhD—and then I'll come to the [XAI database]—but when I started my PhD, I started to get in touch with the community of [VISUALISATION CONFERENCE], which is this big conference about visualization. It's [VISUALISATION CONFERENCE], a big international conference on visualization, and there are specific tracks related to the use of AI with data visualization and the use of visualization for explaining AI.

So, I started to get in touch with these people, but even if, like, the environment was mainly populated by computer scientists and data engineers. So, I started to build a kind of small

community, and I had the chance also to spend a week with a group of them during a seminar where we tried to... I was the only designer there. That was in [a date].

And there, one of the aims of this seminar was to create a common vocabulary or glossary to understand each other when talking about explainable artificial intelligence. And this was super hard because I was coming from communication design, and I was doing data visualization. But they were also doing data visualization because data visualization is also, like, a challenge that the data scientists are currently tackling in artificial intelligence.

But again, this was not... I mean, I was super happy to meet them, but I was not satisfied. So, I then decided to spend my visiting period during my PhD at [a University] —at one university whose member was present at the seminar. They, being data scientists, have developed these building blocks of explainable. It was a very niche paper that I found within the [a leading HCI conference] of [Date], and they were talking about building blocks of explainability, focusing specifically on how the explanation is provided, if it's like a deductive explanation, if it's an inductive explanation, and I found it really interesting because it could be correlated to the type of media that we use for explaining complex information, complex data. And so, the [XAI database] is actually based on these building blocks that are coming from this research center at [a University] in [a Country], plus my interpretation of the different types of media that can be used for explaining complex data related to artificial intelligence.

One of the most challenging things here was that at that time, until [a date], it was very hard to find the explanation for the lay public. All the explanations were mostly dedicated to the artificial intelligence creators, designers, and engineers. There was research on domain experts, and this I think is related somehow to the work that you are carrying on with the doctors and physicians that are asked to trust in this system.

But again, there was little research on the general public. So, the idea of the primer was basically to put everything together—to put all the projects that I found related to XAI at that time. Now it's not updated, unfortunately, but I'm working on that. I need funds for it. So, the idea was to bring together all these projects and classify them according to the parameters that, together with the [a University], we had defined these parameters are both related to the type of explanation: is it inductive? Is it productive? Is it contrastive? But also, like, what is the media that's used for proposing the explanation? Is it a video? Is it a data visualization? Is it a picture? And also, what is the strategy? Are they using gamification? Are they using metaphors? Are they using storytelling techniques that can help in understanding? And all these projects are also classified according to the type of users they are addressed to. And then I wanted to create a space that could serve both as an exploratory space for these projects but also as a source for inspiration. I used data visualization as the medium for representing it.

So, basically, what you see in the [XAI database] is this big space where all the projects are clustered, and you can explore them according to their similarities because of the parameters that they have in common.

00:08:33 Interviewer

I think it's quite interesting that you mentioned struggling for a week to communicate with the data scientists. Do you think XAI means something different to you as a designer? Or, how do you personally interpret the meaning of XAI in your work?

00:08:54 S2

Well, great question again. I think XAI can actually change its meaning depending on the final user. For instance, when we talk about XAI, we can consider two main aspects. One relates to transparency—basically, making the system transparent so we can dissect it and see all the components that comprise it.

On the other hand, there's what's called post-hoc explanation, which involves explaining the results. Starting from the outcome generated by artificial intelligence, you work backward to provide an explanation or justification for that result.

What I've come to understand—and this is something others have also pointed out—is that transparency is much more important for people who work directly with these systems, like data scientists. Meanwhile, post-hoc explanation has often proven to be more important and effective for domain experts or the general public, who may not be particularly interested in what happens inside the system itself.

This distinction isn't just something I've noticed in the projects I collected through the [XAI database]. It's also evident in experiments we conducted here at the university. For instance, we asked students to create posters explaining various artificial intelligence or machine learning algorithms, ranging from simpler to more complex ones.

00:10:44 Interviewer

For the projects you selected for the [XAI database], beyond grouping them and identifying who they aim to explain to, what qualities stood out to you? What traits made each project particularly effective and therefore you selected them?

00:11:05 S2

To be fair, I have to admit that at the time, there weren't many projects, especially interactive ones or those directly related to design. The projects I selected had to have somehow potential relation to design, particularly as a means for communicating the explanation. For instance, if they involved text, that text was often presented in a visual way. This became the first filter I used for selecting them.

You'll find that many of the projects I chose were interactive interfaces, exhibitions, or even interactive websites that used data visualization. These were sometimes addressed

to machine learning experts but incorporated visuals as a medium, connecting them to design. This connection wasn't only about their appearance but also about how the information was provided. So, this design-oriented approach became the key parameter for selection, though I must admit there weren't many such projects available at the time. For sources, I looked at things like the Data Visualization for Explainability Workshop, which is part of the [Visualisation Conference] conference I mentioned earlier (specifically, the [visualisation and AI] at [Visualisation Conference]). There was also the AI Art Google Groups (which may now be the People+AI Research group) that worked on small experiments like the doodles project, which was quite famous at the time. Other sources included conferences like CHI (the Computer-Human Interaction Conference), which proposed several explainers. In fact, the [XAI database] itself was presented at [a leading HCI conference], which seemed like an appropriate venue given the audience's interest in such topics.

So, these were the main resources I relied on, but I have to admit it wasn't easy to find these projects back then. There was definitely a lot of searching involved.

00:13:26 Interviewer

Yeah, I guess it wasn't a lot at the time, and even now, it's not a lot. But I think over the years, with more experience, you've had the opportunity to see many more XAI projects. Do you see any common criteria, or perhaps traits, that you would identify one as a good XAI project?

00:13:46 Interviewer

Say yeah, because right now I guess there may be a bit more.

00:13:49 S2

Well, what has changed... I mean, the big change today is the advent of large language models (LLMs). When I did this research, large language models weren't there, and this has actually brought about a paradigm shift. For instance, most of today's explanation systems are supported by LLMs.

These models, for example, can produce verbal explanations—real-time verbal explanations of artificial intelligence results and mechanisms. This is beneficial both for experts and domain experts. Specifically, the use of LLMs for verbal explanations is particularly important for non-experts because, in some cases, even visualizations were too complex for domain experts to fully understand.

This concept is well described in a paper titled *Beyond Expertise and Roles: a Framework to Characterize the Stakeholders of Interpretable Machine Learning and Their Needs*, which explores AI explainability. I'm not sure if you've read it, but the paper provides a classification of potential users working in the field of XAI. It doesn't focus solely on their expertise but also considers their roles within a project.

For instance, the knowledge someone has within a project can shift, and their expertise can change as well. Take this example: I might be an expert in machine learning, but I could also have a different role in the project, such as being a user testing the system. I remember this paper was particularly interesting for me

00:15:50 Interviewer

But the large language models you mentioned seem to be mainly for textual explanations. What about visualizations? As a designer, I naturally lean toward that graphs are always better than words. I'm curious—what's your perspective on this? And how do you define your criteria for deciding between text and visualization in terms of explanation?

00:16:13 S2

Well, concerning visualization, I can say that there are currently developments, especially this year. For instance, the IEEE conference, which had its last edition less than a month ago, presented many tools that generate visualizations using generative AI.

To my knowledge—and I must admit I haven't gone through all the papers from the conference so far—it still seems complex to create real-time visualizations with generative AI that explain artificial intelligence as effectively as text currently does. Perhaps there is still work to be done in this area, or there might be tools in development that I'm not yet aware of.

As a visual designer, of course, visualization—whether through diagrams, pictures, or other formats—is highly relevant because it aligns with my discipline. However, I also believe that in some cases, the way you present text in an interface is also part of the design work. For instance, designing how the text is formatted, where it is placed, and with what kind of rhythm to read—these are all integral to the design process.

What has also emerged from the [XAI database] is that projects combining different media often perform better. This isn't a groundbreaking insight, but it has consistently shown to be effective. Combining media is a solid strategy for creating explanations, for different kinds of users, especially for users who are not experts in AI.

00:18:33 Interviewer

You mentioned domain experts and the public. How do you see the differences in communicating XAI to domain experts versus the public? For instance, in our project, we deal with radiologists who are the end users. They are domain experts in their field, but they have no knowledge of AI, which is similar to the public. Have you faced any challenges or what are the obstacles to communicate such a complex system to the public?

00:18:36 Interviewer

Well, when I say domain experts, of course, they do not have expertise in artificial intelligence, but they are experts in their own field. For instance, radiologists know a lot about medicine, perhaps specific illnesses of bones or other areas of specialization. They bring that expertise to their work, but their knowledge of AI may be minimal.

Sometimes, for that reason, it's important to rely on post-hoc explanations. With post-hoc explanations, you can explain symptoms, how they were identified, or why a particular treatment is recommended. For example, one of the first explainable AI experiments—dating back to the 1980s—was a conversational system, a sort of vintage chatbot. It helped doctors provide correct treatments for patients through a text-based interface, the old black-and-green computer style.

This system didn't explain how the AI (a rule-based system in this case) worked. Instead, it provided reasons for a treatment, such as, "This is the best option because we've observed these characteristics: X, Y, and Z." It focused on outcomes, which can be more helpful for domain experts.

Another example I'd like to share — that maybe of your interest — is related to medicine. It's a German application called ADA Health (<https://ada.com/>). I had the pleasure of meeting the developers. ADA is a self-diagnosis app with two faces: one side is for patients, and the other is for doctors. There are some explanations of potential diagnoses that are assigned to patients. Patients are asked to compile questionnaires, and the app provides some responses and reminds users to always consult a doctor before making any decisions. For patients, explanations are often rooted in quantitative statistics. For example, the app might say, "X% of people from your sample with similar symptoms were diagnosed with this condition." This kind of quantitative analysis.

Coming back to the main question. Domain experts know a lot about the topic.

Sometimes, the challenge is to convince them that the artificial intelligence can help them. You need to strike a balance or find trade-offs between what the AI is saying and what they (the domain experts) are thinking, and how they can trust AI, of course.

I've never worked in the medical field myself, so I can't say, "This is the perfect solution." However, showing the process by which AI arrives at a diagnosis—perhaps using comparative explanations, comparing the results from other results—can be valuable. Comparing images, comparing cases from the literature or from other cases on the same application on the same system is a path that can help. "Look at this picture. This was ok. For X reason this picture here means...", providing context for the diagnosis.

00:24:34 Interviewer

If I understand correctly, you think the trade-off with the domain expert is not only communicating what the AI is saying but also trying to present their thought process—what they are thinking as well?

00:24:49 S2

Yes, it's a trade-off also for understanding. The medical doctor needs to trust not only the result but also the process. The process that has been followed needs to be made explicit and visible. This is because the doctor needs to understand whether they would have made the same decision. It's not just about the result itself. A doctor can trust the result if it aligns with their own thinking. However, if it doesn't agree with what they are considering, then they need to go back and question it. They will ask: "Why did it suggest this? Why, why, why?" It becomes a chain of "why" questions.

00:25:52 Interviewer

It's also interesting that you mentioned that we have to make the domain expert trust the AI system. What do you think makes an AI system trustworthy for domain experts?

00:26:06 S2

For sure, I think there is a big element that relies on trust in the process—the reflection process, the thoughts, and the logic behind it. Then, maybe the doctors also have to trust the data, the data used for training the AI. The doctor has to be sure that the data that is used for training the AI must be a good representation of reality, not biased, and include outliers in the correct proportion. It should also be related to a specific set of population that is relevant for their studies. So yes, trust in the process and trust in the data are crucial. Of course, they should trust the model, but that's something that we have less control over. If you trust the process and the data, you'll probably trust the model as well because it's in between data, process, and results.

00:27:53 Interviewer

What you mean by process is post-hoc explanation, is not the dissection of the model, right?

00:28:03 S2

Yes. How the information is elaborated, how the answer is given. I'm not interested in understanding which are the elements that compose the model—if there is an algorithm that it's called, if there is a neural network, or if there is an LLP analysis of the data.

I wonder because you did a workshop that engaged more than just the end users but also relevant stakeholders. How do you see designers' roles in this large system, given the involvement of regulatory bodies and various participants in the process? Additionally, how do you think we can foster or support this kind of participation, and is it helping at all?

00:29:04 Interviewer

Great, thank you. Good question. Well, I think that in general, like in all interdisciplinary and multidisciplinary research endeavors, the role of the designer is usually that of a

translator. We need to be able to translate complex information into something that is more legible and readable by others, for whom changes every time.

For instance, during the workshop seminar, I remember that I was the youngest and the only designer, and they gave me the honor of compiling the final report of the seminar. I asked everyone to define XAI and then tried to bridge everything together. I think this is a perfect role for designers—to translate and bridge interpretations of complex systems. In this case, at least for me as a communication designer—maybe I'm a bit biased—I see the designer as a relevant figure in translation.

Another role designers can play is bridging communities while also ideating more creatively than data scientists. This is why, for example, the [XAI database] is built on a generous (?) interface that allows exploration in a serendipitous way. This kind of interaction can promote ideation that doesn't rely solely on gap analysis or identifying pros and cons from case studies in the literature.

The [XAI database] is made specifically for designers—not for computer scientists, of course they are welcome to look at it. The main aim is to support designers entering the XAI field, helping them browse projects and find strategies and tools.

00:31:42 Interviewer

I think that takes us back to the very first question—why did you dive into XAI? And what is your main goal, for example, if you were to shift your career slightly? I'm curious about how you see the future of XAI. You seem to see potentials in it, and I'm also wondering about designers' contributions to it. How do you think about that?

00:32:12 S2

Well, interesting question. This is a bit of my story but probably is needed very shortly. After my PhD, I moved to the [another country] [another university] and worked in an interdisciplinary center that focused on explainable AI. They had strong roots in social sciences and science and technology studies (STS). I started to understand—not the system itself—but the role of everything else around artificial intelligence. When we explain a complex system, we don't rely solely on explaining the system itself. The system connects to many other things we deal with daily. For example, if I'm explaining symptoms about a specific issue to a doctor, I'd probably need to explain where this issue spreads, who else is working on it, and the tools used to analyze the data—even without AI. It's not just the AI system that requires explanation but also the socio-technical system around it.

This is especially important when dealing with people who aren't experts in machine learning—especially the general public. So you need to contextualize the (AI) system, otherwise, it's something without a background and it seems untouchable. Accepting an explanation that is minimal (without sufficient context)—I don't know—can be scary. Because it is very hard for people to recognize themselves without a background,



without an environment, without something that is familiar. So even the familiarity of the background of the explanation, either it's visual or verbal... The contextualization and situatedness of the explanation is extremely important.

This idea is well-researched in STS, where the situatedness of technology plays a very relevant role in understanding these systems. In my case, (I'm still working on this, and I'm not entirely satisfied yet) design was to create a bridge among disciplines—connecting STS, computer science, and design, for XAI. A bridge made of seminars, bringing people working on the same topic together, inviting people to conferences that were not about their topics. It's a lot of confusion.

00:35:58 Interviewer

I was about to ask what are you, what are your method or strategies to be the translator? And then you answered already. But if you have strategies you can also share.

00:36:09 S2

Well, no. I think what I've learned over the years is the importance of bringing people together—placing them in environments that are not necessarily their comfort zone. At the beginning, this could be done physically, but it doesn't have to be. It could even be an online meeting. For researchers, it's often valuable to spend a few hours outside their usual comfort zone.

The idea is to gently push them to think in other terms and stimulate translation or different interpretations of the same issue or term. It's about opening up possibilities for interpretation. I'm not saying I want to push people to do what they don't want to do, but rather create an environment that fosters a little bit of uncertainty.

00:37:30 Interviewer

But how do you actually do it, like in a more specific context? How do you put them out of their comfort zone?

00:37:40 S2

You can conduct focus groups where you involve participants and prompt them with questions like, "What is artificial intelligence, in your opinion? Can we describe artificial intelligence as a sequence of mathematical procedures or not?" Some might say yes, maybe 5 years ago; others might say no, AI is a complex system encompassing various elements such as materials, people, labor, and much more. As a designer, your role is to mediate these discussions. This is a strategy — not a methodology, I would say — to create a little entropy into the system.

This isn't exactly a formal methodology but rather a strategy to introduce a bit of entropy into the system, encouraging diverse interpretations and expanding the conversation.

00:38:43 Interviewer

And do you think this kind of participative method could help with building trust, not just with the end domain users, but also among different stakeholders? Do you think it's helpful?

00:38:58 S2

Yes, why not? Yes, they can be useful. I think it's very important to select people and understand the pool of experts and non-experts you want to include. Creating mixed groups as part of a co-design experience—where you have the final user, designers, and those developing the technology—could be a very good strategy.

00:39:40 Interviewer

I've seen a lot of examples, but I think I also wanted to ask you—how do you translate? I've always wanted to see how you translate such a co-design process into something that is actionable or something you can take to the next design stage.

00:40:03 S2

Do you mean "translate" in terms of how to report or how can I translate this into outcomes?

00:40:14 Interviewer

Sort of an outcome that in the end it will help you design, but I think for us as the designer to be part of the conversation is already helping, but how do you express it in a way that highlights how it's contributing?

00:40:30 S2

OK, for instance, in another project, we created some starting points—maybe diagrams that reflected a potential process the artificial intelligence was using to produce a result. We then asked different kinds of stakeholders to comment on these diagrams. For example: "Where do you see your role being most prominent? Here, here, and here? Why do you think your role is relevant in this part? Which actions do you take to facilitate the work of someone else further along in the process?"

This participatory process helps in a way because it uses a common ground, a shared space, where people can interact and give their own interpretations of it. Since I have a background in visual design, I easier for me to think of this as a big map that people can annotate and discuss. This approach can work for both experts and non-experts, perhaps with different maps tailored to their perspectives, but it can be effective in both cases.

00:42:17 Interviewer

Thank you. And I want to go back to the previous question where you mentioned that when designing, you have to contextualize the system but not oversimplifying it — not a minimalistic one. You can't overload your audience. How do you manage such a balance?

00:42:43 S2

I don't know if this is the answer but it depends. You need to test and validate your design to achieve the correct balance. Of course, you cannot oversimplify, I agree. Oversimplifying is not necessarily using a minimalistic background—you can even oversimplify with a lot of elements. You need to verify which elements are actually needed for that specific user. This requires iterative testing and evaluating what is effective and what isn't. I don't have a special protocol for this, but testing is essential.

00:43:42 Interviewer

Yeah, and one last question, just to understand your vision and outlook on the future of XAI. As designers, how do you see the future of design in XAI?

00:44:04 S2

The future of design in XAI, I see it really connected to the world of generative AI. Probably, prompter design would be a good direction. For example, how can I use generative AI to help me understand complex AI systems, both in terms of text and visuals? Designers could act as co-creators or co-designers of applications that help achieve this, like starting from GPT extensions to other projects using generative AI. Generative AI also helps in producing real-time explanations, which is another big area. It's super useful for real-time explanations and facilitates the process of creating these kinds of explanations. This is on the one side.

On the other side, I feel it's most relevant for the lay public to use design techniques—even with the power of GenAI—to generate awareness about AI: how much it impacts our lives? how much it consumes? How much it is problematic? and how much it's entangled in our daily lives? This don't always need to be real-time. These efforts could include exhibitions or speculative design artifacts that encourage the public to think and reflect on the role of AI today. Such efforts aim to raise awareness about what it means to give data, use data, and interact with AI systems, which is the other side of the power of the artificial intelligence.

[end]

00:46:57 Interviewer

Interesting.

00:46:59 Interviewer

Yeah, and and so nice it was. I learned a lot by just talking with you. And I, there's so many, so many things I need to go through. And you were giving such a yeah.

00:47:15 S2

No, thank you. Was great. I mean, thank you for. I mean I was, it was like. One year that I was not talking about this project anymore. So it's it was nice for me as well like to to to yes, to to to yeah. To talk about it again.

00:47:35 interviewer 2

Now, what are you focusing mostly on?

00:47:38 S2

And now I'm working on a project that uses AI, but for sustainability awareness, like for climate change adaptation, awareness is still about AI. There isn't some extensibility.

00:47:57 S2

So we need to produce an explainable system for citizens.

But the the main topic is towards like a climate change adaptation and climate change awareness. So it's just another application, yeah.

00:48:11 interviewer 2

OK, we have a project in our lab.

Neither me or Wen are working on it, but it's also about like sustainability in water management. And yeah, maybe when when we meet in a car like, could you speak with some of our colleagues that are working on that? No change.

00:48:28 S2

Yes, that would be great actually, yeah.

00:48:31 interviewer 2

The source we have? Yeah. Explanation of data. Water is used. So your communication with the citizens but also. You know politicians, policymakers. Yeah.

00:48:41 S2

Yeah, indeed. Even in our project I'm working now. It's. Yeah, it's it's a big project. And of course, there are many kind of stakeholders, even politicians. Yes. And of course, you need to deal with them when when working with topics like that very.

00:48:57 S2

Very specific for citizens for yes, for the city in general, yeah.

00:49:04 Interviewer

I can tell you're trying to harness the power of AI to address topics that really matter, like sustainability issues or challenges posed by AI itself. I think it's a valuable exercise and a great tool to tackle such concerns.

00:49:26 S2

Yes, I mean, these are the good things. But we need to be aware of other aspects, not just in terms of the data we provide to AI. We need to be aware that AI are helpful, but it can also be used or misused potentially. So yes, while the positive aspects

are evident and we're all trying to do our best. It's people who are working on it, so you don't know the intentions.

And power. Power is another relevant element in the AI dimension—power, politics...

00:50:23 Interviewer

Yeah, I think the data... I'm just chatting, but I feel like a big concern I have about AI is related to power. Specifically, how it seems to widen power gaps. That's one reason I asked about participative design earlier. There are some criticisms about participative design is how to balance the balance the power between participants...

00:50:59 S2

This is another project I worked on which is *shaping AI* where we together use participatory methods with the group of extended domain experts. We had the developers but also activists and artists—people who were working in the field of AI, but not citizens. It was not a civil society so they are not included. It's a poor design.

But even in that case, it was kind of evident that who designed the artificial intelligence was thought to be more powerful, “we are the data scientist producing AI, so we have more power”. And all the other people who recognize that “you're the creator, so you have this power.”

Many of the issues and the controversies that have emerged related to AI are because of power. For example, the use of large language models and the big biases that are within them. The problem is that those models are trained according to data set that are representing the western side of the world. The white men that have the power and so basically they will produced biases according to this information that they are provided. So in this case, of course, even participation can have this kind of problems. What we have done also was to use participatory methods not to find solutions, but for discussing. Because the problem sometimes can be using participatory methods to solve a problem. Probably there will be always one person that is more expert than other that say “I know how it works so I can have the ultimate answer”. We work on controversy mapping and controversy elicitation. If you work on a problem and you ask people to reflect on how the problem can potentially be more problematic, you generate the discussion and avoid the situationist approach, or at least some times. This is what happened, but then there are another options.

So yes, the participatory methods in the way in which they are coinfecting (?), but maybe not solution-oriented.

00:53:45 Interviewer

When taking that interesting conversation... The idea that some participants may hold more power. How do you balance that translating the results?

00:54:04 S2

First of all, we trace the conversation. In this experiment, for instance, we used data physicalization tokens to collect information and track the discourse and how the discourse evolved. We also asked people to take notes and write their own reflections. These tokens made it visible if someone participated more than others. While we couldn't connect the tokens to actual person, but we can tell that some voices are more relevant than others. The way we analyzed this depended on the axis that they are analyzing and the parameters that they are looked at.

We always included a moderator in each group during participatory sessions. One of the moderator's roles was to avoid or mitigate situations like this. But, it still happens, even with a moderator. For example, if someone decides this thing is horrible, the group will not do it. And power appears in that way.

Balancing power isn't simple, though. This isn't my primary research field—I'm not an expert in participatory methods. I use them and study what I need to apply them effectively, but this isn't my main area of expertise. Perhaps others are more knowledgeable in this regard.

### **S3**

00:00:03 Interviewer

So maybe we can start with describing your typical day at the lab. Like what? What do you, what do you work on, and who do you interact with usually?

00:00:14 S3

So normally I'm working on the [NATIONAL MS COHORT] project, and we are correcting the masks that are automatically generated. So we go through, you know, [NATIONAL MS COHORT] is MS Cohort that takes patients from different hospitals in [the country]. So we collect images from [City 1], [City 2], [City 3]. Now it's also [City 4], and, of course, [coordinating centre]. So they all sent the images, I think, in a pace of, this I don't know, less than a spatter regularly. And then we run the algorithm. The masks are generated, and we correct them, so we remove false positives or we add whatever is missing. And I do also try to find out other findings that do not fit to any. So if the patient has had also an infarct, or if he has some meningioma or whatever, or a trauma. Also, they cyst that may interact with analysis that people want to make afterwards. So this is all the information we make, and then when this is corrected, it's the data are sent to the statisticians, the CTU. But this also, it's not me who's doing this. [Colleague1] takes its work. He does. So normally I'm on this part of this [NATIONAL MS COHORT]. I'm working alone. And only when I have gone through the patient, normally it's 400 every three months, around 400 visits. If I have found things that do not fit to MS, so trauma or whatever, I verify this with a neuroradiologist, so

that whatever I write there, it's been verified by someone that has the qualifications. I'm just visually trained, but I'm not, as I say, I'm not a physician.

So these are the [NATIONAL MS COHORT] part, and now I'm working up on a project with [Colleague 2] and [Colleague 3]. They are both engineers. But basically my part is the same; it is to check masks, and we are trying to fit them the best way possible to follow them longitudinally in the most reliable way. So to see whether changes... to find a way to reliably follow the patient's longitudinally.

00:03:13 Interviewer

So if I understand, these hospitals will send you a scan, and then you correct the masks. Does it ever go back to the hospitals, or...

00:03:27 S3

No, not because this is just for research purposes.

00:03:31 Interviewer

OK. OK. OK. And what do you mean by following up the patient longitudinally?

00:03:32 S3

Yeah, because you know the patients are regularly... they have to visit regularly. Normally, it's once a year. This is my understanding, yeah, because I'm not doing that myself, but they are followed yearly. So the patients that agreed to enter into this cohort, this [NATIONAL MS COHORT] cohort, they are followed yearly.

I think in the hospitals also they have this kind of routine. The patients are visited regularly every year. And just to know if they have a crisis, if they have an event, they may go more often. All these scans that are for clinical purposes are collected, and this is what we receive.

In [coordinating centre], because [coordinating centre] decided that it was going to be the center that would coordinate this. So these are the scans that we receive. We have at least one visit per year per patient, but in some cases, we may have even more. Not many more, but sometimes more.

And then, normally, the way we do it is cross-sectional. So the masks are treated as they come. We do not take information comparing what was in the visit before.

00:05:07 Interviewer

Do you have any patient information when you annotate?

00:05:13 S3

At least we don't. But we can ask for it. So all these masks are corrected and made for research purposes. When someone wants to perform whatever project, they can use these masks and the information we have created. Then, if you need clinical information, we can ask for it from the hospital. Normally, [coordinating centre] patients are the ones that most

of the things are made with because these are the ones that people have easy access to. But I personally don't have any information. I just have a number, and I know the center where they come from, but that's all.

00:06:02 Interviewer

Only the numbers. Em.

And if I understand correctly, your role is annotating the MS scans, and then it will be used maybe for training a new AI system or whatever?

00:06:16 S3

This is set for statistics for correlation between lesion load and cognitive deficit, or it can also be used to mask the lesions in the brain in order to perform atrophy investigations. And this is why these other findings are also important. If the patient does have an acute stroke or trauma or whatever, this is information that is valuable also, in order to dismiss the lesion load that is not due to MS but has other different origins.

00:07:08 Interviewer

And you mentioned, like, sometimes it may not be MS, and you have to confirm with the neurologist. But I'm curious, because I thought the hospitals would only send scans of MS patients. Or is it a mix?

00:07:25 S3

So the scans are only of MS patients. There are patients that have been diagnosed with MS. But of course, these patients have MS, but they can also have other things. So these are additional findings to MS. You may have a patient that fell and had a trauma, so part of the lesions that you see on the scan are not due to MS but rather to trauma, or a hemorrhage, or whatever. But they are all MS patients.

00:08:04 Interviewer

OK. And then you have all the information collected for one patient. Does that mean you have the longitudinal follow-up of that same patient?

00:08:17 S3

Yes. And of course, this information—if there is a trauma, hemorrhage, or whatever—the hospital has it. The only thing is, we are not getting all this information. We're just getting the scans. So we do this further; we add this comment so that the people who are going to use this patient or these masks have this information. But of course, we are not telling anything to the hospital. It's just something that is for our use.

00:08:58 Interviewer

Ok. In that case, I want to ask, when you read a scan, how much time do you usually take?



00:09:05 S3

So, because the patient's masks are already created and the segmentation is already available, it depends very much. It can go from 3 minutes to 20, depending on how many corrections we have to make. There are patients that need more corrections because they have a bigger lesion load or because of the quality—you may have more false positives, or there is movement. So it depends. Let's say you can go very fast: few lesions, very clear masks. I don't know. Sometimes if you have to do a lot of corrections, it can take up to 15–20 minutes, maybe.

00:09:53 Interviewer

And if the scan is not so clear, or there's noise, or they moved a little bit, it will take more time?

00:10:03 S3

It takes more time or no time at all because then it's discarded. So if the quality is really very low, I discard it because I think it makes no sense to keep this information. You are not going to be able to compare.

00:10:20 Interviewer

And how many scans do you run through, for example, in a week?

00:10:28 S3

Well, this is also very... for me, this kind of thing is very difficult because... let's say in a week... I'm working [part-time] now, I'm starting [a smaller additional part-time appointment] with [Lab Lead]. I don't know. I cannot tell you.

00:10:45 Interviewer

And maybe how many in a day or in one hour? How many?

00:10:48 S3

In one hour, in very good conditions, I may... I don't know, 20 patients maybe.

00:10:59 Interviewer

That's very fast, like three minutes.

00:11:01 S3

Yeah. As I say, it depends very much on how good they are, of course, because the masks are already... I'm only looking at masks. So, yeah, it depends.

00:11:13 Interviewer

And then the mask is the current AI algorithm you are using in the research center?

00:11:22 S3

Yeah, these are... I can't. This is [Colleague1], yeah, who's doing this. I cannot tell you which algorithm, but yeah, the algorithm that's in-house. So they, I think, it was created by

[Lab Lead]. So when I say 20 patients, really, I'm talking like this—I have no idea. Normally, I get my list, and I go down. And it takes me, I don't know, whatever, I take the time that I need. I cannot tell you. As I say, I'm very bad with this. It's always been a problem for me already. [Colleague 5], when we had to say how long do you need?

So, we didn't have any provided masks. For me, it was very difficult. You may have patients that you just open, you go through, there's nothing, so they go very quickly. Or you have patients where you need a lot more time because there are also specific problems. For instance, you have patients where you have lesions that don't appear in FLAIR.

You don't see them in FLAIR anymore because their hyperintensity is not that hyper anymore, but you can see very well the black hole in T1. So you know that there is a lesion there, but you cannot see it in FLAIR. For me, it's important to note this lesion. So I'm segmenting manually, and I normally do it in a different color so that we can distinguish lesions that are seen in FLAIR or lesions that we know are there but we do not detect in FLAIR.

For such patient, it takes really very, very long. So if you have five of those, then in your morning, you're not going to go very far. This is why I think it's really very difficult. I never try to see how many. I go through the list, and I say, well, I know at some point it will be over, yeah.

00:13:10 Interviewer

And I wonder, for each patient you get, sometimes you get FLAIR, sometimes you don't. And then most of them, you get at least one type of scan?

00:13:18 S3

We need to have FLAIR. We always have FLAIR. We need FLAIR and T1.

00:13:24 Interviewer

OK, so most of them you have to... And how do you compare between the two?

00:13:33 S3

You mean, I have them one next to the other, and they are co-registered. So whenever you have your pointer, you know where you are in T1 also. This is really important because you need both to see the borders of the lesion. Because FLAIR... I think T1 normally has really valuable information also. So we have both, always. We need to have FLAIR and T1 also because, I think, in order for the mask to be generated, you need both. And then sometimes we also have gadolinium, but this is not mandatory. We do not always have gadolinium. Sometimes we do.

00:14:23 Interviewer

And I'm not an expert, so I just want to make sure. FLAIR and T1 are scanned separately?

00:14:32 S3

Yeah.

00:14:33 Interviewer

Oh, so they might have a little bit of mismatching, or is it possible to have something like that?

00:14:40 S3

This again, this is also not my field... but I guess they are co-registered.

00:14:46 Interviewer

So for the algorithm, it also corrects them and matches them.

00:14:51 S3

Yes. Yeah, they have... yeah. And when I have them, they are co-registered. So you are always... you see the same on both. This is here because, when I was in... well, this was always like this, yeah. T1 and T2 are always co-registered.

Because I was working in [a clinical trials imaging centre] for clinical trials, and there we were doing a lot... we worked for clinical trials, so it was normally longitudinal studies. For the longitudinal assessment, we do not have them co-registered. So you have to look for the lesions in different positions, which makes it a little bit more difficult, of course.

00:15:43 Interviewer

So for a patient, a year-ago scan and a scan now—those ones are not co-registered?

00:15:48 S3

Yeah, but the visits are not co-registered.

00:15:52 Interviewer

Ah, that is very interesting.

00:16:01 S3

I think because they want to... When you visualize the patient longitudinally, of course, you don't have the native information.

00:16:09 Interviewer

What do you mean?

00:16:16 S3

So... They don't like very much the co-registration because, of course, there is some information missing or changing. When you move your scan, when you resize, or when you register, you know what I mean?

00:16:31 Interviewer

OK, yeah. But why for FLAIR and T1 is it OK to co-register?

00:16:38 S3

As I say, this is also one of my... but I guess... I have no idea how they make this. I cannot tell you. You'd have to ask [Colleague1] or people that know better than me. I don't... I get the images, and really, it's good when they are like this. I don't know how they do it, whether it's T1 that's co-registered. Because, as you say, I think they are acquired separately in the scan. So, I guess... I don't know if this is

enough. They're done on the same day. If the patient doesn't... I cannot tell you this. I don't know. I don't know how they are. Maybe I should ask.

00:17:16 Interviewer

And you also mentioned, if you register—if you want to co-register—you may lose some information. That's why...

00:17:23 S3

Yes, this is what I've been told. Because I personally don't know either, but I imagine that when you move slightly, you don't have exactly the same... yeah. It's not exactly the same information in intensity and also in size, most likely. Yeah, when you move something, when you try to register your patient, it's not really the same, I guess.

00:17:52 Interviewer

And for the longitudinal ones, to read those scans, like because it's not co-registered, how much time does that take?

00:18:02 S3

But you know the problem is that we don't do longitudinal studies—it's cross-sectional. So this information is treated as if the patient were new. You do the patient that was here a year...

So this is why you're going to have, probably, fluctuations in the volume because the patient is not exactly in the same position. And also, in this particular case, [NATIONAL MS COHORT], the parameters should be followed, but they are not always exactly followed. So you have to take into account all this variability from one scan to the next, from one visit to the next. It's not like in a trial where all the parameters are controlled, and then you have a better research scenario. Here, it may fluctuate quite a lot.

00:19:03 Interviewer

So for a patient who has a second visit or third visit, when the scan is sent over to you, do you compare it with the previous scan, or do you just read it as if it's a new one?

00:19:12 S3

Yes, this is it. We read it as a new one. I do compare when... I mean, we go through all the visits when we are in a particular research project. Then, we take them all. You can follow and see if the patient had new lesions, or if there are lesions that are smaller, or if you see atrophy increasing.

But for [NATIONAL MS COHORT], for the lesion, for the mask evaluation, every patient is treated as if they are there for the first time. Right now, of course, the goal would be to have longitudinal follow-up. I think the masks would be more valuable, but we are not there yet. It takes really a lot of time. It's a huge volume of data, and we are still thinking about how we could do this.

00:20:15 Interviewer

So if I understand correctly, to compare them from the previous scan to the current visit, it would take way more time than treating it as a new patient?

00:20:28 S3

No, for me, it probably would be faster. But as we said, we have to find a way to co-register. And all this takes a lot of space—the big data. So you have to find a way to... yeah, we don't know yet. It's also a huge data bank.

So I think we are not there yet. I think it will be for the future; it will be great because it's a lot of data—all these [NATIONAL MS COHORT] cohorts. I mean, it would be really nice to have them longitudinally available, but right now, it's not easy. We're not there yet. But this is where I would like to have a bank where whoever wants to use these patients can sort them depending on whatever criteria we have written there, and it should be easy to... Now, it's a little bit... it's still not optimal, let's say.

00:21:39 Interviewer

I think you only recently started adopting this AI tool in your work, no?

00:21:49 S3

I do not understand.

00:21:50 Interviewer

So for the masks that are automatically generated by the algorithm, since when did you start adopting this tool?

00:21:55 S3

Yes. I think this is... I cannot tell you. I've been with [Lab Lead] since 2020 so four years. But they were doing this already before; someone was doing this, but not in a very serious way.

But how many years? I don't know, maybe... I cannot tell you, maybe seven, or I really don't know. But for me, I've been here for four years. When we arrived, we started with the masks that were from 2020 or 2019, the very first masks. And now we are trying to fill the gap with all the masks that we have there.

So [NATIONAL MS COHORT] has patients, some of them are from 2016. As I said, I started four years ago, and we were trying to fill the gap now with the masks... all the masks that are available and have not been automatically assessed yet. So it makes a pretty long follow-up. We have patients that have 8/9 years of follow-up.

00:23:21 Interviewer

And I want to ask because I feel like your role is kind of the frontline of the AI system, where you annotate all the scans.

00:23:30 S3

Yes, I'm the very first, yeah.

00:23:31 Interviewer

Yes. So I wonder, how do you start trusting the tool and also correcting them? Or is it because you are correcting them that you don't necessarily need to trust them?

00:23:45 S3

Yeah. Uh...

How to say... I get the masks, and then you see how good or how... of course, it could be better. I mean, it's working pretty well. It could be improved, and this is also something we are thinking about. Maybe now, retraining the machine again with some masks that have been corrected after being generated. Try to re-run them.

But I have to say, it works pretty well. So, I don't know. You mean... of course, I've been working for 13 years now on MS with masks, so I'm pretty well-trained to see lesions. This is my trust in... yeah, my experience in MS and correcting masks. As I said, I've been working for 13 years now on MS, and my very first years were spent segmenting manually. So you are detecting, you have to look for lesions, and you have to discriminate—this lesion is this, false positive is that. We were segmenting manually, so you get a pretty close approach to the lesions, maybe more than if you start already with automatically segmented patients.

00:25:23 Interviewer

What do you mean by that?

00:25:26 S3

For me, we had to have an empty brain—not empty, a clean brain—and we had to look for lesions and segment them. So you have a system that you look at every slice, and you look for lesions, and you have to discriminate: Is this partial volume? Is this a false positive? Is this cortex? Is this following?

Maybe you get trained. When you have all the masks already, when you have a brain full of red spots, some things are easier, of course—it's already done. But maybe it's different to get this training. I don't know. For us, I want to say my trust is in my former knowledge of MS lesions. I don't know if this answers your question.

00:26:30 Interviewer

So when you were doing the segmentation manually, you think you learned more, and you trained your skills?

00:26:36 S3

Yes, I think so.

00:26:39 Interviewer

You developed experience, and you take that experience and then use it to correct the AI-generated masks?

00:26:47 S3

Yes, yes, yes.

00:26:53 Interviewer

Do you think the AI tools make the job easier, but also it has some drawbacks? Or maybe you think it sometimes neglects some lesions that you find harder when there are already red spots?

00:27:11 S3

Of course, I find it really great to have them already generated. I have to say this makes the work a lot easier, of course. The problem is sometimes you get things when you segment manually. Depending on the quality of your image, you don't have the resolution that you would like, not always. There are always regions (??) with things. So if you have to segment manually, when you are not sure, and you don't take it—you leave it.

When you get it already, then you have to make the decision: "Do I leave it?" The machine is seeing something there, seeing hyperintensities that I have problems identifying. So then you have to decide: "Do I trust this that I'm not seeing properly?" But apparently, they are there. Or do I remove it?" This is a decision sometimes where you have to, as you said, trust the machine.

00:28:12 Interviewer

Yeah, that's what I think is super interesting, where, for example, one part of the vision that even you are not so sure, but the machine may mark it red. How do you decide?

00:28:28 S3

So this is it. I have to go and look. Normally, it depends: is it a place where it makes sense that this could be a lesion? And for me, I always check also with the T1. T1 gives you really very good information because I think, yeah, with T1 you can really...

But then about the borders, and this is the main problem with lesions: where does the border stop? The machine is very generous here sometimes. Either it depends—I have to say it depends on the patient, it depends on the overall of the patient. How does the whole thing look? Not only this very one slice but the whole patient. And then you decide to remove it completely or leave it.

I cannot tell you; it depends. But this is the thing. Normally, when you segment manually, there are cases where you say, "OK, I don't put my fingers in there because it's too complicated." And then when you have it already, you have to decide. But in general, I

would say, for me, it's really a very good thing to have them ready, to have a pre-segmentation.

00:29:50 Interviewer

And previously, when you tried to do it manually, how much time does it really take?

00:29:59 S3

A LOT. I tell you, this is really A LOT—a lot of time. And this is why, to have them already generated, it speeds things up really a lot.

Not only that, I say the problem when you segment manually is, as I said, to decide where you put the border of the lesion. If you do two more rows of voxels, your lesion is going to be a lot bigger; two less, smaller. But for you, it may still look good, your segmentation—your manual segmentation.

But you cannot—you know better probably than me—the difference. When you see this gray, depending on the surroundings, you're going to see more or less, depending on the voxels that you have. So, with the automatic segmentation, you have something that will always be the same. You have the borders defined on a threshold or whatever, and this is going to always be the same. If you run your masks two months later, you're going to get the same result.

Manually, you're going to have a lot more things that will change your segmentation, even if not a lot. For some lesions, it may really vary a lot, or from one person to another—someone may see the borders larger than your next colleague, who will see them smaller. For this, I think that the masks that are auto-segmented are really a good tool because you have something that relies on something objective.

00:31:45 Interviewer

Uh, so the border—the machine has a threshold to decide. So it's very objective, like for even across different doctors and across different times, it's all very objective?

00:31:58 S3

Yes, yes.

00:31:59 Interviewer

And it's helpful when measuring the progression or quantifying things?

00:32:04 S3

Yes, yes. I think you have something that is objective and reliable, repetitive. You will always get the same result. For me, this is really... because, as I said, I was working manually for years—years. And you see the differences between people can be really very important. And you cannot say, "This is wrong; this is right," because there's no objective standard. It's only the way you see it. Whereas if you are presented with the segmentation and they just ask you, "Is this a lesion or



not?” then everybody’s going to agree. Of course, there will be again lesions where you may not find a consensus. But for most of the lesions, everybody would agree this is a lesion.

00:33:12 Interviewer

And I want to ask, if there are new tools, how do you compare which one is sort of better or more helpful? Like in terms of maybe it has more false positives—how do you measure? How do you compare two different tools?

00:33:29 S3

Yeah, we’re not there yet. But this will be, again... I don’t know. This is something we will go into as well. I don’t know

It depends. I guess in the end, you have to do statistical measurements. Does this particular tool have more false positives, or was it missing lesions, but the ones it found were better? The other one might find fewer, but it has a lower... I don’t know. This is something that I guess one has to discuss in the group—what’s the best solution to have more accuracy, but maybe less sensitivity, or the other way around? I don’t know.

00:34:21 Interviewer

Personally, like in your work, which one would you prefer?

00:34:25 S3

If I... but I think if I had to correct... I don’t know. I should see what would be better.

I guess if it’s to remove, it’s probably easier than to add new ones. So maybe something that gives you more false positives but captures all the lesions is probably easier to deal with. But I don’t know.

As I said, up to now, I have only one method to check. And normally, when I have to check for masks, I’d rather remove than add. So, for me, it’s better if I have a false positive than if it misses something, because removing something always takes less time than adding it. Also, I think you introduce less variability if I have to add manually than if I just erase.

00:35:29 Interviewer

And do you think having too many—like having a machine that is very, very sensitive, but it will have too many false positives—do you think it will make you trust the algorithm less?

00:35:45 S3

Yes, probably. Yeah, that too. I think you need an equilibrium.

00:35:51 Interviewer

You need a balance?

00:35:53 S3

Yeah, yeah. Because if everything is taken, then in the end, it's not helping you very much. You start to wonder... Yeah.

00:36:03 Interviewer

But if we enter this kind of imagination space, how do you think an AI tool could be perfect and help you the most?

00:36:09 S3

Where it's the most accurate. I don't know what you... yeah, you have... of course, if you start to find too many false positives, as you said, then you start to wonder, "What's the point?" If you cannot trust ...

Because, of course, for us, up to now, this algorithm is to help us with something that we should do manually—it would take us a lot of time. We don't use it for diagnosis or anything like this.

So I guess, for diagnosis, probably dealing with false positives is worse than the other way around. But I'm not a doctor, and there, I cannot really help you. I'm guessing. I have no idea.

For us, where it's a tool for a manual task—in my case, to remove—it's always easier than to add. But it's only a part of the whole.

For us, it's just an instrument, a tool to make our task easier. But it's not very important for the patient, let's say—it does not have a big impact on the health or the medical issues of the patient.

If we had to make a diagnosis out of it, then for sure, it would be a lot more difficult to deal with false positives, I guess. But I don't know, really. There, it's out of my scope. I cannot help.

00:37:56 Interviewer

Thank you. I will take that to the doctors, thank you so much for that. And I wonder, because right now there are only masks appearing on the scan, would you like the AI tool to offer more explanations or more information? Or do you think the current way is working great?

00:38:08 S3

For us, for what we are doing, this is... yeah, this is enough. But as I say, it's a really very, very small aspect of the whole. This is why I think... I don't know how much it's really important for you.

For us, it's really a very tiny thing that has no big impact on the patient or on the decisions you have to make for the future of the patient, the health, or the whole system. We're in a very small part.

00:38:59 Interviewer

But I think, for me, this is super helpful because we never hear the story of how an AI system started to develop, like how it's evaluated (??). And also, you remember [Colleague 6], right? [Colleague 6] always says there's some... for example, she tries to, how do you say, improve even the accuracy of what the annotators did.

00:39:32 S3

Yes.

00:39:37 Interviewer

So I think that's a super interesting sort of loop of information. Now it's more complete, so I think this is really, really helpful for me because I want to see it in a big picture, but I didn't know... I thought it's the medical doctors who did all the annotations, but apparently, it's not.

00:39:50 S3

No, no, no. Normally, yeah. No, no, no.

I mean, I don't know how they work, or they may not even use these masks. For them, I don't know. In our case, it's more for research purposes.

Umm. And...

Yeah, I don't know. Sometimes there are some things... I think for us, it's more for... yeah. In our case, it's for research. And then, of course, if you do a clinical trial, also the masks are used, but also only in volumetrics. So it's the amount of volume—if the lesion load is increasing or decreasing, or if you are getting new lesions. And for this, the longitudinal aspect is really important, yeah.

But normally... So in the clinical trials we had also, there were also radiologists. Normally, they check, and they do this. So for us, we are in research, with [Lab Lead]. It's only research. We do not... so our things stay in the research group. But for the clinical trials, we also had a neuroradiologist to, to say, to find out incidental findings—what I told you. If they have something, there was communication with the hospital. So there, it's really different because it's a different framework.

You find something that maybe... I don't know, what looks like a tumor, or something suspicious that would be very important for the patient. Then there will be a communication with the hospital. But normally, you would expect that they have already seen it. It's not that they didn't see it, but just to make sure you found something in case they missed it.

In my case, I write these things in order to know, when we do research, that there are things we have to take into consideration. But we do not communicate. I mean, we assume that this work has been done by the doctors.

00:42:28 Interviewer

And for the... I want to ask, do you get any feedback from the engineers who work on using the scans that you annotated in their work? And do they have any feedback back to you, or is it just you complete it, and then it's done?

00:42:49 S3

This is what we also tried to implement, and this is very difficult—the communication, you know, is really very difficult sometimes. Everybody does their work, you send it. I do my Excel table with all these things that I find: this one has a cyst, this one has a hemorrhage. And I would like this to be taken into account, but I'm not really sure that it's always done, unfortunately.

This is one of the big problems: to get this communication flow. Normally, it's stuck. So, this is why, as I say, we would like to have this kind of databank where, when you go and collect your patients, you get all the information. Because normally, people don't take the time to go and check, "What did you write there?" They just take the patients, whatever it is in there, without thinking, "Oh, but this one had some problems in migration of the cortex, so it may have also seizures, for instance." So, their cognitive impairment may be due also to these other things. They don't really look at this, and this is still what I would like to get before I leave. So that this information we collect were fine and it's really used.

And then, with the engineers, I'm working now with them in this longitudinal assessment, and I find it very interesting. Because I see, we see the different ways we have to look at the patient, at the problem. And we are slowly bringing them to see the scans in the way I see them. Not as numbers, or in this very rigid, mathematical, physical way, but in a more organic, biological, not-that-clear way—at the borders, the way... So it's very interesting. It's been really long to say, yeah, but this different approach to the lesions, they're not only numbers, but what is behind the numbers that I can see myself.

00:45:26 Interviewer

And how do you think this collaboration, for example with the engineers, having this feedback loop—how do you think it will evolve in a sense?

00:45:36 S3

I hope so. I think it's absolutely necessary. Yes, it has to. Yeah, they have to come together, for sure.

00:45:45 Interviewer

And how do you see it, for example, becoming in 5 years or 15 years?

00:45:53 S3

I don't... I think... I can't tell you. But I hope in five years because, yeah, I think it should... It depends on how much time you have to invest. I think sometimes now, you have to want to get results. If you start to think too much about the problems, the difficulties, or the

challenges, this is going to put your end result later in time, and you may not have time to wait that long. So you cut out, and you go fast to results. And this, for me, is one of the problems sometimes. If you start to think about how some of the manipulations you do are affecting the result, to take the time to...

I'm a biologist, so normally, in this field with human people, you work the other way around, you have the data. For us, you create an experiment, and you go from there. It's a little bit different.

I think... To take time to understand how your different manipulations are changing the end result is, for me, the starting point. And from there, build up. So, yeah, I hope. But I think now, maybe one does not have that much time. Time to check everything—you have to run.

00:47:26 Interviewer

Yeah, I think that's indeed a very hard challenge—to push things faster but also to make things more correct.

00:47:36 S3

Yeah.

## **S4**

00:00:00 Interviewer

You had a background in interaction design and how we are curious, how did you jump into investigation mostly on AI?

00:00:12 S4

Gosh, that's that's um... The short answer to that question would be that when I was still in practice, I found myself increasingly working on products and systems that included mostly machine learning as a technology. I was collaborating with machine learning engineers and data scientists. What struck me back then—I think this moment was around 2016 (mostly in things like recommender engines and that sort of things)—was how challenging it was to collaborate on products that involved the technology for a variety of reasons. That kind of put me on the path of thinking about... OK, clearly, this technology (these techniques) is increasingly popular. Clearly, they demand new things from designers. So that was the initial kind of impetus to dig into that more proactively. In parallel to that, I was also involved in community organizing in the tech sector locally in the [a country] around the same time, around issues of fairness, equity, and that sort of thing, raising awareness among peers in the tech sector—designers and technologists—about their responsibilities, but also the potential that we as a professional group have to

positively influence the products and systems we work on, ensuring they benefit the majority of people rather than a small minority.

So those two streams—the professional work on machine learning-enabled products and the organizing on the other end—kind of merged into this plan to do a PhD at [a university], which is where I'm now at [a university]. The PhD didn't start as a PhD on contestable AI. The contestability aspect was something we latched onto some time after I started. But yeah, that was the motivation to start working on that and to do research for design or constructive design research. In terms of the methods, it was a natural fit obviously for me because I come from practice. I have all these practical skills and expertise, and it was a good opportunity to bring those to bear on scientific questions around AI accountability and that sort of thing.

00:04:05 Interviewer

I think you also trying to make this knowledge spreadable. You also try to teach responsible AI. And can you share a bit more about what you teach or what methods you give students for example?

00:04:17 Interviewer

Yes. Great question. Most recently, I've been working together with a colleague, [colleague 1]. She's a professor here, and her work focuses on feminist generative AI. We developed a new master elective for industrial design engineering students, which is called [*a master's elective on responsible AI*]. It's not a very imaginative title, but it does cover the subject matter quite neatly. In terms of the methods we're teaching, it's more or less a blend of things taken from value-sensitive design and design fiction. More broadly, of course, a whole range of design futuring approaches could be applied, but we chose to focus on design fiction specifically, as it has been developed mostly by the people at Near Future Laboratory. That kind of strand of design futuring, if you will, is the main thing that we're using there.

Before this course, I taught another master elective for several years called [*a master's elective on AI in society*]. In terms of methodology, it was similar but had a different flavor to the course that we are teaching now. The focus was a bit more (as the name implied) on the sociopolitical context within which students do AI projects and helping them orient themselves to the sociopolitical implications of any project they do. It was explicitly in the context of public sector AI. For this course, we used a range of methods, primarily from strategic design. A bit more on the strategic side, a bit more on product strategy side, as you would perhaps apply in industry when doing product strategy there. A bit more strategic in nature. The new course is more focused on the imaginative elements of designing responsible AI.

00:07:20 Interviewer

I think, for example, public sectors, AI system are all very complex. How do you communicate? How do you explain clearly to your students?

00:07:38 S4

Absolutely, the complexity is a big challenge, and there are multiple sources of it. There's, of course, the technological side of complexity. Then there are societal aspects, like understanding the relationship between technology and society in the first place. There's also complexity related to normative theories—how do you do ethics? How do you do ethics in the context of design?

In both courses that I developed, we had a lot of complexity to unpack. It's also a part of the challenge, me, as an instructor, to decide what to focus on and what to leave unpacked (black box) basically, "well, there's this the other thing that's also complex, but we're just going to trade it as a black box".

In the previous course, we focused very much on educating students about the nature of the technologies involved in AI, mainly machine learning, and more specifically, computer vision. We used a mix of hands-on experimentation with tools like Google Teachable Machine. More generally, is to find a way to let the student to play with these technologies in a very direct and hands on manner, so they can get an intuitive feel of how these technologies behave. I believe that intuitive feel can then be used to aid in thinking about the behavior of systems at a larger scale. For example, if you play around with Google Teachable Machine, you train a little computer vision model, and you try to do something. You start experimenting and figuring out where it fails, where it succeeds, and why that might be. That's a very productive way for these types of students, at least, because they have no formal training in computer science. It's a productive way for them to enter into thinking conceptually about what is a machine learning model, what does it mean to train a model? What's the relationship between data, training data, and the model, etc.? So that's what we did there primarily. Then, of course, it was complemented with fairly straightforward materials like readings and lectures.

With the new course, interestingly, we spent way less time talking about the specifics of AI technologies. In part, this is because we now have the luxury of a new broader context that has changed. Basically, the programs have been updated to include more training around AI technologies elsewhere. Whereas before, with the previous elective, there was hardly any AI teaching happening in the faculty when we started that elective. That has dramatically changed between when I arrived here in this faculty in 2018 and now, it has completely flipped. There's a lot of AI courses offered now. Our students now come in with a basic grasp of the technologies involved. They even get training on the bachelor level now in those technologies. So we have the luxury of saying, okay, we don't have to deal with that complexity anymore. We can spend more time on complexities related to, what are values? What are moral values? What is the relationship between technology and society?

What does it mean for people to participate in technological development? More on the politics around it, etc., which before, we weren't really able to cram into the course on AI and society because there was only so much time we could spend on that. The emphasis has shifted a little bit, and I'm actually quite pleased about that.

Then the design fiction brings in this whole idea of caring for the future—understanding responsible AI design as caring for the future. That's why we bring in the design fiction aspects, because it's a way to explore futures as a designer and to take responsibility for the futures that you envision and propose. But then, of course, students also need to handle the complexity of what the future is anyway and how the work they do now relates to future developments in technology and society, etc. There's lots to unpack and a lot of different complexities.

I guess, in general, with any type of teaching and design, the trick is to somehow make it practical—make it something that you can work with as a designer in some way, either conceptually or physically or whatever. But you need to kind of work with it in a hands-on way.

00:14:01 Interviewer

You mentioned a little bit of incorporating moral values and engagement. I see you posting on having the book *Value-Sensitive Design*, do you also incorporate methods or frameworks?

00:14:12 S4

Yes, of course, so value-sensitive design has its own kind of methodological framework. It comes with a whole bunch of baggage, the worldview, if you will. So we teach part of that as well.

In terms of methods, we teach them stakeholder analysis that you typically do in value-sensitive design. They do case study analysis and learn to distinguish between direct and indirect stakeholders and all that stuff. They also learn about how to reason from harms and benefits to values and how to conceptualize values. What does it mean to conceptualize a value? How can you use normative theory to do that? It's a little bit of philosophy of technology that they kind of learn without really being aware of it.

We don't really talk about it in those terms, but in fact, that's what it is.

One of the big challenges for us was to connect that type of work with the futuring aspects. How do you bridge those two? There is some work in the value-sensitive design space around value scenarios and that sort of thing. These are more scenario-like anticipatory activities, but they're not very imaginative or designerly in the way that you would do in a design fiction project. We teach them ways of developing briefs around future visions or future worlds. This starts with articulating different worlds and then developing briefs around those. These future worlds are connected to the stakeholder analysis and value



analysis activities that precede them. Those scenario-type activities are the methods we teach to help transition into the design fiction work properly. Of course, design fiction is not really a well-described method itself or a set of methodologies. It's very fuzzy, and there are many different ways of doing it. There, we lean heavily not so much on a recipe for doing it but more on exposing students to many examples and analyzing those examples together—like what makes them work, what they're made up of, etc.

So yeah, those are some of the methods that we teach them in the course.

00:17:25 Interviewer

Actually, I'm not very familiar with normative theory that you mentioned. What does that mean?

00:17:31 S4

I mean, just generally, like ethics or the legal side of it. Legal philosophy, political philosophy, moral philosophy—those for me are all flavors of normative theory. We also expose students to a variety of responsible AI frameworks. At this point, there are way too many frameworks and principles out there, and they have different standings. Some come from policymakers, some from industry, and some from academia.

We've collected a few examples. For example, there's the assessment list that has informed the AI Act. We also expose them to frameworks from industry, like the *Microsoft framework for responsible AI*, and frameworks from academia, such as the work Virginia Dignum (*Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*) has done on this topic. These different types of frameworks typically have in common that they enumerate a number of principles that need to be somehow adhered to. We teach students how to interpret these principles and apply them, either for assessment purposes or in more generative ways—like when they're doing the design work itself. These are concrete examples of normative frameworks that come from various fields.

00:19:34 Interviewer

Since we are, we are working on Explainable AI. There's so many different terminologies and you have expertise Responsible and Contestable AI. I wonder, have you ever encountered the word Explainable AI. How do you define it and what does... ?

00:19:49 S4

Yeah, of course. I mean, I don't have a definition ready to go.

The contestability stuff, for me, emerged from initial work on Explainable AI or Transparent AI. Those are kind of closely related terms for me. In the context of contestability, explanations are one of the means for enabling contestation. You need to understand something about the AI system you are confronted with to be able to appeal a decision.

You need to be able to make sense of why a system does the thing it does. That's the explainable aspect of it.

In fact, one of the very first publications on contestable AI was in the context of a healthcare training tool for psychoanalysis. The authors identified contestability as a desirable quality, (this was, I think, back in 2016 or so) building on explainable and transparent AI. (unable to identify the literature???)

For me, explainability is a necessary prerequisite for being able to contest the system. Of course, that implies a particular type of explanation, which also gives you an idea about what an explanation should include. It needs to include the things required to contest the system, which is helpful because otherwise, it risks becoming an end in itself. Then the question becomes, "When have you exhaustively explained the system?" The contestability perspective is very helpful in focusing on what we need to explain about the system in the first place.

00:22:32 Interviewer

What do you think contestability focus on, which aspect of explainability?

00:22:42 S4

So two sides, I mean, OK, well, it depends a bit on what the perspective is we're taking. I always split it out into a local and a global perspective. So the local perspective is one that someone who is subjected to an AI system decision output is impacted by. That's a local contestability. And then the global is more the societal perspective. OK, we have an AI system doing things in society that has consequences for society at large, and society needs to somehow be able to scrutinize and monitor or inspect the system generally. For the local level, which is kind of what I was talking about up to this point,) the explanations are really about giving a rationale for an individual decision, a behavioral explanation (if you will) of why a system has come to a particular decision, which might include, if it's done on the basis of particular types of particular inputs, data inputs, then of course you would need to explain those, but also generally the rules that were applied to the data to come to a particular decision. So it's quite demanding because it might be the case that certain black box approaches just resist this kind of form of explainability, which is kind of on the interpretable side of explainability, that it needs to be interpretable for humans.

So even if you use a black box approach, then you might want to derive some set of human-interpretable rules from that black box model. That's kind of the type of explainability that we typically focus on. Or it's more on the counterfactual side of things. So then you focus more on exposing people to alternative inputs that might lead to a desirable output. So then that's how you do the explanation.

But for the type of contestation that I'm most interested in, that's usually not sufficient because it doesn't give you the adequate amount of agency really to change something

about your circumstances. You can't really appeal the decision on a more explicit basis. So the idea of the counterfactual explanations, of course, is that you, well, OK, you don't like the output that you got? You applied for a loan. It's rejected. And now we give you some counterfactual explanation. Oh, if you make this amount of money, then it would be approved. Let's say, for example. And then the people working on that type of stuff would say, well, it's contestable because you can go out and try and change something about your circumstances that would allow you to get the outcome that you want.

That's not really the type of contestation that I'm after, because I want to enable this dialogue between stakeholders. So you need to be able to have an exchange of views about a system, and that's really hard to do merely on the basis of a counterfactual explanation, because as a recipient of a particular system output, I would still not be—I can only guess why a particular change in an input would lead to a particular change in an output. So it's insufficient for me as a basis for articulating my objections. Because I'm still left with just saying, well, I don't like the output that I got, and I want to change it, but I don't know why it is that I got this output.

00:27:08 Interviewer

And I want to make sure that I understand. When you say explanation, do you mean the end part or the entire process? Because you mentioned the exchange of opinions between stakeholders, I'm not sure if I understand with what you are speaking about.

00:27:23 S4

Yeah. Alright. We're still staying with this local view of an individual being impacted by an AI decision. And then this exchange (exchange of opinions between stakeholders)—literally, what I call a "*decision subject*," are someone subjected to the decisions that are made with the help of an AI system... This is mostly in the context of public AI systems, which is what I've been focusing on, but to a large extent, it would also apply to commercial settings.

To contest such a decision fully would mean that, at some point, you are able to enter into a dialogue with a human on the side of the organization operating that AI system. And that's because only humans are able to have this type of debate. This type of debate is the essence of what it means for something to be contestable. It's not something that you would be able to fully automate. When it comes to the point where you need to have that dialogue, as a *decision subject* who objects to a particular decision, you would need sufficient information to make your case basically. The type of explanation that you need is what would reasonably allow you to articulate your case in your defense.

That's the distinction I was making just now with the counterfactual approach. The counterfactual approach presumes that it's not necessary to have the possibility of an exchange of views or dialogue with a system operator on the other side. From the perspective of a counterfactual explanation, it would be sufficient to give a *decision subject* information about what they would need to change about their situation to get a different outcome. In that view, contesting a system would simply mean changing something about your situation to get the output that you want. But in my view, you're not really contesting the basis or grounds for a particular decision—you're just manipulating the system to produce the output that you want. That isn't the same thing. It's a form of control, sure, but it's not conversation. Conversation is a more expansive notion than what the counterfactual approach would afford.

00:30:54 Interviewer

And in that sense, if I understood correctly, I think this kind of explanation would facilitate, for example, the *decisions subject* to be able to have sufficient information to go back to a human to have this kind of human-and-and dialogue.

00:31:15 S4

Yeah. So understanding "why," right? The work that I've done is grounded in a particular idea of what it means to be autonomous as a human, and there are two components to that: practical agency and cognitive agency. Practical agency is about being able to make plans and act on them, you have the ability to enact a plan that you have. In the context of an AI system, changing something about it would mean you need to have the levers to change something about it. That's one aspect.

But the cognitive agency element is also crucial, and that connects to explainability. Cognitive agency is about being able to accurately evaluate your circumstances—to have a proper understanding of your circumstances. If you have insufficient information about a system that is shaping your life, then your cognitive agency is limited. In my view, this is the shortcoming of counterfactual explanations. I'm just using it as an example, right? It's not the only case, but it's a helpful one to make the distinction. The counterfactual explanation gives you some practical ability. It tells you, "Oh, if you change this particular input, then you would get a better output." But it hardly increases cognitive agency because you still don't fully understand or know why that is the case. That's why, from the perspective of autonomy, counterfactual explanations fall short. For me, contestability is closely linked to autonomy. It's a way to respect people's autonomy, and you need to do both: provide the practical means for people to act, but also the cognitive elements that allow them to properly understand the world they are in.

00:33:29 Interviewer

I suppose this is mostly on post-hoc explanations, but what if the system already have embedded biases? How to make such a system contestable as well? I'm not very sure.

00:33:45 S4

Yeah, I mean, bias and fairness are related concern. In a way, contestability is also a means to address harmful biases. On the post-hoc side of things, if someone is adversely impacted based on a group characteristic that they feel is unfair, then contestability provides a way for them to appeal the system on those grounds. This is why it's so crucial that people get a sufficiently accurate picture of why the system is making decisions about them in the way it is. That's on the individual and local level. I guess what you're asking about is also on the global level. On the global level, you need a separate set of techniques. For me, contestability at this level involves participatory approaches to system design and development. It's about being more inclusive in determining who gets to shape the system as a way to raise awareness about potential harmful biases. That's a separate stream in the work I've been doing, which is about developing new tools and methods for machine learning system design and development that enable people to participate fully. To achieve this, you need to help them understand the decisions that go into the design of such a system and evaluate whether the trade-offs being made as part of the engineering process are acceptable. That's a whole separate set of challenges, but it's essential for ensuring systems are fair and contestable at both the individual and societal levels.

00:36:16 Interviewer

If I understand it correctly, the postdoc explanation acts as a loop and goes back into the system where you could potentially improve with that.

00:36:28 S4

Yeah, that as well, yeah. That is, I would say, a third element. There's the ex-ante phase for a new system where there are opportunities to use more agonistic or adversarial, inclusive, or pluralistic approaches to participatory machine learning design and engineering. These approaches act as a safeguard against harmful biases to begin with.

I operate under the assumption that even if you implement all of that—which isn't often the case in reality—it's likely you will still miss things. And anyway, the world is not static; it's dynamic. At first, the system might operate within acceptable boundaries, but as circumstances change, it could become harmful in some way. That's where post-hoc, local forms of contestation come in.

Thirdly, yes, indeed, those individual contestations or appeals can serve as a signal, essentially acting as a feedback loop. This loop can potential signal the structural deficiencies in a system and provide input for ongoing system maintenance and further development.

00:38:15 Interviewer

I wonder the speculative design project that you did, where does it situated in?

00:38:24 S4

Ah, you mean the [a speculative design project on AI] project that was published at [leading HCI conference]?

00:38:25 Interviewer

Yes.

00:38:32 S4

Well, OK, so that grew out of some theoretical work I've done on contestability. I wanted to translate that into artifacts that would embody some of those ideas in a way that would be more intuitively graspable by audiences. That was the initial motivation to create a short speculative video. The video was like a near-future scenario or vision of a more contestable AI system—what it might look like and the benefits it might have. That was the starting point.

Then I paired it with more scientific motivations or projects that focused on the challenges public organizations face when seeking to implement contestability ideas. I used the video as an instrument, a prompt, as part of an interview study. What it brought to the table was that it grounded the conversations I had with civil servants in a concrete idea of how things might be. This approach allowed them to speculate about the challenges they might face if they were to pursue such a vision.

In most cases, the people I spoke to were working on AI projects in the civil service.

However, they were not explicitly pursuing contestability, at least not yet. What the speculative element did was allow for conversations about a future that hasn't arrived yet. That's how I combined it with the other aspects of the work.

A nice side benefit of doing it this way is that I now also have a video that stands on its own. It helps people quickly grasp what I'm talking about. Many of these ideas are abstract, and it's often hard to think of concrete examples, but the video provides that clarity.

00:41:21 Interviewer

So I think that project offers a fictional space where you could evaluate what contestability could do and get some response from them.

00:41:30 S4

Yeah. Well, it's actually one step further because I preempt the question of what could it do. I just tried to show it and make the case tangible—why I think it's a good idea. Then it lets people explore or think about what it would take to get there.

Of course, you can have a debate about whether I'm right—would it indeed offer these benefits? That's also possible. But you can also, if you grant that it is desirable to pursue

this vision, consider what it would take to actually achieve it. That's the direction I took with this particular study.

00:42:17 Interviewer

So, you're kind of planting a seed in their head that this might be desirable, this might be preferable?

00:42:24 S4

Yeah, I'm definitely not shy about putting my cards on the table. I think it's a good idea for particular reasons. I'm not trying to sell them on it as such, but I do try to make the argument and be transparent about my agenda. I think that's the right way to do it. Some people might say, "Well, you're priming them," or something like that. That's maybe true, but within the context of the type of research I'm doing, that's not actually an issue. Because, again, I'm asking them to go along with that line of thinking and help me think through some of the challenges that might arise. These are challenges I might face myself when I try to implement this going forward.

00:43:17 Interviewer

I think most of these projects, efforts trying to make an ethical or more responsible AI, involve this participative aspect, whether it's during the validation phase or throughout the process. Do you have any suggestions for involving participants? For example, you involved civil servants, and I bet they are hard to reach. How did you approach that in your project?

00:43:41 S4

Yeah, it is hard. And it's even harder when you try to engage with, for example, citizens. I'm currently working on a project in [city 2] that is in part inspired by the work I did on [a speculative design project on AI]. It's also about a computer vision system. What's interesting here is that it's a technology project being developed by some in-house development teams in [city 2]. They've also got on board with this participatory idea and put together a citizen panel for the duration of the R&D project, which spans over a year, I think. They selected a group of people through an initial survey. These people were then engaged throughout the R&D project. They were also given a voice in the reports that the R&D team prepared for the city government. Not only were the citizens included in the engineering process, but if political decisions arose that needs to be handled by the city government, viewpoints of those citizens were explicitly included in the requests sent to politicians.

That approach is pretty new—it's not typically done that way. It's also incredibly time-consuming, resource-intensive, and not necessarily scalable to all technology projects that a city undertakes. However, it illustrates how hard it is to include citizens in such a project. That's why I'm more inclined to focus on existing institutions already in

place that represent larger groups, rather than fixating on direct participation. For instance, labor unions could be one such institution, if that makes sense.

On the research side, when it comes to recruiting participants for studies like this, which are deeply embedded in a particular organization, I'm only able to do it because I've spent years building relationships in that particular organization, building a network, and investing in those relationships. At some point, this pays off (if you want to put it that way) when making requests to talk to people.

But you need to... I need to spent a lot time, just being present, talking to people, and doing my thing, and not expecting much at first. Slowly but surely, your network expands. It's really about meeting people in person and building up that network that enables you to do things that would otherwise be impossible.

Similarly, I recently conducted a study with several design agencies in the [a country]. Honestly, I was only able to do that because I had spent 10-15 years in the design industry here. Some of the companies I approached were former clients, competitors, or whatever. Otherwise, it's very hard. I have to benefit from that network in that case and that makes it easier. Again that that means investing a lot of time and energy in that network—it's not something you can shortcut or rush.

00:48:44 Interviewer

Since we are running out of time, I don't know if you are ok with just two more questions.

00:48:50 S4

Yeah, that's fine.

00:48:52 Interviewer

Since our project is about trust, I wonder how, like all the aspects you talk about, contestability and also the ethics, how do you think these can contribute to building a trustworthy AI system?

00:49:07 S4

I have this thesis here, and trustworthiness is something I'm still not 100% sold on. I'm not sure it makes sense as a design goal. I actually have a paragraph in the concluding chapter of my thesis around trustworthiness. Let me... yeah.

For me, trustworthiness is linked to contestability. I think when citizens feel in control—or more generally, when people feel in control of their lives or the systems that impact their lives—they're more inclined to trust those systems.

For the same reason, I think it's insufficient to focus solely on explainability and transparency. There's this line of thinking that comes up quite often, particularly among technical experts and people in government: "If we just explain what we're doing better, people will trust it." But I think that's a fallacy. Explaining doesn't, in itself, allow for sufficient control.



In political science, when people talk about trust in political institutions, they often discuss how trust is something that tracks other things, namely the predictability and reliability of institutions. I think the same applies to AI systems. If those systems work reliably and predictably, people are more inclined to trust them.

But this idea of trustworthiness as a design goal? I'm not so sure about it. One of my favorite papers on this topic talks about trust in terms of a leap of faith (*A Leap of Faith: Is There a Formula for "Trustworthy" AI?*). It's by Matthias Braun and others from 2021. They argue that trust always involves some leap of faith. At some point, you need to make the jump, and that's an indication of trust. But this also implies there's always a residual lack of trust. There's always a lack of trust that remains. I don't think that's necessarily a problem. In fact, it might even be desirable. That's why I'm a bit ambivalent about having trustworthiness as your primary design goal. It might lead to practices that mislead people about the true nature of the system. You could start fixating on making sure people aren't concerned about what's going on.

From a democratic perspective, it's perfectly fine if people don't fully trust a system. You need a modicum of trust, but people don't have to completely trust any particular system. Instead, if you care about trust, as someone building or operating a system, you should focus on being reliable and predictable. If you focus on those things, trust is likely to follow.

00:53:55 Interviewer

We're designing a system where the decision subject is the patient but the primary users—who check the decision of the AI system are radiologists who reads the MRI scan. The people who are most affected by these decisions are not included in the loop. How do you think, for example contestability could be included in this system where there's dual decision subjects?

00:54:36 S4

Yeah, I mean, this is actually not that dissimilar from many public AI systems where you have a frontline civil servant who typically uses the inference from some kind of model to inform a decision they need to make as part of a policy execution task. So, it's actually quite similar.

I think the nice thing about many of these diagnostic aids is that there's more opportunity, and there tends to be less distance, between a caregiver and a patient. There's more opportunity, in theory at least, for this three-way dialogue between the AI system, a caregiver, and a patient. This dynamic could support making sense of a diagnosis provided by an AI system.

There's some work on contestability in the medical field that I've referred to in my theoretical work (*The four dimensions of contestable AI diagnostics - A patient-centric*

*approach to explainable AI*). For example, there's a piece by Ploug and Holm. They explicitly discuss this relationship.

In my work, there's some overlap, but I've had few opportunities to really focus on that particular dynamic. It's quite interesting and connects to the idea of interactive machine learning—where you can interact with a model in real time and manipulate it in various ways, as either a patient or a caregiver. That's really interesting. I did notice your project is in the healthcare domain. What was the condition again? Was it MS?

00:56:55 Interviewer

Yes, MS.

00:56:59 Interviewer

So, I want to ask about something you mentioned earlier regarding how hard it is to engage citizens for AI in the public sector. I think healthcare faces a similar issue in engaging patients. How do you address this kind of challenge where you cannot really engage the people who are most affected by AI decisions?

00:57:25 S4

Yeah, like I said, for me, we managed to do that at least in part in this most recent project by having close collaboration with the organization developing and deploying a particular system. Working with a city government who has means for connecting to their citizens—at least in theory. Some are better at it than others, and some have more infrastructure for doing that. [city 2] is quite advanced in that sense, so that's a benefit we had there. I mean, I don't have... In the medical context, for instance, [a close family member] happens to be a medical doctor, though in a completely different field. There are always patient associations and similar groups that, at least in the [a country] context, are typically partners for these types of collaborations. But it's quite hard because you often face more ethical demands when conducting studies involving people suffering from a condition. It adds another layer of complexity.

So, I don't have a quick fix for that. It's a lot of hard work. What can I say?

00:58:51 Interviewer

I'll wrap up with the one futuring question, how do you envision like designer's role contributing to the AI systems in 5 or 15 years in the near future?

00:59:11 S4

Those are two very different time horizons. The future role of designers, yeah. I don't know, I mean... I come from a UX and interaction design world myself. When I did this most recent study, where I reached out to consultancies working on at least data-related projects or public sector projects, in the [a country], we have a pretty mature design field,

but it was still quite challenging to find companies with substantial experience or profile in that area. That already indicates that it's still quite early for many.

Talking specifically about UX and the interaction design field, I'm a bit worried they are missing the boat—or perhaps have already missed the boat. They seem stuck reacting to whatever other fields impose on them. Missing the boat, not necessarily in terms of not adopting the latest technologies, but more in terms of not wrapping their heads around what UX or interaction design has to offer, specifically for the types of problems people face with AI systems. That's one part of it.

The answer for me to that question is... UX and interaction design has always been about, at least in part, agency and autonomy for users originally. These systems have far-reaching consequences for people's autonomy, for me this is clearly something that UX people can clearly contribute to, making sure people still feel in control of their lives while interacting with these AI systems. I can't think of any other discipline better positioned to address this problem. But I don't see many people talking about it in these terms. Maybe they're nervous about talking about it in those terms.

Locating design in the first place in many AI product development processes is quite hard. Some often quite marginal in terms of what is explicitly labeled — obviously that there are many that is happening — as design activities. but they are done or made by teams or groups with no formal design training or job titles that include “design.” But design is still happening.

For me, the question becomes, should I focus on people who identify as designers, or should I focus on the people making the type of design decisions I care about?

If they don't happen to call themselves or don't identify as designers, who cares? That's not the point.

For example, just the other day, I spent half a day in a workshop with ICT architects. Most wouldn't identify as designers, but many of the decisions they make—like planning ICT systems that include machine learning components—relate directly to the ideas of contestability we've been discussing. So perhaps I should be talking to architects, rather than designers. Or maybe I should be spending more time with engineers. I don't know, what do you think?

01:04:17 Interviewer

I think It's interesting sort of training the people to utilize design research, for example. They are necessarily have to identify themselves as designers. Maybe that's another practice. And also engaging more designers in the field that's another thing.

01:04:38 S4

Yeah, totally. Back in my practice, the perspective I took was that the design is what teams do, groups of people do. A team might include one or two trained designers — who has design in the job title —but ultimately they're useless unless they bring the whole

team on board. As a senior designer, I often take on the role of facilitating so that the team would be able to design properly. I think that's a much more productive way of thinking about it. So more as an activity, something that people do rather than a fixed role with a monopoly on design—that they do the design, and the rest just follow. I don't think that's a very healthy or productive way of approaching it.

I also like the idea of designers as stewards. This aligns with what I wrote in my thesis. Thinking design as stewardship, that also mean design isn't just a stage in system development—it's something that should accompany a system throughout its lifecycle. A designer's role should extend beyond spending a chunk of time at the start and then handing the project off to another group of people. OK, I will leave it there.

[end]

01:06:32 Interviewer

It's super interesting. I I don't know if you know the lab in London called Dark Matters Lab and the director called himself head steward of the lab.

01:06:37 S4

Yep.

01:06:44 Interviewer

Yeah, not not as someone else. So I think that's that's really interesting to hear like how to say to to implement in the the area of AI and us designers as facilitators and then help. Sort of. The training, training people with more, for example, empathy and and awareness of the decision that they make for the decision.

01:07:14 S4

Yeah, yeah. Oh, by the way, this is like Höök and Löwgren (*Characterizing Interaction Design by Its Ideals: A Discipline in Transition*). And in this citation, they also talk about that in relation to design and and machine learning so that.

That. Helpful one for you.

## S5

00:04:28 Interviewer

And I think maybe I just go back to to what you mentioned earlier. I'm very curious about why did you choose to work more on XAI and interpretability? What was the initial goal of focusing on this?

00:04:48 S5

For me, the main reason of working on XAI is not to understand the black box of neural networks, but it's rather to ensure patient safety. Doctors have this statement, “do no harm”, basically everything they do should first put patient's safety at the top priority

And there is where XAI... The hope is that we can provide that security because we can catch biasing models, we can catch shortcut learning, we can confirm the concept of human-machine alignment, which is basically how an AI system looks at the data and whether the features used by that model to solve the problem are similar to those used by a human. If an AI system behaves like a human, a human might understand the behaviors of the machine because it mimics humans, say visual perception or the visual cortex. Typically, we know now that neural networks use very high frequencies that humans cannot see. Therefore, there is a misalignment there, for example, between what the human and the machine use as information to solve a problem. So, there is where XAI can elucidate things, can bring that safety where we understand how the system behaves. We can perform quality assurance, or we can audit a model. So that for me was the first motive. Why I wanted to start knowing more about it (XAI) patient safety. As I mentioned, we have an FDA system approved that's being deployed in the US and a nightmare is that you get phone calls saying, "the system just doesn't work. We don't know why or the system is really causing harm because we missed lesions and the tumor grew up and the system completely failed there." So, as an engineer for me, it was important to place myself in that situation of a real-world application and it's my responsibility to find mechanism to audit that technology.

00:07:26 Interviewer

I think for different people, the understanding of explainability and interpretability is a little bit different. Can you share what do you think the interpretability or the explainability of a system is?

00:07:40 S5

Yeah, to be honest, for me is potato-potato. It doesn't matter because (what I said before) it ultimately is about patient safety, so I can interpret, or I can explain a model using the different distinctions. Ultimately, I want to have a statement about patient safety. So we have had discussions. There will be a Springer book soon on trustworthy AI, where we have a chapter on interpretability and we have a chapter on explainability. For me, this is an academic exercise. But ultimately, (as I mentioned) in a real-world application, it doesn't play a substantial role because both should aim at providing patient safety.

00:08:31 Interviewer

Since we are working on the idea of trust and I wonder, how do you think of this concept or what you're working on contributes to building trustworthy AI?

00:08:46 S5

I think it's a double-edged sword because the problem with trustworthiness is that it is very subjective to humans. "I trust you, but I don't know you, and I just trust you because you show me some data. That's OK. Oh, now you know your things." It's like someone's pitching to you and sounds fancy, but maybe it's very hollow after that. So the trustworthiness is quite subjective and it's very subjective to also backgrounds. Maybe a younger person that's more tech savvy and I say, "yeah, I trust this system because I like technologies", whereas a senior person says, (you know), "I don't wanna work with a machine. I don't trust it."

We have seen papers where for example a user first see an AI result with an explanation and then another situation where you ask the doctor to perform a diagnosis, and then the AI and XAI data is shown. In the two situations it's not the same. In one aspect is that people say, "well, I already made-up my mind. I'm not going to change it just because this AI system is telling me otherwise." There is a (confirmation bias, no it's called) anchoring bias, where the doctor say "I know my profession better than a machine. So I'm not going to change my decision." There you have cultural and very subjective aspects that make this whole trustworthiness very difficult to handle.

(There is also and I mention) it is a double-edged sword because it could be that in the future people have been working with this AI system in the hospital for years, they don't check anymore. Like I don't check the calculator on my phone because I think "well it's just a calculator". This could also happen that with AI, the trustworthiness would get to a point where people don't even check anymore. Is that a good thing?

That's (sort of) the pivotal point where (you say, well,) doctors don't even think about that result and they just blindly trust the system. It's like death by GPS, people follow the GPS signal, they don't even look ahead and people have died because they just don't look ahead. They don't drive, they just follow. "OK, now I should turn left. OK, left." And accidents happen. So that's basically the same situation where people (sort of) turn off their brain, just trust the system and bad things can happen.

00:11:56 Interviewer

How do you think we can balance this over trusting and not trusting? How do you think we can try to achieve or what do you think is the best level of trust?

00:12:11 S5

I think the best level of trust comes with transparency, where there's the (whole) demystification that AI is true intelligence, that AI... Basically people don't know what the error rate of a system is. AI doesn't know what it doesn't know. It's also very difficult for AI system itself, then we go back to uncertainty, it's very difficult for a system to reliably say, I don't know.

I always say it's like a student in an oral exam saying "do you know this question and the student might say, yeah, I know". (You know) the student has an objective which is passing the exam. Asking the same student to self-evaluate himself or herself, it's a design problem, in my opinion. I think this is where you need information that is statistically reliable that by design does not depend on the same system you're evaluating,

And then it's also from the side of the user to demystify any wrong perception about the system, how the system will figure out, like wrong annotations like the human needs. For example. I'm thinking of systems where you need some human interactions and people they say ah, "the system will figure it out." They just very quickly provide annotations for example. Showing the limits of the system to a user during a training session for example, showing pitfalls of the system (like let's say) this system for MS (Multiple Sclerosis) is very susceptible to patient motion. The system is very susceptible to very small lesions that are close to that area or very close to the skull. That brings transparency and informs the user about "Ohh this system is not perfect, but at least I know those failure modes."

00:14:31 Interviewer

Can you give us some kind of a context? Your work related to this, trying to achieve a balance of trust during (for example) trainings and also more information provided to the doctors.

00:14:51 S5

Yes, we were just talking about it and. For example, as an analogy, if I tell you "I have a high uncertainty in predicting the temperature in 2032". You will say, "I don't care about 2032". But if you tell me, "I want the uncertainty on the temperature estimation for tomorrow". That's more relevant. The same in medicine. Many times we calculate uncertainty values, but we don't connect that to the clinical relevance. I think that any information that is provided to an MD should be connected to the impact of that decision. If the system is saying "ohh this is a cancer patient and (that let's say) with a 60 or 40 yes or no cancer patient decision. It's (kind of) uncertain. And that decision leads to, for example, radiation. You need to take chemo with surgery or whatever. That decision has a high risk. Then if let's say in another case where, "OK, just take an ibuprofen". The risk or the consequence of that uncertainty evaluation is way different.

And I think that's what many colleagues they miss that they calculate something. They show that this number is better than the one from previous year. And being sarcastic and then to connect then they say, "well, what's the deal with this number?"

What is the ultimate consequence when I show it to a doctor? What we've been discussing is how to connect the dots between an engineer number and the clinical utility and impact of that number.

00:17:13 Interviewer

I think this is related to what you what you said before it's all related back to the patient safety, isn't it?

00:17:21 S5

Right, correct. I had to tell the patient, "you will start chemo" and then you have this little of patient safety. And it's there's kind of an end. It's like showing a certainty in places where it's really relevant, right. Elon Musk was once asked about would you show uncertainty maps on what your sensors are telling. Like, it's a dog crossing the street and you want to tell maybe with some heat map on the certainty of the detector. And he said no. Because it would be too much information. It might confuse the user. The user is driving has to know "the machine is not sure"?

Ultimately, you have this balance of information and relevance of the information. The same with doctors. Many times we engineers want to provide a lot of visualizations and data and we overdo it and (I think here) the key is to show whatever is relevant, "show me in red things that I should be careful about".

00:18:38 Interviewer

Except for the training that you mentioned, how do you communicate the uncertainty or the relevant information should be provided to the doctor is through what forms?

00:18:53 S5

I stopped working on uncertainty estimation because I had one PhD student working on it, and I realized that uncertainty estimation can be calculated and there are different ways, but it's not reliable because models tend to be overconfident. Neural networks are (really like) they don't know what they don't know. Then I realize I have another problem. The first problem was the clinical relevance.

I can calculate something, but I need to place it when it really matters, but then on top I realize oh some of these numbers, sometimes they work, sometimes they don't. The system will tell me 99%. I'm sure this should be like that. And then sometimes the system was right. But then in the next case, completely overconfident. It was bluffing. And I realize we have a problem, what we call in the neural network, model calibration. It's still an open problem. There are ways to do it, but still an open problem.

So there is where again you start like "umm I I have two problems now and one is more important than the other, but it's very hard". So instead of calculating a number and then hoping this number is correct, finding when to show it.

What I decided to do... I work in cancer, the AI systems need to provide data to radiation therapists. This was for brain cancer. The brain needs radiation. Then I realize if I can inform the doctor on areas that show how their work might affect negatively with that extra



radiation, that's already information they may use. I'm not showing them when or where the system is not correct, I'm showing them where their decisions might have or might not have an impact.

Many times they need to operate on the data or they might need to correct an AI system, say "OK, here I would correct this and now let's use this data to plan the treatment instead of showing them all this. The AI system might not be put here, here and here". I'm telling them, do you know your corrections. And I'm telling you whether that is worth doing. Because sometimes they do corrections, they spend an hour, here "I would just modify the AI system result in this way". And the modifications or the extra time they spend is completely useless. Wouldn't make any difference after treating. So basically the whole workflow all the way to what really matters there is where I want to measure, relevance. That's a mindset change, we have between let's show where the air system is not sure" to "let's show them whether their workloads or their time spent is meaningful".

00:22:06 Interviewer

How do the doctors respond to that?

00:22:14 S5

They had a lot... Because they spend less time in front of a screen, less time on the phone with the other colleague who has to run all these optimization processes. They cannot do themselves and they directly can see it, if I do this half an hour work and the system is telling me "you won't see a difference in the final outcome". That's a really half an hour saved that they can spend talking to the patient.

00:22:44 Interviewer

From what you understand, what do you think are the challenges to implement such a tool in the clinical setting?

00:22:57 S5

User interfaces for sure. You want systems that are well integrated. (You don't want systems where you need to export, import, click here.) Half an hour you can start using it. It's all about smooth experience and minimalistic UI. Understanding what they need to know and everything else should be removed. That's one challenge. We are not experts on that, so we need partnerships to think about and all the interactions with them, "what kind of color maps do you want? Do you want blue to red? How do you imagine it?"

And doctors have this problem that, they have their daily problems, they have their daily routine, so it's very difficult for them to move out of their comfort zone, to move out of what they already know, so the typical answer is "I like to have something similar to what I'm already using (from each company)". That's a very common pattern that

they don't want huge changes. It's like they have been driving this car with buttons everywhere and you give them a Tesla car where everything is one screen. For them, it's like, "Oh my God, I can't handle that".

And the other challenge is you are calculating things to then tell them "Don't worry about that". So I have to be sure that what I'm calculating (again) is reliable. So robustness and patient safety (again), so I also have to think, "I'm calculating something, what is the impact of that calculation? What is the worst-case scenario if I messed up?". And that's hard for an engineer because you are putting yourself in a situation where your system failed and you don't want that. As an engineer, you always think about this is going to work, but no, here you need to put yourself in that situation that the system is failing. What we actually do is that we have one fellow, and their job is to make the system fail. We are not in charge of making the system fail because we are full of biases. It's engineers, we create something and that's your baby. We need someone else who is not conditioned by that and the job of that person is to find ways to make the system fail and tell you when it failed. And I think that's a much healthier approach to dealing with the problems of robustness or reliability.

00:26:01 Interviewer

Do you do this process before giving it to the doctors and to check and verify its reliability?

00:26:11 S5

I would say in parallel. We also want the system to be looked at by the MD and that typically is the feedback like, "how do you see it? how do you feel it?" In parallel you want to start like checking your new tool and seeing when it fails. I call it check and see, "you pock it from here, from this side, from the other". In our case, for example, we add noise to the images we simulate patient motion we down sample the image like crazy to see if the system still works. We simulate the patient rotating the head during the scan in a crazy manner. "You stress the system". Normally we do this in parallel.

00:27:02 Interviewer

Do you think in that sense, the doctors developed more trust by giving more information or having this verification process? Do the doctors tend to trust more?

00:27:18 S5

I've seen 2 profiles of doctors, one say, "this is none of my business. I'm not an engineer. So I don't wanna know. This should work." It's kind of like a buyer. Doctor that says, "I'm getting this. You are responsible of making it work. And if it doesn't work then I don't trust it anymore." That's the harder kind of consumer, because you don't have that opportunity of creating that feedback loop, that trustworthiness builds up where I

could say, “look, we run all these tests, the system tends to fail in these circumstances. So try to avoid that.”

Some doctors appreciate that (as I mentioned), because you're providing the blind spots of the system. (as I mentioned) other doctors, they don't wanna know, they don't wanna be involved. They think this is like “I buying this device and it should work. And it's none of my business to know all the tests you run.”

But I think in AI in medicine due to the complexity and the critical aspects (we mentioned on safety), they have to be together.

00:28:41 Interviewer

How do you communicate this system failing to the medical doctors?

00:28:49 S5

Visualizations. We pick metrics they know. For the given task, let's say “you are classifying something or you are delineating something in the image”. Then we show them this is with and without the perturbation. What is really important is how good the system is, and this is the integrated variability, “this is where the humans are”. Basically they don't wanna see the system is below the band of human performance, they wanna see you are in between. So that the humans can rely on that system. It's kind of a virtual colleague doing the same task. Then you should “Look! if the patient would have moved, then the system doesn't work anymore.”

00:29:44 Interviewee

And we go back to challenges in clinical integration. Previously you also mentioned the doctors like the previous system that you developed because they think it saves time. Do you think it's an important thing to consider in developing?

00:30:03 S5

I think time is super critical. We are an aging society. We have short shortage of MDs, less MDs are going actually in medicine and we don't have enough doctors. They are getting more and more data because maybe 20 years ago, having an MRI scan in a center was a big deal. Now they have 3 or 4 here in [a city], for example. So the amount of data also increased. They have to look at more patient data. There is that imbalance. That's where these technologies try to save time. And there are boring tasks. (Many of these things) automation is quite boring. I would not be a doctor. Just for that reason.

00:31:07 Interviewer

For other barriers except time, do you think for the adoption of the tool? And what factors do you think is important? We talked about time and what else you think it's very important.

00:31:30 S5

Ease of use and then we go back to similarity to things they know. You should not design things that are too off from what they know or what they expect. In a way, I think that makes the whole development quite incremental because you have to bring aspects from other systems and you need to bring innovation with caution. You need to understand this group of users, they can deal with paradigm change where they don't use the mouse anymore, they can use these goggles and do things with their hands.

So easy of use and the whole friendliness.

It shouldn't be an added, "Oh my God, I have to learn this new thing and it shouldn't be like (it's kind of also human nature) "This might be better for the patient, but that's half an hour extra for me. Oh my God. We really have that half an hour." That it's not just about patients' sake, but it's also the user's sake of adopting a technology. if you say "in terms of patient-care might be the same result but it will save you time". They will, "Ohh yeah. Yes, I like that."

And then cost. Many times decisions of the new technologies go through the financial department and unfortunately there is where you say well you reduce your time for the patient may be similar or better. But for the hospital, that will require an investment or require an extra machine. It's complex and one needs to consider also the business component because many hospitals, they need a cash flow. So many super good ideas are killed just because actually go against the cash flow.

I have colleagues who for example take CT scans because they need to amortize the price of the CT scan. That's the dark side of healthcare, where you do things just because you can charge the insurance company.

00:34:11 Interviewer

Can we go back to the process of uh of implementing a tool and can you walk me through it? Who are involved? what testing or what kind of method you use during that process?

00:34:28 S5

To implement an AI system, you need certifications. You need to prove that your system doesn't harm the patient, that it has a level of performance that's equal or superior than a close competitor. And that requires... in the case of AI, you need to show that when you test your system, your test data set is representative of the population it's going to be used on, and so they have some kind of rules of thumb. (for example) For FDA, if you want to use European based data, they say, "well, it shouldn't be more than 50% of the total data you're using". And the data in general should be

representative of the cohort. You need to show numbers, metrics you need to typically compare yourself to (in the FDA they call it) predicate device, which is kind of “This is my closest competitor. These are the metrics that the system has in terms of goodness”. And then you say, “OK, I’m as good as this one or superior”. And then is documentation. So that’s the procedure in a nutshell.

00:36:03 Interviewer

For developing a system that is more trustworthy, I think I’m sure that you involved more people in the test. I wondered how does that process works?

00:36:19 S5

Trustworthiness is not evaluated when you want to have an FDA or CE labeled device...

00:36:30 Interviewer

I mean you mentioned how you measure the robustness, how you test the robustness of the system and there are all these contributing criteria to trustworthiness? What are the criteria of trustworthiness? How do you measure the meanness?

00:36:52 S5

I think the closest is robustness. Robustness is how the system behaves under harder conditions or how they behaves, how much its performance varies. Let’s say you have a cancer AI system that you know, a classification model of cancer patients. You want to measure metrics of classification accuracy, and you want to measure how that classification accuracy varies across subgroups or across different machines (let’s say for imaging). You have different vendors and you want to measure robustness of the performance across vendors and across protocols. Here I think the keyword is measuring viability under such different conditions.

And then trustworthiness... Trustworthy device then will be called a device that it just works independently of the vendor, it’s very stable in terms of performance. It’s not super good and they’re super bad. I would say that’s a main criteria for trustworthiness, “it just works”. A very reliable device that... (We were talking about robustness.) Imagine your phone, the moment it’s -2°, it just shut off. And then “oh my God, I don’t have a phone anymore just because it’s -2°”. That would be an example of an unreliable/low robustness device.

00:38:45 Interviewer

Other than robustness, what else do you think contributes to building trustworthiness?

00:38:53 S5

That’s something that we have discussed. When we develop systems and we have open challenges. We invite teams around the world to compete. We never place the results in terms of complexity of each case. What I mean is that in medicine there are easy

and hard cases. For doctors, there are easy and hard cases. Hard to tell. They need to show it to another colleague. A system that fails on the easy case is that a first year medical student can solve. You can imagine that produces some level of distrust. If you can demonstrate that how your systems performance varies, as a function of complexity of the input of the problem. That is, in my opinion, quite important. What we tend to do in academia is that we just report average values independently of the complexity. That's, in my opinion, a mistake. If you could have a measure of complexity or you could rank your data as a function of complexity and then show how the model performs across these levels, I think that can also help in the evaluation of trustworthiness.

00:40:25 Interviewer

You mentioned you're not really sure the doctors would even want uncertainty. I wonder how do you think about that? Could you elaborate more on that?

00:40:38 S5

The question is whether doctors want uncertainty?

00:40:46 Interviewee

MHM

00:40:48 S5

I think it's an "if" type of answer because if you say I will show you uncertainty and this will improve patient safety. if you give me that connection between your uncertainty numbers or your heat maps with colors and patient safety. I want it right and I'm glad to look at it for maybe few more seconds before making my call.

I think it's a matter of if you bring values to the numbers, in terms of time savings, or patient safety, showing consequences, "OK, the system might not be sure, but what's the deal if I go one route or the other? I have to make a call between this or that." If the system would help me in "OK, I'm not sure about this decision, but if you would go this route. These are the consequences or pros and cons. If I took the other one, these are the other pros and cons.

I think that information is even more valuable than the system is 60% or 40% uncertain about this decision. Because we still rely on the human after the AI system. Guidance on consequences of decisions I think are more impactful than how sure that system is.

00:42:28 Interviewer

I think similar question applies to robustness. Do you think doctors want that as well? Based on your feedback from the doctors, how do they take it?

00:42:48 S5

I'm very convinced that many in medicine, robustness is even more important than an average performance. In computer vision we see everyone runs for who has the best

average (with whatever metric)? Who gets the next CVPR (Conference on Computer Vision and Pattern Recognition, it's a top conference)? Many times it's all about that average value, those little deltas, but people care very little about their robustness, that variance. When you push for the super performance, there is a tendency to make the system more brittle, less robust. When you push the training process of a AI system, you really want to sort of squeeze everything to get that push. This is where you have things like shortcut learning, where the system is pushed to a limit, "OK, I need to find another way to solve this problem". There is spurious correlation in the data and then I got the best average but in terms of variations in the result maybe you're compromising.

Imagine that you have a phone that is super-fast in downloading YouTube videos. It's the fastest and you buy it. But then in the next YouTube video you want to open, it just starts lagging. And then the next one, boom. And then in the next one, it stops. You don't want that phone probably. You'd rather go with a very simple phone that it just maybe 2 seconds on every YouTube click. You click, but you have it. Reliability and that robustness is what doctors need. Remember they are super stressed, they don't have time, they have too many patients, that they need reliable systems. Robustness (I say) is more important than super top accuracy.

00:44:55 Interviewer

Do you think that also helps with contributing to patient safety?

00:45:02 S5

Yeah. Imagine that system said "you need to treat this patient with chemotherapy". And it completely failed", then patient safety is completely compromised as well.

00:45:19 Interviewer

I think this topic of implementing these AI tools into a clinical setting involves many different people. And I wonder and in your previous work, who are involved in implementing it and in the process of developing such a system?

00:45:43 S5

It's 80% engineers and 20% other people. We like the tools to involve the other the non-engineering persons. And many times it's about the willingness. The engineers like to work in their corners and they don't like to talk much. Doctors are very busy and they might have only half an hour of their time. It's very complex. You need to find a mechanism. You need to find a convincing arguments to make it happen.

In that 20%, we have doctors... We also work with psychologists. Typically, you run a subjective analysis of AI system. Doctors need to give stars, "I like it. I like it a lot. I don't like it." There you need good designs, "how do you present the data under which conditions to show it in the smaller screen or in a larger screen?" This is where the work of other non-experts typically take place.

There are situations where you want to test the system not only with the MD, but also experts around them that are non-engineers. In the case of radiation oncology, you have those physicists who are not engineers, but they they know the physics that are used to treat a patient with radiation. You go through loops of iterations, showing them (the phsicists) the technology because it affects them as well.

But it's quite imbalanced (as I mentioned).

00:47:38 Interviewer

What are the challenges to maintain this kind of collaboration with people from other domains?

00:47:50 S5

They don't have time, so when they accept, you need to be super careful in your claims and set the expectations, realistic expectations.

Errors that I see (is like) “could you give us the data to create this AI system? It's going to be the greatest”. Then you disappear for three years because you are working on it. And then at the end you say, “oh, we have some results”. And people are “Oh yeah. Remember, I did a lot of work and they just show up now”. You want to be fair to them and say, “OK, do you mind meeting every months? Because we want your feedback to improve the system”. I see in that situation where engineers they go once to the doctor, then they disappear, they do their business, they run their agendas and they might go back to the doctor just for the final check. I think that's a challenge and something that we need to improve. I know [a consortium member], for example, she's closely working with doctors and I think that's important to keep that.

Then understand the needs of the doctor, “Do you want a paper publication out of this work? Or are you expecting an actual tool? And if yes, to which level would you like to have?”. If the doctors, “Oh yeah I want a product deployed, it should work on my patients. I'd like to run a prospective study with that.” I cannot do that for you, because I would need a company to do a whole certification of that technology. I can give you a prototype that can do this and that, but that would signify that you have to do this and this in order to use it. So setting the expectations, in my opinion is super important because some doctors they think, Oh yeah, I give you this hundreds of cases and I'm going to have this AI technology for my patients to simplify my life at work. There you need to talk about “what are your expectations?” (and again), demystify the situation.

00:50:08 Interviewer

I'll wrap up and I ask a question about how do you envision the AI model that you are working on and implemented in the clinical setting like in five or ten years? How do you envision that?

00:50:30 S5



Well, that's a tough question. I predict the future.

I still think that we are far away from the “Swiss knife system” that solves many problems at once, and still this is very specific to the task. In five years, probably we're still in that situation of I need an app for this problem or I need an app for that problem and that's one thing that I would still imagine happening in five years.

I think that the communities now are... First of all, acceptance that AI technologies are happening in the clinics. We don't hear this anymore, “AI will replace radiologists”. But rather that, “Radiologists who don't use AI will be replaced by those who do use AI”. This leads to... In five years, the adoption level will continue increasing. But at the same time, there will be pictures of systems that are created by startups with this brilliant super cool idea, but then it just doesn't work, so there're still explorations in the applications, leading to wrong expectations. In five years, you will really start stick to what works and things that just unfortunately didn't work, “it was just a cool project maybe some good preliminary results, but it didn't fly.” I can imagine that in five years.

And then maybe in 10 years... I think in 10 years the job profiles will evolve to a point where humans are monitoring. Today they do a lot of work. We expect to automate that, but in the future, they will have a little apps communicating among them, like minions, say “I have App 1 and 2. They need to process the data and I'm the orchestrator”. I'm the one saying “you and you process this data. Give me some analysis. Give me a summary.” (like) you have some kind of ChatGPT in between that helps you to orchestrate and there is a verbalization interaction in the next 10 years. Before we click buttons, and I think the whole large language model wave, they are super strong will lead us to that verbalization interactions. I hope it's gonna be that that situation where you talk to the apps and then “OK, just run it again. Or give me a worst-case scenario for this patient. Or give me a summary of the consequences of that uncertainty you just raised.”

00:54:21 Interviewer

You even envision integrations between different tools, between different apps as well, AI plays a role in explaining them or interaction?

00:54:36 S5

Yeah, we wrote a commentary paper about it, we call it orchestrator [a commentary paper on XAI orchestration]. In medicine is longitudinal and is multimodal. When you go to the doctor, they might take different scans, so there is a text report, there is a picture, there is a blood sample and so on, but it's longitudinal. They like to compare. Doctors are very good at comparing what happened between time points. So when you add those agents that work at different modalities at different time points, the complexity of the black box is just larger. Because it's multiple AI working at multiple time points working on

multiple modalities. There is where we what we propose is that orchestrator that can verbalize and can bring things to a human level of understanding to explain things. The whole system of AI minions is saying “this patient should start treatment. I want more evidence.” At the same time you as expert will start challenging as if you were challenging a colleague saying, “I’m not convinced about this decision. Show me the facts from the CT scan. Or show me the facts from the blood. Or show me the guidelines and compare the guidelines with what you are showing me.” There’s that bidirectional communication between your orchestrator element that deals with all those minions and yourself as an expert.

**S6**

00:00:02 Interviewer

OK. And I’ll start with a with a simple question. If you have any work or experiences related to AI and and legalising them.

00:00:16 S6

I’d say more... well, let me think. Actually, it was more on an academic side.

I would say a legal academic side, but practically, I used to work like 3–4 years at the [The country] Institute of Bioinformatics. And I guess that actually I worked on some legal aspects where AI was involved, but... yeah, it was not the main focus.

So mainly on the academic legal field.

00:00:51 Interviewer

And for... I’m not real sure if it’s more or less related to the academic part, but do you see... how do you think there are other challenges or barriers to implementing, for example, AI tools in healthcare?

00:01:09 S6

Oh yeah. There are a lot.

00:01:11 Interviewer

Just general. Simple basic question, sorry.

00:01:16 S6

I can. OK. So, first, accessing data.

With the data protection rules, you know you need to access sensitive data. So that’s one of the main challenges. And, you know, you have all these classical—I would say classical—data protection regulations that promote some principles like minimization. So, it means that you can only access and process data or collect data only where it is necessary, but no more. And now, you know, that’s this big data issue with AI wanting to access and process more and more data—the most possible data. So, you know, you have

a conflict here with these principles on accessing the data. On top of that, it's sensitive data because it's patient data. So, that raises a lot of questions.

Another aspect with data, for example, is usually, if you take [The country] for example, we distinguish between processing or accessing data for research purposes or other purposes. And now, when you bring these devices, like AI devices for radiology, for example, it's becoming more complicated or more complex to distinguish between what is research and what is clinical. You know, because you're training an algorithm in... yeah, right now exactly when you use it. And this classical distinction is more difficult to see now, for example. So, data, data access.

Then, I would say everything related to security and medical devices regulations.

Why? It's a bit different because, well, using software is not new in this field. But what is new and maybe a bit more difficult—well, challenging—is the evolution of the devices.

When you want to regulate and make sure that one device is safe, you need to make sure... What is the field where the software can... well...I would say evolve. And for AI, it might be a bit tricky here to see exactly because it will depend on how the device is used, trained in real life, etc., etc. So, medical devices, regulation, and security might be a bit challenging.

But there are a lot of agencies internationally that are working on that, like the FDA in the United States, for example, or in Europe. In [The country], we didn't touch it that much, but it will come, of course, some medical devices and security.

Then you would have everything that is related to patient rights and fundamental rights—I mean in the public health approach—because there are these risks of discrimination. For example, depending on the biases and how these tools are trained, for example. So, you could have biases in the dataset you use, but also in the algorithm and how you will build this algorithm. Everything related to fundamental rights. You could also see risks in the algorithm, but also how it is used externally, for example, because you know certain people or sections of the population can use it differently. For example, it might require some IT literacy, for example, you might miss (??) some population, for example, for older people, it might be a bit more tricky, or richer people will access it more easily. So, you could also have some discrimination in the population depending on how you use these tools.

For the discrimination, I see a risk with insurances, for example, because now we have a logic where they say, "OK, we reimburse this kind of surgery." But now, they will use these tools as well and say, "OK, we need to use it and see what are the risks and chances that the people will recover if we use that technique or not." So, you will not just go and say, "OK, I have the right to get this surgery," but probably they will make all these assessments with the AI before (beforehand), and it could lead to some forms of discrimination based

on the algorithm. And one of the questions is to see how people can defend themselves while facing this new trend. That's one of the aspects.

Another one is liability when something goes wrong. You need to find who is liable for the damage occurring. Now, you use classical rules of civil liability, for example. But you see that now we have a big number of actors along the chain. When you develop this tool, it's just not one company developing one tool, but they—that's not my job exactly, but I know—will reuse parts from other sources and put it on the market. Then, you have the hospital or the doctor who will use it. Sometimes, the patient themselves is part of the chain. So, first, it depends—if something goes wrong, who is liable?

And then, what we see is that it will be more difficult for the patients to prove that something was not OK, for example, because they will have to explain or prove that, at some point, maybe one or two years before, the algorithm was trained in such a way that something went wrong back then. It's highly technical. Now, they're thinking at the European level, for example, to make it a bit easier for patients to access proof and technical documentation. For example, they can show that the device was not exactly in line with the regulations. Then, the producer will have to bring some technical documentation explaining why they were. They're trying to make it a bit easier for the patient. That will be a new regulation at the European level, probably next year.

So, that's probably the most important topic, I would say: data, safety with medical device regulations, fundamental rights, discrimination, biases, and liability.

00:09:21 Interviewer

And maybe we'll go through that one by one. But I want to ask, are there—because we all know there's a GDPR for the European Union countries—but are there [The country] equivalents for that kind of law right now?

00:09:35 S6

The [The country] situation is a bit tricky as well because of the federalism. So, we have federal law with the Data Protection Act, but it only applies to data processing by federal bodies and private persons or private companies/entities. Then, you have [The region] data protection laws at the [The region] level for public bodies. What's tricky is that health and hospitals, for example, are usually public [The region] bodies. Usually, like [UNIVERSITY HOSPITAL 1] and [University hospital 2]—they need to comply with the [The region] laws because they're [The region] public bodies. So, you see that we have this sharing of laws. Depending on if you act as a private entity, we have the federal law, and if you act as a public body, you will have the [The region] law. They're usually not so different, quite an equivalent to the GDPR. The federal law was revised last year to try to be in line with the GDPR, but in the end, they modified it a bit, It's not exactly in line with the GDPR.

00:11:05 Interviewer

And you mentioned there's a difference in [The country] for data used in research and clinical settings. How do you think that would be an issue in bringing AI tools into clinical practice?

00:11:20 S6

So, you have—if you want to reuse data for research, you will probably have to comply with the Human Research Act, which is a federal law. To reuse data, it depends. Here, the focus is on what is defined as research. Human research is defined as the research on the human body, but it could also be from data to understand how the body works. Actually, the legal position is a bit complicated. But now, for me, if you just want to use data to treat patients with an algorithm improving... in clinical settings, one of the questions is: is it still research or not research? If it's research, you have some rules that allow you to reuse the data with general consent. People need to give—usually, there might be some exceptions—but they need to give their consent for research in general. And if you're out of that realm, it could be different rules depending. You could also use data in a general way, but only if you are able to anonymize it, which is quite difficult with health data now. Otherwise, you will have to get specific consent for each research or non-research use of the clinical data.

It's quite difficult to make a very short conclusion because experts do not agree exactly on how we should apply these laws now. But it raises new questions about where you stand when you use an AI tool, and this should probably be clarified in the next year. But now, we are not exactly sure where we stand. Some people will say it's always research. Others would say it's not research. These are the kinds of new questions we need to face now.

00:13:57 Interviewer

And a very specific question. For example, for gathering the training data for the AI algorithms, all the patients agreed it's for research purposes. When taking it to clinical settings, for example, the AI algorithm is definitely trained by data that was agreed for research purposes. Then, when taking this algorithm to the clinical setting, this kind of usage sort of changes. How does that work?

00:14:29 S6

Legally?

00:14:27 Interviewer

Yes.

00:14:32 S6

That's... Exactly, exactly. That's what I was trying to underline because we don't have any concrete answer, and that should be the topic of legal research, probably.

00:14:50 Interviewer

And do you have any suggestions for researchers to make the process easier? For example, what are good practices researchers should follow?

00:15:09 S6

Yeah.. So, I think what is actually [UNIVERSITY HOSPITAL 1] is doing now is that you need to put in place some clear processing internally. So You need to identify early—at the early stage of the project—if you want to access sensitive data or not. What is important is to make sure... probably because it has to go—if you make research, you need to go to the Ethics Committee, [The region] Ethics Committee. So, you need to put in place clear processes exactly for what you need to do from the start. Knowing from the start exactly what you need to do and making sure that there will be a smooth process with the Ethics Committee. But this is not the researchers who can decide—this has to be made within the institution in discussion with the Ethics Committee. That's what they did, for example, at [UNIVERSITY HOSPITAL 1], I think, with their Bureau du promoteur de recherche, they are trying to implement that. For example, I was talking about personal data, but now they're working on how to de-identify data. They're trying to make sure that when they process the data, they de-identify it in a reliable way. Putting processes in place is critical. Otherwise, you would just have teams evolving in their project, and at the later stage, they would realize, "Oh, by the way, we should get ethical approval." And if they did nothing to identify the data and meet the ethical and legal requirements, it's usually very time-consuming to rebuild this from the beginning.

Another big topic I didn't mention is that when you want to access a lot of data. It's quite rare that you have all the data you need in one institution. Usually, you need to collaborate with other institutions to get more data. I used to work on this type of framework in the [The country] Personalized Health Network, a federal initiative, which was led over the past eight years, they developed an infrastructure to share data or access data in different university hospitals in [The country]. Actually for each project, you need to make contracts and agreements between institutions to make sure the data is shared properly. Today, we do not have a federal or official structure allowing data sharing between institutions, so making contracts is very time-consuming. You need to deal with the legal offices of each institution, university, university hospital, and they will challenge every clause and specify the conditions under which the data can be used and not used. What is more difficult is, once you've completed the project and created your dataset, using it or reusing it for another project often requires going back to all the institutions for approvals again. So today, it's not impossible, but it's very difficult to share data between institutions in [The country]. Now, they've launched a new program, called DigiSanté in [The country] aimed at facilitating data sharing over the next 10 years and removing all the hurdles and obstacles in [The country] for sharing health data. But I don't know. We'll see how it

works. I'd say that's one of the biggest challenges if you want to access data from other institutions.

00:20:03 Interviewer

And you mentioned—I want to make sure that I understand it correctly—“de-identifying the data from the patient” means sort of disassociating the patient from the data?

00:20:16 S6

Yeah, actually, if you take the Human Research Act, and I talked about general consent, people...When the patient can just say, “OK, I agree that my data can be reused for research purposes,” it's possible if the data are sufficiently anonymized. The laws are coded so it means that you can have a link between de-identified data and the patients—you are able to go back to the patient. But for that, you need to process the data and transform it in a way that, if someone else accesses the data—a researcher, for example—they're not able to re-identify the person without disproportionate effort. So, it means that you need to de-identify the data to be able to reuse it for research purposes based on this general consent. The law is not exactly very precise on how you can de-identify this data. It just says you need to take into account the risk that the people who will access the data can re-identify the person. And they should not be able to do it without disproportionate effort. That's what the law says. Now, one of the challenges is: how do we de-identify this data to make it comply with the law? At [UNIVERSITY HOSPITAL 1], for example, [Professor 1] is working on this. I don't know if you know him. He's working on these de-identification processes and trying to make... I used to collaborate with him. It would be interesting for you, maybe. They've developed some tools where they have several criteria, and you can say, “OK, we use such techniques, but also we made some agreements and ensured that the people who access the data will not try to re-identify it.” Plus, they use secure infrastructures to ensure no risk of accessing the data. At the end, you have a global score and say, “OK, the risk is quite low now that we can re-identify the data.” So, that's what I meant when I said de-identification.

00:22:53 Interviewer

And I want to ask, because you mentioned ensuring the fundamental rights of patients in terms of bias and discrimination. Most tools have assessments before they are launched in clinical settings. How does that work legally to measure whether a system has discrimination or biases?

00:23:24 S6

Good question again. Now, at the European level, they adopted the AI Act. Probably you've heard about it. Actually, this regulation was meant to ensure we bring safe AI devices into the European territory—but not [The country], because [The country] is not part of the EU. And they added a few things for protecting the people, especially the fundamental rights.

In [The country], we don't have any AI Act for now. The Federal Council will announce, at the beginning of next year, where they want to go with the regulation. Whether they want to follow the European trend or go in another direction, we don't know exactly. They could just say, "OK, we do nothing and stick to the usual rules." Or they could try to get a federal law on AI, which is a bit difficult with the federal competencies in the [The country] Constitution. Maybe they'll say, "OK, we will just regulate specific sectors like health, research, or the environment." Let's see we don't know yet. But for now, we don't have any specific regulation in [The country]. The challenge is to see whether these kinds of biases would be covered under current laws.

If we talk about fundamental rights, it's usually when the state takes action to ensure they respect fundamental rights. One of these protections is protection against discrimination. If you take the [The country] Constitution, it's Article 8, which states that the public body should not treat people in [The country] differently based on age, health status, or other criteria. It's not absolutely impossible to do so, but if you do it, you need a very good reason to justify it. So it's mainly on the relationship between the state and the individuals. For example, a dermatology scan that works well for white skin but not for darker skin—that could probably be considered discriminatory tool by the state. It's hard right now to see exactly how to assess that, as we haven't faced it yet exactly. But as much as possible, the state should be very careful when using these tools and ensure they don't create this distinctions.

Once again, there's no specific rule in [The country]. But it means that when they use this AI tools, they need to make sure that, for example, the AI tool was developed with data that is representative of the population in [The country] where they will be used. That's theory or goal. It sounds simple. But in practice, it's a bit hard, even within [The country], populations in [City 1] and [City 2] can be very different. So it's very very hard to see. These are the new questions we're facing now.

And you may have a lot of different situations, as I mentioned, because you have the development of the tool, but then you also have the way you will use the tool in practice. So, the external efforts to use it and how you will use it. At each of these stages, you must try to ensure that they avoid, as much as possible, this kind of discrimination. But it's hard to give you a more precise answer. I don't—I'm not able to say, 'You should just do that, and then it will be fine.' So, these are the new questions we're trying to address now.



00:28:46 Interviewer

But thinking from us... from developing the tool stage, do you think developing—I'm not very sure if you are very familiar with the concept of explainable AI and transparent AI. Do you think for researchers aiming to achieve such things, it would also benefit from, for example, from a legal perspective in the future to bring them into clinical practices?

00:29:18 S6

So you mean ensuring sufficient interpretability of the results?

00:29:24 Interviewer

Yes.

00:29:24 S6

Yeah, yeah, yeah, it's very... So yeah, then you could go into the theories between explainability and interpretability and all these concepts. But yeah, that's—I'm quite sure that in the future this will be a requirement. Even at the stage of certification probably, when you will go through the certification process based on the Medical Devices Regulations, you'll probably have to explain how it works as much as possible. And, well, there are some legal scholars who propose to have impact assessments on fundamental rights. And if these kinds of tools develop in the future, of course, they will have to understand how a device comes to a certain result. And this will be central, probably, in the legal aspect to be able to interpret correctly the results. So, yeah, I would try to... because it's hard to explain—okay, that's a fact—but at least try to interpret the result and make sure that the results reflect what you want to achieve and what you wanted to have as results. It would be probably very important in the future.

00:30:00 Interviewer

And also, from a legal perspective, how do you think a system, for example, trying to achieve interpretability, can contribute to the trust from the legal experts? Do you think it's possible?

00:30:27 S6

Yeah, yeah.

No, no, it's... Do you mean trust from the beginning—either relations?

00:30:40 Interviewer

Trust by the legal experts.

00:30:43 S6

I mean, I don't—I don't know if we need trust by legal experts.

Ah. It depends, you know. Legal experts... there are different kinds of legal experts. There will be the ones working in the companies or lawyers who just try to go and be paid by the clients and the companies and try to make sure the results will come to the market. And

then you have the legal experts, and then you have the politics. So, it depends who you want to convince at the end. Because the legal expert, I'm not sure that's the one you need to convince, actually, probably the legal scholars, and if you can explain, yeah... Because then they will say, okay, that's a requirement, then we achieve the requirement. But then, who do you need to convince? That's the question: the hospitals, the doctors, the insurances that will have to insure the risks, or the politicians who will make the rules and decide what the requirements are and under which conditions you can put that on the market. So that's probably the people you need to convince at the end.

00:33:03 Interviewer

Yeah, I think first of all would be hospitals, and then would be the person who approves this tool to be in the clinical—to be used in the clinical setting or not. And then it's the insurance companies. So, I think these are the three legal experts we are aiming to convince the most. And do you think, in that sense, that explainability or interpretability is going to help construct this kind of trust? Or do you think, in some cases, this kind of lack of trust is also a good thing?

00:33:48 S6

No, honestly, everyone is calling for explainability, so I don't have any reason to go against that. I mean, I would be in favor. So, if we talk about trust... yeah, no, I think it's necessary to try to improve that. So, for me, it's a central aspect.

00:34:13 Interviewer

But I want to ask, for example, maybe for the population, also there's a question of adverse trust, like over-trusting. Is there any—how to say—measures or aspects from the law that could ensure this thing wouldn't really happen in a sense?

00:34:40 S6

Not really. I mean, the law... I'm trying to think where the law might have a role here, but... I mean, you have the certification processes to make sure that's where it happens because you have these medical devices regulations, and then you have some private—notified bodies, as we call them—who make the assessment. Usually, they are in the European Union, and that's how the system works. Then you have these bodies, private bodies, who are regulated by the law, who give the certification, and the whole process is supervised by an authority that makes the control. And you have national authorities, like [The country]medic in [The country].

So, I would say, if you talk about public trust, probably the public will have trust in these authorities that oversee the overall surveillance. So, in [The country], probably people would trust [The country]medic. And I don't feel—in the medical field, no one knows how IRM works. I don't know, and I just trust it. I'm not asking the doctor to explain to me how it works. But I trust the authorities who make sure that the whole process is followed and that people can understand.

When we talk about interpretability, I hope as a citizen of [The country] that the authorities ensure that the companies who bring that to the market can explain how it works and then get the certification. So, that's where the law is working, probably—putting a framework on how this process works, what the conditions are to put that on the market, and how there's surveillance by the authorities. That's where it happens. For example, in France, they adopted a law a few years ago, modifying it to make it obligatory for hospitals and doctors to explain that they use AI—an AI device. I don't know exactly how it's phrased in the law, but an evolving AI device. And then it was introduced in the law to allow the patient, for example, to get a second opinion from another doctor or to get some information on how it works. But at the end, it does not work because hospitals just do not inform the people, and the people don't ask. So, it's not really feasible. And I think there's distrust—not trust in the people and the doctor, could be, but more trust in the process and the authorities who oversee the surveillance probably.

00:38:06 Interviewer

And from a patient's perspective, and then talking about consent and being informed, for example, if we're talking about the scenario where a patient goes to a doctor and, for example, their scan is being read with the help of an AI system, how do you think the patients should be informed, in a sense, to give their consent or to give their informed consent?

00:38:38 S6

So, there's two—there's information and consent. It's one thing to be informed and say, okay, usually when you go to the doctor, you will not consent. You will give general consent to be treated and have the surgery and all, but you will not give your consent for each tool being used. Like, for example, using one tool and another. So, it's one tool among many. But it's a different case when you use AI. If you go, for example, for diagnostics—you have a cancer diagnosis, radiology is a good example. For instance, if you have breast cancer and then they find something that might be cancerous, and the AI found it, I think it would be important to inform the patient how they reached this conclusion. And what's important here is to allow the patients to react or maybe go and ask for another opinion to ensure they can think through this situation.

So, I think with AI, it's important to reinforce the duty of information, probably. But probably not an additional obligation of consent specific to AI, you see. Because, at the end, the hospital and the doctor need to treat the patient according to the state of the art. And if AI is the state of the art, they will use it, and they won't just say, okay, a patient decided not to rely on AI, so we'll go with the more dangerous way to treat the patient.

So, I would say there's a distinction here—maybe an additional or reinforced duty to inform about this specific tool, which has an important role in the treatment. But for consent, I'm not sure it would be very different. Unless maybe you use—I mean, this is maybe a bit futuristic—but if you have an AI that would autonomously perform the surgery, maybe at some point when we face these questions, we will probably need the patient's consent. But where we stand now, I see it more as an issue of information rather than consent.

00:41:12 Interviewer

And another thing, maybe we can go back to the example of the patient's data being used to train the AI. How are their data being protected, and if they ever withdraw their consent, what will happen to the AI system?

00:41:29 S6

Yeah, that's exactly the new conflict that I see. So, usually, if they give consent—if I take a research project—you can keep using the data until the end of the project. That's the rule. But you cannot reuse it for another project. That's how it should happen.

But when you enter into this big data or AI training context, it's a bit trickier. In a way, you should be able to remove the data from a dataset if you want to use it for another purpose or project. It's not always possible because you're not able to differentiate between the data at that level. So, in a way, we ask for reasonable efforts to remove the data if you want to use the dataset for something else. But otherwise, the law allows you to use the data until the end of the project, if you have the consent.

00:42:49 Interviewer

And also from a tool-development perspective, what do you think developers and researchers should prioritize to ensure, for example, that patients are going to be informed and things like that? What features of the AI system do you think we should be prioritizing?

00:43:18 S6

In the development phase?

00:43:19 Interviewer

Yes.

00:43:22 S6

Okay, in view of the information of the patient... I don't know...

It depends on the device actually. If the device is supposed to be used directly by the patient or if it's the doctor that will use the device. So, if it's directly used by the patient—well, if you take the European legislation, for example—there's an obligation to ensure transparency and make sure that the people who use it know they are interacting with an

AI. For example, it could be a chatbot or some other interface. It must be designed in a way that the end user knows they're interacting with AI.

Now, if you're working in the radiology field, I see it as less likely that it's directly used by the consumer. But we don't know. The problem is that it should be directed to healthcare professionals, and they need to be informed, for instance, that it's an AI tool. They also need to be trained—maybe on the risks and on what can happen. I

mean, that's somewhat similar to classical devices, but they need to be aware that the device may evolve depending on how it is trained and so on. So, everything specific to the AI device should be conveyed as information to the healthcare professionals who will use it, including, for example, how it was trained and what population it was trained on. If, for example, healthcare professionals know that the tool was trained with a focus on the [The country] population, even if it's difficult to precisely define, they might know that when treating someone from another part of the world, perhaps in an emergency setting, they shouldn't completely rely on certain aspects of the tool. That's part of the risk that should be taken into account and why proper information should be provided.

In the development phase, really, I'm not sure—I mean, maybe that's not exactly my area.

00:44:51 Interviewer

But I also want to ask—you mentioned all these different terms, transparency, interpretability, and explainability. I wonder what these specific terms mean to you, as a legal expert. I know they mean different things to different experts. What does those terms means to you from a legal perspective?

00:45:07 S6

Yeah, for me, transparency is more about what I mentioned earlier. So, when you have a user interacting with an AI, they should know they're interacting with an AI tool. So usually, we use it as well as transparency. And it could be also a deep fake or something else, and you know that's not—it's not the reality, but it was conceived by an AI. It's more transparency.

And then, yeah, you have the interpretability. It's more, for me, it's more the fact that explainability is often difficult to reach. And then legally, I would rather use interpretability as something which we should try to achieve. Because you will use a device in a certain way, to obtain certain results, and you need to make sure the best effort to understand that the answer you get, the results you get, is what you wanted to achieve. Because you know, there might be a difference between where you go, what you can get as information or decision, and what you wanted to—or you think you want to—you think you get this information. So, the explainability is to be sure when you have like this device, making it very clear, as much clear as possible, to explain what it should be used for and how, in a way, you can maybe, that the results you get are in line with what you are expecting or

searching when you try to get a specific decision in the field of health. And I think that's important for the healthcare profession and for the hospitals to be in line with that. And at the end, you know, I don't know if it's so important for doctors or hospitals to know exactly how it works, because we know—even for technicians, people, developers—it's hard to do it.

So, we should know maybe that there's a part that we cannot explain or else. At the end, we need to connect the expectation and the result, and that's what is important in the field of health.

00:47:01 Interviewer

Do you think there are aspects of interpretability that are very important in the legal context, but that people from other fields might overlook or ignore... aspects of interpretability?

00:47:29 S6

It's hard to answer because now, yeah, we have this European legislation, but people are not exactly aligned on what should be done based on these requirements. In [The country], we don't even have these requirements.

So, it's hard to give a precise answer, I'm sorry about that. But I just—and it's maybe should be looked in the light of the risk, or the liability. And if something goes wrong, you see, you need to—you need to show that you acted in line with the state-of-the-art of the medical art. So, it should be defined as something that is commonly agreed in the scientific field, scientific medical field. And we will come to the conclusion that you violated the state-of-the-art of the medical art if you went against what is commonly agreed.

So here you see that if you use a tool that you absolutely ignore how it works and what it gives as results, probably then you would act against the medical state-of-the-art. And this would have to be defined for each specific case, depending on what are the common knowledge, what are the developments in the research field. And also, it's difficult to give you a precise answer here, but what is sure is that if I just use it because it's—I think it's cool, it will—I use a medical new AI device because it gives great results, and I don't know exactly how it works, probably I wouldn't be informed (??) as a doctor.

Not a very precise answer, but I don't have any.

00:49:07 Interviewer

Maybe I'll just wrap up because the time is running out, and I want to ask your visions in maybe 5 or 15 years. How do you think the legal aspect will change to protect, for example, the patients' fundamental rights of them being informed? I think you mentioned

that maybe next year the patients will have access to legal technical documents of all those tools, and I wonder how you think this will evolve in 5 years.

00:49:52 S6

Well, there's no access to technical documents at the beginning of next year. It's just the Federal Council that will announce the big direction where they will go in [The country] with the regulation. And it's not specific to health—it will be for general-purpose AI in [The country].

So, yeah, based on that, it's very difficult to make a prediction (??).

I could try—well, and for the regulation in [The country], I think that based on the Constitution and the federal competences—so what can the federal state do in [The country]? I see it's not very likely that they will adopt a big law like in Europe for [The country], like covering everything, every AI tool in [The country].

Probably not—maybe I'm wrong—but probably not. So, what they will do, they will adopt some sectoral laws, and probably will be for health or for research and have specific rules for using AI. Maybe they will reinforce the medical device regulation specific to, you know, the certification process when you need to make sure that the AI device is safe.

And for the patients, it's hard. We don't even have a national federal law on health in [The country] on patients' rights. So, what we could see here—and I don't see some people are advocating for such a law—but we would need to change the Constitution to do that. So, I don't see it in a very close, yeah, not the following next years.

But what we could start to see, that in the [multiple regions]. So, we have 26 [multiple regions] in [The country]. Maybe in the [multiple regions], in the health regulation, they may introduce some right for the patient—maybe the right to be informed, the right to, I don't know, to have a second opinion if an AI tool is used, for example.

We may start to see these new rights coming, maybe in the [The region] laws here and there in [The country]—maybe not in a very harmonized way. So, that would mainly concern the use in the hospitals and by the doctors, but maybe not the development phase. The development phase would be more in the medical device regulation.

And for the data, as I mentioned, there's this big program for the next 10 years in [The country], DigiSanté, and now they're revising—they're trying to spot all the federal laws in [The country] that might be impacted by sharing of health data and trying to improve this data. There is also maybe a project of a framework law in [The country] to allow better sharing of data in [The country]. I will—I think we will go in a trend where we'll facilitate sharing of data in [The country] in the future. I hope, having better data governance as well, but it will take time for sure, so not be too—so that's a big—it's not, once again, it's not very precise. But it's hard to make a good prediction right now where it will go in [The country].

00:51:09 Interviewer

But we might know next year when they're announcing their plans.

00:51:13 S6

Yeah, in general.

00:51:15 Interviewer

OK.

00:51:17 S6

But not very precise, I fear.

00:51:22 Interviewer

And it will always depend on the specific case, or do you think it's mainly either way going to be related to healthcare? Or do you think all this, for example, to get tools to be approved, is it more focused on the specific tools? Or do you think—oh, I don't know, I'm not really making sense, sorry.

00:51:50 S6

What they would say, the Federal Council next year, it will be very general—some big words. Should we go in the path of the European Union with a general law or not? It will be probably very general.

But in the next years, yeah, they will have to adapt this sectoral law in any cases. For example, the medical device regulation—probably they will adapt it. But the medical device regulation just ensures that you have a safe device.

They're not taking care of discrimination or the patients. That's not—it just ensures something is sufficiently secure on the market.

And then if you go to the data, you have the data part, and maybe here we'll see some modifications as well. And then you have the patients' rights, and maybe it would be more in the [multiple regions] that they would try to improve the position of the patient.

So, going—spreading a bit, depending on the type of regulation. Yeah, you will have some specific regulation for health devices and maybe one day in the Constitution, we'll say that we have some fundamental rights protecting the people against AI tools.

But there is no concrete plan for that now.