AI-GR Pod 36 11.13.25 Laura Zwaan

[00:00:00] Cognitive biases, they often bring us good. Having availability bias, diagnosing a patient with COVID in COVID time is a good thing because it is more likely it's based on expertise. If we would get rid of that expertise part, and this quick recognition part, we throw out the baby with the bathwater because that's the skill that people need.

But of course, we need to keep learning, and we need to keep making sure doctors are well trained and, and provide, get feedback and get corrected when they're off. And, um, that's something that AI can do. It's also something that should be done more in the workplace right now by providing each other feedback and looking back at what ultimately a diagnosis was.

Hi, and welcome to another episode of *NEJM AI Grand Rounds*. I'm Raj Manrai and I'm here with my co-host Andy Beam. And today we are delighted to bring you our conversation with Dr. Laura Zwaan. Laura is an Associate Professor at Erasmus University in the Netherlands. [00:01:00] She's an expert in cognitive psychology and epidemiology.

And, Andy, I think this was a really fun conversation about not just human psychology and decision making, but robopsychology or machine psychology.

Yeah. Raj, I've been a long time, like, cognitive psychology nerd. A lot of my early interest in AI was from folks like Doug Hofstadter and like really amazing cognitive science folks.

And so, this was just a super interesting conversation that combined so many fascinating topics. How do people make errors, such a really important topic in medicine. And now that's generalizing to how do machines make errors. It turns out, as listeners will hear in this episode, they've inherited some of our cognitive biases, which is, again, just so fascinating to me when you think about the robotic picture of AI that many of us had as early as 10 years ago.

So really, really enjoyed this conversation. Laura was amazing and really helped us better understand error-making from both humans and machines.

I totally agree. I think you can see the expertise and sort of the thought that she's put into these really [00:02:00] nuanced questions about how to define errors and all the biases that enter into our labels, around what is an error and how hindsight is 20/20 and defining an error and therefore framing our ground

truth. This was an amazing conversation and with that, we are delighted to bring you our conversation with Dr. Laura Zwaan.

The *NEJM AI Grand Rounds* podcast is brought to you by Microsoft, Viz.ai, Lyric, and Elevance Health. We thank them for their support.

Laura, welcome to *AI Grand Rounds*. We're super excited to have you on the podcast today.

Well, thank you for having me. I'm very happy to be here today.

Laura, welcome to *AI Grand Rounds*. So, this is a question that we always get started with. Could you please tell us about the training procedure for your own neural network?

How did you get interested in artificial intelligence? And what data and experiences led you to where you are today?

Yeah, thank you for that question. I think I will [00:03:00] answer this question very differently from the usual people in your podcast, because when I was in high school, I had no idea what I wanted to study.

But I knew one thing, it was not gonna be psychology or medicine. Which is interesting because I end up exactly at the crossroads of those two specialties. But to allow me to make that decision, I went abroad for a year after high school I went to England and Spain to do some language courses, and I worked in a Spanish hotel in the reception for a bit.

And I enjoyed it very much. So, I was thinking maybe I should work in tourism. But I got some really good advice there from my supervisor in Spain who told me, Laura, if you want to work in tourism, you can always do that because you speak your languages, you can always pick up on that. Go learn something real first.

So, I had to call my parents who had to change my application from applied sciences to psychology because I got interested in psychology during that year, and I decided to go for psychology [00:04:00] after all. And my first course was cognitive psychology and to learn about how people make decisions, how does memory work? How do we allocate attention to certain information while disregarding it for other information that comes to us was really interesting to me right away.

So, I was hooked and I really loved that topic. So that got me into psychology, but still not in medicine or AI. Medicine came to me after I did my master's in cognitive psychology. I wanted to do something that had an application, so not the fundamental cognitive psychology work. That's when there was a Ph.D. position to study medical errors and the cognitive aspects to it.

And I applied for that position and got hired and started to work on medical errors and in particularly quickly moved to diagnostic errors. And the diagnostic error field was very new at that time. There was very little that had been done in that domain. So, I had the opportunity to contribute to the [00:05:00] development of the fields in both the research and to the community building.

Well, so then we have the medicine part, but not the AI yet. The AI is very recent for me. I always say I'm really not an expert in AI. Because it came to me quite recently when we did a lot of the studies and we found that diagnostic errors are likely to occur because of knowledge deficits. And a lot of the more general interventions, such as de-biasing strategies, didn't really help much.

It needed to be content specific. That's what AI comes in, that's what AI can do. So, that's how I got interested in AI to see what can AI do to help reduce this diagnostic errors. So yeah, that's how I ended up there, or here.

That's great. And actually, I think that's a good framing for what we want to dig into next.

So, we really want to go through some of your recent work together and also some of your, I think, seminal earlier work. And just as you said, I think what strikes me in reading your papers is that [00:06:00] there's really a lot of thought and focus on how to define what a diagnostic error is, what a medical error is.

And I think maybe you can walk us through maybe the first one of these papers, just title the challenges, and I think it actually tells a story there, "The challenges in defining and measuring diagnostic error." So, of course doctors disagree, right? Clinicians disagree about what is a medical error. And I think there are often reasonable, divergent perspectives on whether even an error occurred.

And so, maybe you can take us through the general problem in a little more detail. What has been the challenge in this field? And then, maybe, specifically also what that contribution was in that paper and what we should take away from your paper.

Yeah, sure. That paper was written quite early on in the field of diagnostic safety when there hadn't been much thinking about the concept of diagnostic error. It was often considered a diagnosis is right or wrong. Patients get harmed, but it's much more complex than that. [00:07:00] And we organized it one of the earlier diagnostic error in medicine conference an expert meeting to discuss this topic.

And there were several things that came out. And one of them that I think is very important to realize for people who are not in medicine, like I was also new. I didn't really know how it worked very well, but that its diagnosis is evolving over time. So, it really depends what moment you look at a specific case, whether something is a diagnostic error or not.

And I often describe it in a way that when I go to the doctor and I think I might have a pneumonia, I'm coughing a lot and I go to my GP and does some tests and say, no, you don't. And then I come back next week and then I do. Often people would automatically say, oh, this is a diagnostic error because this was not found a week before.

But the real question is, was it already there? Was it the pneumonia at that time? And could you have known that it would develop into that? And that's something that is not often taken into account in assessing diagnostic errors, and it's something very important to [00:08:00] realize. Another thing that is really challenging in the field of diagnostic error is to identify, in hindsight, I mean, we always determine the diagnostic error in hindsight, whether a diagnostic error happened, and that leads to hindsight bias.

Once you know the outcome, it's very easy to say that it was an error. So that results in, in—.

But you almost have to put yourself in the mindset of the clinician who is actually going through it without the benefit of that final diagnosis, right? Yes.

And that's very—.

Can you rid yourself of that hindsight bias in defining a diagnostic error?

No. You cannot do that in methods that are commonly used to study diagnostic errors that are based on record reviews or identification of an error afterwards, or of harm afterwards, or a mismatch in label afterwards. And then it's really difficult to determine was there an error or not. And we did another study

actually, in which we studied the effect of this, and we gave people clinical cases and we asked them to identify process errors.

And what the people didn't [00:09:00] know was that we changed the ending for half of the people. So, for some people it ended— Wow. —in a diagnostic error saying you thought it was pneumonia, but it turned out to be a pulmonary embolism. And for the other people we said, oh yeah, it was right. The patient got better.

It was a pneumonia. And then we saw how many cognitive biases people assigned, how many process errors, people assigned to that specific case. And we found that they've assigned many more if the outcome was bad compared to when the outcome was right. So, it's very outcome based and yeah, that is a challenge in measuring.

So, we need to have a good combination of also studying things in the lab more prospectively. Comparing cases that go well with cases that go wrong, not only looking at cases that go wrong, and really be aware that this hindsight bias plays a role, but yeah, very difficult.

Can you talk about the adjudication of diagnostic errors and where the field is today? And what is the state of the art for defining if resources aren't an issue? If [00:10:00] you have unlimited access to clinician time, which is, of course scarce, very valuable, but you have unlimited access to expertise. What is the gold standard today, in 2025, for taking, let's say again, going retrospectively now, you know, forgetting the prospect of studies for a second, but going retrospectively and defining whether or not a diagnostic error occurred.

Is it some combination of kind of consensus, individual votings plus consensus across the physicians where they discuss it? What is the gold standard? Is there a formal framework that, I know there are a few for defining and thinking about the taxonomy of errors, but how do you define a diagnostic error if resources aren't an issue?

We still have a challenge in defining diagnostic errors. The field has moved much more towards improving diagnostic safety as opposed to identifying diagnostic errors. Hmm. But of course, it's still worthwhile to see when an error occurred or not. I think for sure what you need is, is [00:11:00] multiple methods.

Just one type of method doesn't work. We need to have triangulation. Recently with one of my Ph.D. students, we did a study where we also, we looked at

patient records, we included patients prospectively, and then we interviewed clinicians about their decisions. That's already getting a little bit better because we've seen that a lot of the errors actually turned out not to be errors because there was some information available elsewhere or the clinician could clarify.

So, I think that is a very nice way to improve the measurement. Very costly and time consuming. There's not really a gold standard. It stays messy for, for the purpose of, we have the disciplinary board in the Netherlands, and they decide if a doctor made an error or not. And also there, it's quite biased.

There are experts, often actually from a slightly different domain, who are evaluating if an error was made or not. What I would really like to propose for a purpose like that is to have a case description without the [00:12:00] ending, have a lot of experts review it, and then see if there's a majority of the people who would say, I would do this differently. I think this is not appropriate care.

And it should be mixed, not only with cases that were errors, but also with cases that were correct, so people can actually make the decision based on the process rather than on the outcome. Right. That would be my ideal, I guess. I think you, you said something there that I'd like to follow up on.

So, like, often when we think about medical errors, we imagine one doctor sitting down and they're the sole decision maker and the primary point of care for a patient. But medicine is much more of a team sport, where there will be nurses, often there will be a physician who's the continuity physician who provides the bulk of the care, but on the weekend, someone else may come and provide weekend coverage.

And this is gonna be increasingly important as AI is part of the team. But how, when we're thinking about medical errors, is fault assigned. You know, that's, like, kind of a crass way to phrase it, but how do we [00:13:00] know who made the error and who's therefore should be held responsibility in these complex team-based health care scenarios.

Yeah, that's very challenging. It's almost never one person isolated making an error. It's the system, it's the team, it's the collaboration, and that all plays a role. I would really not like to assign any blame because I think it's very complex process that is really, really difficult. I understand sometimes that there need to be regulations and testing if something is up to the standards, but for me in studying it, I would very much like to focus on improving and see how we can make care better.

And of course, the teamwork is crucial in that. That's also something that patients really notice and that patients really want research efforts to focus on. On the continuity of care and transfers of care, because they also see and feel that that's where a lot of errors occur.

Yeah, I do wonder, I mean, so I'm married to a physician, and I think the assignment of [00:14:00] blame for medical errors results in a lot of defensive medicine being practiced, and it actually changes the way that a lot of doctors interact with their patients. I just wonder how AI is gonna interact with sort of that dynamic.

If they're gonna be less likely, and we'll, we'll dig into this a little bit more later, but yeah, I totally agree with you that focusing on improvement, but it does impact the doctor's psychology about if they document something in the wrong way, then during a disclosure process or something like that, it could be used against them.

Yeah. So, maybe, going into more constructive or forward-looking, improving safety as opposed to assigning blame direction here. There are these numbers that are quoted. I think there was this landmark report that was called, like, "To err is human" right around 2000.

Then there's been various updates to that estimate both in the United States and outside of the United States for how many deaths and other morbidity that's, that are due to medical errors annually. What is your sense, as definitely a leader in this field, in defining [00:15:00] diagnostic error as to sort of maybe just the direction in which we are moving? Do you think diagnostic errors are reducing? Are we getting safer over time?

Since 2000? Since 2010? Since 2020? Are we heading in the right direction? What would be the story you would tell from where we were with that landmark "To err is human" report to 2025?

I think we are making progress, but I think it's very slow. We did a study in 2005 looking at adverse events and diagnostic adverse events where one of the group, and we see that if you look at trends over time with the same method, it doesn't really reduce much. It reduced initially a little bit in the Netherlands at least. Yeah, but it's not really reducing much more recently. It's a difficult thing to state because these are really small numbers and it can be really influenced by hindsight bias, by who the reviewers are.

But I think we have made progress. I think we are much more [00:16:00] aware of it. You see the culture changing, so I think it's better. Really difficult to measure though. There's also still a lot of room for improvements because there's still a lot of errors that happen and occur, and I think, yeah, it deserves attention still.

Yeah, it's in some sense, the question I ask is even harder because I'm asking you about the direction in which we're going, and it's even arguably reasonably hard to discount period, right? Let alone understand the direction over time with various studies, with different ways of reporting or defining diagnostic errors and so on. Andy, do you want to talk about the next topic here?

Yeah. Actually, I was gonna ask a question about the "To err is human" paper, and it's led to this catchy headline that medical errors are one of the leading causes of death in the hospital. I guess, outside of exactly the number of mortality attributable to error making from a gut check perspective, does that feel approximately correct to you or is that over counting or undercounting?

I know that's a little bit of a spicy [00:17:00] question, but it's always, whenever you hear about that study, that's one of the things that is quoted.

Yeah, I think it's a bit over counted. I think hindsight bias plays a major role there. What do you call a diagnostic error? And also, another big difficulty is to determine if the harm was actually caused by the diagnostic error.

And that's something that is almost impossible, I would say, to assess. There are some methods that are exploring this, but for example, if a cancer diagnosis is diagnosed three months late, it might be very clearly because of an error that it could have been diagnosed earlier. But what difference is this three months gonna make in terms of the patient outcome?

If the patient dies, ultimately, would the patient not have died if this was found three months earlier? Yeah, this is a really challenging topic, so I think I would like to stay away a bit from those numbers and say, thinking about what the consequences are. I think it's a serious enough topic to study [00:18:00] and to improve, but I would not say that it's the third leading cause of death in hospitals. I think that's probably a bit of an overestimation.

Thanks. So, now we're gonna transition from human error making to machine error making, and you're a cognitive psychologist. I think that's such a fascinating field. I kind of think of what we're entering now as the era of, like, machine psychology. We're trying to understand the ways in which machines

make decisions, but they are very black box and opaque, and we can ask them questions and interrogate them.

So, I'd like if you'll entertain me, like to sort of start diving into the world of machine psychology through the lens of a cognitive psychologist. So, what has studying human error making taught you about the ways in which machines make errors?

I think the interesting thing, and I, I would say I don't really know that much about how machines make errors, but there seem to be quite similar mistakes.

They both show errors in line with cognitive biases. They both [00:19:00] sometimes go into the completely wrong direction. They have different strengths and limitations. The biggest strengths of humans is that they're more flexible and they can take the context into account and adapt quickly in very complex settings but are limited in the amount of information that can be processed.

And because of that, they rely on experience and take some shortcuts, sometimes often, that are actually often based on a lack of content, specific knowledge. And that can lead to biases and errors. AI actually can process a lot of information. That's its strength, the calculation power, but of course it's biased by the information that they're trained on.

They show similar errors in terms of outcomes, but the reason is different. The causes are different, I think.

Mm-hmm. Could you give us some examples from your own work about how cognitive biases have shown up in machines? 'Cause I'm imagining a listener might be surprised to hear that an AI has anthropomorphic cognitive biases. [00:20:00]

There was a study that wasn't mine, but I think it was Lieder and Maier's study. And they have checked whether a large number of well-known cognitive biases in humans would also be present in large language models, and they found that most of them were actually more strongly even present in large language models than in humans.

An example would be the framing effect. So, if you frame an outcome of a treatment in this many patients would die. People evaluated very differently from saying this many people or this percentage of the people will live. 90% will survive the surgery, makes the surgery sound very attractive.

While 10% will die during the surgery, will make it sound less attractive. And humans are influenced by this framing in terms of what decision they make. And actually, LLMs were similar in the ways that they would make decisions, and they were also influenced by framing, for example.

That's interesting. A known [00:21:00] problem with large language models is this problem called sycophancy. And because they are trained to be chatbots, you know, pleasure, uh, chatbots—. Sometimes the problem with humans as well.

Exactly. So, people pleasers are the human equivalent of this. And so, they're actually trained to kind of agree with you. And I guess that the evidence that you're pointing to there shows that how you frame the problem may induce specific kind of biases, maybe because the underlying model is trained to be pleasant and agreeable.

Yeah, I think that's probably true. I think this is more pleasant and agreeable. I think this error is probably because the AI is based on how humans would respond. It's quite a well-known bias, so it'll probably be like that reflected in the data.

But what you mentioned that the AI is trying to please, of course, is also a problem when you think about confirmation bias, trying to anchor on information that is provided, and that's something new indeed, where we have to look into the machine psychology and see if they make similar decisions.

[00:22:00] I have a follow up question on the framing.

So, the framing point is fascinating. There's actually this *New England Journal of Medicine* paper, I think back in the 1980s, the early 1980s, by Barbara McNeil and Amos Tversky and a few others that explored exactly the same scenario and a few others on framing a decision in terms of probability of living versus probability of dying after the procedure. And so effectively, logically the same outcomes and the same probabilities attached to those outcomes but framed completely differently.

And you got very different answers from respondents about what they would prefer as a function of how the same probability was framed against the sort of good outcome versus the bad outcome. And what's fascinating is then starts me pulling on this thread of like Tversky and Kahneman and all their work in the sixties, seventies, I guess really the seventies, right?

And this seminal work in psychology and in studying humans and how humans both use heuristics and [00:23:00] biases for making sometimes quick, good judgements, and also sometimes really leading them astray. Do you think we're seeing now, and I think that, uh, Lieder-Maier paper is a good example of it and it's maybe an existence proof of sorts, but do you think we're gonna see a sort of revival of that work of studying humans from the Kahneman-Tversky era, but now applied to machines, uh, or should we be thinking about this a little bit differently than just sort of porting the Kahneman-Tversky work to robots?

I think it would be very interesting to study this in machines and see how it differs from what commonly humans, how humans would make systematic errors in cogniti, like, cognitive biases and how machines would do that.

I do think, ultimately, this is not the way to go completely, because this is all based on vignettes. And all of the relevant information is available. And cognitive biases are also always only determined in hindsight if the process went really well and it's based on expertise of the [00:24:00] physician. If they get the right diagnosis, they say, well, this is a really good display of expertise.

If they get the wrong diagnosis, they say, oh, it's a cognitive bias. They took a shortcut. And in such vignettes, it's always possible to find the right answer. I mean, not in the framing one, but for example, with availability bias, it's often trying to lead people astray to make this mistake. And it is insightful to see how physicians, but also machines make these decisions.

But ultimately, to improve in practice, we need to see what it does in practice in the complex and and fuzzy environment because we see it doesn't work there in the same way we did this study in Switzerland. It was led by Wolf Hautz. I dunno if you know it, but we tested Isabel Healthcare. It's a decision support tool for diagnosis.

And in experimental studies it had been showing some positive effects. And then we tested it in real clinical trials in four emergency departments in Switzerland, and then we did not [00:25:00] find an effect. There's a lot of other things at stake and not only the case description. So, yes, we need to study how these decisions are being made.

Yes, we need to understand how humans have different biases and different systematic errors from machines because it can help us provide insights in how we can improve the human-computer interaction and the collaboration. But we need to also test that in actual clinical practice.

Again, I find this topic just endlessly fascinating. If you think about the mental picture we had of what AI would be 10 years ago, it was this perfectly rational Dr. Spock-like entity. And what we ended up with was something that functions completely differently than human cognition but has nonetheless inherited many of the same biases that people have.

Your point is well taken that we need to think about them in very different ways, but it's still what a fascinating place we've ended up at in that we have a shared set of biases between man and machine. [00:26:00] One forward-looking question I'd love to ask you is you've done amazing work on the consequences and definitions of human error making.

We've talked a little bit now about how machines make errors as you're looking forward the next five or 10 years, which set of error makers, the consequences of those give you the most pause? Like, are you more worried about humans making errors or the consequences of machine-made errors?

Yeah, that's a good question.

I think based on individual errors the consequences would be similar. But if you look at it at scale, probably AI is a bit more at risk because of the errors are a bit more systematic than they are in humans. Individuals make more, I think, different mistakes if we look at, yeah, so the scale of the harm that can end up in practice, I think AI is a bit more at risk.

We also probably need to see it how we can detect errors. I'm not sure if that's easier with AI or with humans. Humans still [00:27:00] have a lot of feedback opportunities, but not as many as we would want. Um, and then the third one is improving. If we detect the errors, can we actually make humans or AI do better?

I think in that sense, probably AI wins because we can retrain much easier than how we can train humans because humans are pretty hardwired in those biases, and, but ultimately, we need to of course have the collaboration and see how they can cross-check each other.

Yeah, that was gonna be my next question, which system is easier to correct?

Like you said, we can get a new pile of data or get some corrections to machine errors and kind of like reba the model and get something. It does end up being a little bit of game of Whack-a-Mole, and that like, there's always gonna be

something that doesn't work. But correcting human cognitive biases feels significantly more challenging.

In some sense, that's what the whole field of psychology has been thinking about for a long time. I think if I heard you correctly, it sounds like that you would agree with [00:28:00] that. Yes.

Yes, I would agree with that. Cognitive biases, we are not gonna get rid of them, but also because they often bring us good.

Having availability bias, diagnosing a patient with COVID in COVID time is a good thing because it is more likely, it's based on expertise. If we would get rid of that expertise part, and this quick recognition part, we throw out the baby with the bath water because that's the skill that people need. But of course, we need to keep learning, and we need to keep making sure doctors are well trained and, and provide, get feedback and get corrected when they're off.

And that's something that AI can do. It's also something that should be done more in the workplace right now by providing each other feedback and looking back at what ultimately a diagnosis was.

I have one last question that combines the two topics that we've been discussing here. So, would you be now thinking about using machines to define diagnostic errors from health record data?

[00:29:00] Would you be, I think people are already doing this, right, they're using LLMs to review electronic health records, see if they can spot errors? I think it's still pretty early in the work here in the sort of subfield, but would you suspect that LLMs are subject to the same sort of hindsight bias that humans are in defining diagnostic errors?

So, giving an LLM the final result, will it be harsher on the same exact process if it knows that it led to a bad outcome versus a good outcome?

Probably, yeah, I think, I think probably. Yeah. I think so, too. I think so, too. I think so, too. I think, yeah. We should test it though. Yes, but I, I think so, too. It would, it would be actually a very nice machine psychology study to conduct.

Excellent. *AI Grand Rounds*: we come up with research projects here on the, on the fly. Yes. This is very good. Alright, should we, uh, Andy, should we move to the lightning round? Yeah, let's move to the lightning round. [00:30:00]

So, just a quick recap. These are gonna be short, rapid-fire questions and you can respond, but hopefully briefly, so that we can get through as many as possible. So, the first lightning round question, I know you're not a clinical psychologist, but I still would love your answer to this one. Will large language models be net positive for mental health over the next five years?

Yes. Yeah, I think so. I think it can really help. I think it's, you already see a lot of people are using it that way. So, I think, I think it will.

Awesome. Alright, here's the next one. Will AI be used routinely as a second opinion in clinical practice in the next five years? Yes, I think so. I also think it's already heading in that direction.

I know of a lot of people who are already using it, and I think it will be used even more, and hopefully we can get better insights in the risks and the benefits of it to do this well.

Awesome. So, our first non-LLM lightning round question. What was your first [00:31:00] job?

Oh, oh, a very Dutch one. I was like peeling tulip bulbs because—

Nice.

—I'm from a family that my grandparents had a tulip farm, so I was very, I, I think I was nine years old when I was helping there. Yeah.

That is a very Dutch answer. Yes. Amazing. Alright, our next one. What is your favorite novel? Oh, I'm rereading Pride and Prejudice right now, which I is one of my favorites.

Yeah. Yes.

Excellent. If you could have dinner with one person, alive or dead, who would it be?

Yeah. So, many people that pop up when I'm thinking about that, but probably my grandparents, who are deceased, they would be so great to see them again. Very personal answer, but yeah, I would like that the best.

Alright. Alright. And here is our last one and just as a disclosure, this was generated [00:32:00] by GPT-5. Uh. What's the biggest cognitive bias clinicians should watch for when working with AI recommendations?

Uh, I think confirmation bias. So, I think indeed the tools will tend to agree with you and you have to be very careful not to also get very happy and excited about that. But be, be critical. Be critical of the answers that you get and try to get an alternative explanation rather than one confirms.

So, I'm gonna go off script here and just maybe ask a follow-on question 'cause I think your answer's so interesting. I think confirmation bias is the big risk for all of us in medicine, outside of medicine, with these tools. And it relates to what Andy said, too, this sort of sycophancy that's baked in, they're designed to be helpful and they're designed to be in a way things that you want to keep coming back to using and what do you want to keep coming back to using something that agrees with you and sort of confirms your beliefs.

Right. And so, does that [00:33:00] inform, from your perspective, does it inform how you would test or deploy AI in actual clinical workflow? Like, is confirmation bias sort of, you know, if the doctor or the clinician knows what they want to seek out a second opinion for? That's sort of a radically different paradigm than just having a second opinion triggered outside of the clinicians by their own volition, right?

One that's triggered based on either a note that they're writing or one that's triggered based on sort of observation of the clinician-patient interaction or something like that. Do you think, therefore, it's important to test these tools outside of just sort of a button or the desire, essentially the clinician initiating the interaction with the AI second opinion?

Yeah. I think we should really make sure that AI, when it's used, it tries to provide alternative explanations. And we've seen in the Swiss study, I think the qualitative part, some first [00:34:00] result that I understood that actually a lot of the residents were using it to confirm their diagnosis before to go going to their supervisor.

Wow. And I think that's something that we need to prevent. So, we need to build in, in the human-computer interaction, the way the interface works to see if we can try to refute it, but also not only use it for the ultimate diagnosis, also think about next steps. What do I need to do to test if this might be something else?

So, we should focus on trying to making it disagree with us.

Amazing. Alright, so, uh, Laura, you did fantastic at the lightning round. We actually, we just have a few final kind of big picture last questions for you to close out. So, the first is, what are your predictions? We're gonna ask you to speculate here.

What are your predictions about fundamental shifts, if any? So, you know, an answer is medicine is slow. There are not gonna be any, but what are your predictions about fundamental ships, of fundamental shifts, if any, in [00:35:00] the nature of human computer interaction in the next one, five and 10 years?

Yeah, I think, it is slow. So, it will be a process before human computer interaction will be really adopted in medicine in the way that it's integrated in the workflow. So, that might be the 10-year time prediction. I think, I hope by then it'll be integrated in the clinical workflow, but before we get there, it'll first be tried out, hopefully in, in a year or so, that there's more and more use of LLMs in clinical practice to improve the diagnostic process.

And I expect that it'll be a bit, doctors still being overconfident, maybe ignoring the advice, but that's a process. It takes time to build trust, build calibration, knowing when to trust it, when not to trust it. And that might be more up to speed, say in the five-year prediction.

People know that they can rely on it sometimes, but may still be critical. And then after that, we'll have to watch [00:36:00] out for automation bias and trusting it too much. But if it's integrated well in the right way, hopefully well tested and evidence-based, it'll be used in practice. Yeah. As a team, I would hope the clinicians working with the AI as a team, discussing and collaborating a case, collaborating on a case.

Cool. So, final question, and this is a little bit more abstract. So, I'm always fascinated when new kinds of technology give rise to new kinds of scientific inquiry. We had Eric Horvitz from Microsoft on the podcast a while back and he made this point that some of the stuff happening in mechanistic interpretability, which is how do we understand how information flows through a large language model, kind of as an analog to neuroscience and neurobiology.

They have this very complicated neural network and they're trying to sort of reverse engineer the principles that dictate how it operates. And, I guess, I'd love your take on how seriously we should take the [00:37:00] sort of machine psychology idea.

Like, plausibly, hundreds of billions of people are already affected by machine decisions and therefore machine errors. Is this a new scientific field that we should be thinking about seriously? Do you encourage your graduate students to start thinking about machine error making and machine decision making or just, um, yeah.

How does this, how does this look from your perspective?

Yeah, I, I think it's already starting quite a bit to study machine decision making in a similar way as we do in psychology indeed, to sort of identify the black box decisions that are being made and see if we can find systematic deviations. I think it's a good way to study this and to see how the machine psychology is different from the clinician psychology because that is what's gonna help us improve the collaboration.

So, I would definitely encourage people to study machine psychology, and I don't think we need to know the details of how the process works, but we need to know how it [00:38:00] behaves. So, the psychology of it.

Cool. Awesome, Laura. Well, thank you so much for coming on the podcast today. It was, it was a great conversation.

And again, just like such a super fascinating topic.

Well, thank you so much for having me. This was really fun.

Thank you, Laura. That was amazing.