AI-GR Podcast 14 01.12.24 Roxana

[00:00:00] I mean, it's not like the models magically learned bias by themselves. These models learn from human data. And so, the biases in human data get picked up and regurgitated with large language models. People will say like, well, humans are biased and, therefore, the models are going to be biased or that's okay. And I actually don't think that's okay.

[00:00:27] I mean, it's not like the models magically learned bias by themselves. These models learn from human data. And so, the biases in human data get picked up and regurgitated with large language models. People will say like, well, humans are biased and, therefore, the models are going to be biased or that's okay. And I actually don't think that's okay.

[00:00:50] Technology is not going to save everything.

[00:00:58] Welcome to another [00:01:00] episode of *NEJM AI Grand Rounds*. I'm Raj Manrai, and I'm here with my co-host, Andy Beam. And today, we bring you our conversation with Roxana Daneshjou. Roxana is an assistant professor of biomedical data science at Stanford University, and she's also a practicing dermatologist. She's been working in AI for dermatology and for medicine more broadly for several years.

[00:01:22] Andy, I think this marks the beginning of season two and Roxana is a wonderful guest to kick off our new season. So, to me, I think Roxana is a clinician scientist who truly and impressively draws and mixes together both her clinical and her technical skills to do unique work at the intersection of AI and dermatology.

[00:01:42] She also cares a lot about algorithmic bias and health care disparities. We really got into that in the conversation. I thought she had a lot of insights to share there about both the potential for AI to exacerbate disparities, but also how it might be used as a tool for good. You really feel, I think, the urgency of [00:02:00] these problems when speaking with Roxana, and I learned a lot from our conversation with her.

[00:02:04] Yeah, Raj, I agree. It was a great pleasure to talk to Roxana. I first became aware of Roxana, you know, like I do of a lot of scientists through Twitter, where she's prolific. The first time I met her was at NeurIPS and I was very pleasantly surprised to find that she's as nice in person as she is on Twitter.

[00:02:22] I agree that her work on algorithmic bias is timely and urgent. Um, and when I think of Roxana, I really think of her as an exemplar for future leaders who are clinicians who are really pushing the boundaries for AI and medicine. I often get approached by junior clinicians saying, how can I mold myself to have a career and really make an impact at this interface of AI and medicine?

[00:02:46] And almost always I point to Roxana as the template to do that. She's clearly a rising star in the area. And I think she is uniquely good at combining her deep clinical expertise to inform her research directions in [00:03:00] AI. And so, for a lot of reasons, it was really great to have her on the podcast. The *NEJM AI Grand Rounds* podcast is brought to you by Microsoft, Viz.Ai, Lyric, and Elevance Health. We thank them for their support. And now we bring you our conversation with Roxana Daneshjou.

[00:03:24] Welcome to *AI Grand Rounds*, Roxana. We're really excited to have you here today. Thanks for having me. Roxana, this is a question that we always like to get started with. Could you 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?

[00:03:45] Well, I guess, how far back do we want to go? I mean, um. As, as far back as-- We like, we like to go back to when the neural net was initialized. The pre-training that happened, basically I grew up to, uh, [00:04:00] to Iranian immigrant parents who really liked to incorporate science very early on. I have memories of doing these long road trips with my dad and mom and my dad would always play these astrophysics recordings.

[00:04:16] And so I had a, you know, from a young age, just developed this curiosity for how things work and how to build things. And I actually, of my own volition, I wanted to go to essentially nerd school. I went to a high school that was the Texas Academy of Mathematics and Science, and it's just as nerdy as it sounds.

[00:04:36] High school students get to live on a college campus and take college classes together, and it's a boarding school. And that was actually when I first got into research, because I joined a neuroscience research lab with Dr. Jannon Fuchs and was doing a lot of mouse work at that time. And then I went to Rice and I studied bioengineering because I was [00:05:00] very interested in understanding how the human body works, but also how to build things that could improve the human health condition.

[00:05:12] And I think that that was a really formative experience for me because as a bioengineer we talked a lot about design thinking, we talked a lot about identifying problems, and then developing solutions and testing. And I think that formed a real foundation. And to be honest, in college, I waffled a lot about whether I wanted to get an M.D. or a Ph.D.

[00:05:37] I waffled back and forth many times. I ended up actually applying to go to medical school and get an M.D. And I landed at Stanford, which has a heavy research focus. And I took, I was taking a bioengineering seminar with Dr. Russ Altman, and he talked about using computational methods to study [00:06:00] human genomics.

[00:06:01] And that really sparked my interest. I was a first-year medical student. I emailed him. I met with him. I started doing research with him. And then at some point, I realized two years into that and then after having done a one-year research fellowship that I wanted to do a Ph.D. So, I actually didn't go into medical school as an M.D. Ph.D.

[00:06:21] I was an M.D. only who suddenly had struggled with that decision before coming and had just picked one and then decided to add a Ph.D. in the middle of medical school. So, I did that with Russ and I completed my medical school. And right about that time, there was all the interesting work coming out around computer vision.

[00:06:44] And I had become clinically interested in dermatology. And so, to me, it seemed like a real opportunity. AI seemed like a real opportunity to solve problems in dermatology, which is a very visual field and all of health care in general. [00:07:00] So, I ended up doing a residency in dermatology and then a postdoc with Dr.

[00:07:06] James Zou in health care AI. So, I was in training for a long time, as you can tell from that story. Can I just follow up? What was it about dermatology that sort of drew you to that? Well, I think, I think there was a couple. So, on clinical rotations, I really enjoyed basically every rotation I went through.

[00:07:26] The one thing that's really nice about dermatology is that it's a very visual field. You can walk into the room and see what's going on a person's skin and put the pattern together with some history. Sometimes you don't even need the history. Sometimes you walk into the room and you can just tell exactly what's going on.

[00:07:49] And actually, one of my favorite diagnoses is this phytophotodermatitis, which is they get geometric brown or red patches on their hands and they've suddenly [00:08:00] appeared. And it turns out that this diagnosis is usually after handling certain plant-based products and then being exposed to the sun. And, so, you start to ask them questions.

[00:08:13] So, lime juice is one of the things that can do this. So, you ask them questions like, oh, so this past weekend, did you hang out with your friends? It's actually called margarita dermatitis. Did you handle limes because you were making margaritas or, you know, having a barbecue or doing something else and their eyes will just widen because that's exactly what had happened.

[00:08:35] And so, there are other diagnoses in dermatology like that as well, like when you see the pattern, it corresponds to some exposure. So, it's pretty fun to like be able to make diagnoses like this. I think that's part of it. I think also I build, I have really great longitudinal relationships with my patients on an outpatient basis

[00:08:56] So, I see them regularly. I know their stories. I know all about their [00:09:00] families and what their travels are and so I enjoyed also like the aspect of being able to build these really long-term relationships with the patients that I take care of. That's a great overview of your training. What are you up to now?

[00:09:15] Yeah, so I am an assistant professor of biomedical data science at Stanford and also joint with dermatology. So, I practice half a day a week seeing patients and then I run my AI for health care lab the rest of the time. Which is essentially like the physician scientist dream, is to be able to still have your hand in clinical work, because I think, one, it's very gratifying to work with patients, and two, it gives you a window into exactly what's happening into the health care system.

[00:09:48] And the rest of the time I get to work with my amazing, the postdocs, the graduate students, and medical students who are in the lab. Awesome. I think that is a good transition to talking about some of your research. [00:10:00] So the first paper that I want to talk about is called "Disparities in dermatology AI performance on a diverse, curated clinical image dataset."

[00:10:08] So, one, could you give us the setup for this paper? And then I think we'd like to dive into some of the details. The setup for this paper is that

[00:10:16] I'm interested in AI fairness and bias. And Dr. Ade Adamson, who is a friend and colleague, had written a perspective piece in *JAMA Dermatology*, essentially saying, early on, that he was very concerned about whether or not datasets that were being used to train dermatology AI models were representative of the full skin tone spectrum.

[00:10:43] And I think that his concerns were really appropriate because as many other dermatologists had pointed out before, like Dr. Jenna Lester, the education in dermatology, like the textbooks, the [00:11:00] training, in and of itself is not very representative of diverse skin tones. So, you're looking at the training being like that for humans, and then you start to worry about the training for models having the same problem.

[00:11:15] And before we actually wrote this paper, we wrote a different paper where we actually went through and looked the literature that had been published in the dermatology AI space and found that almost none of those papers actually even reported what skin tones were used in training or testing, and that for the ones that did report that, they basically either significantly underrepresented or excluded

[00:11:43] Brown and Black skin tones. And, so, the impetus for this paper was we wanted to create a dataset, sort of like a benchmark, so that we could test how algorithms perform across diverse skin tones. [00:12:00] That's great. So, could you actually, for the non-dermatologists who are listening, like, how do you measure skin tone?

[00:12:06] There's a scale. Could you walk us through actually how you assess variety in skin tone versus just like dark versus light? Right. So, there is a scale. It's actually not a great scale, and there has been a lot of discussion and efforts on developing new scales because the current scale is not very inclusive.

[00:12:27] However, much of the machine learning world has been using that imperfect scale, and that's actually what we do in our paper as well. But I just wanted to put that out there, I think that scale is problematic. It's called the Fitzpatrick skin tone scale. It was originally developed to help dermatologists look at how easily someone might burn,

[00:12:51] sunburned to in order to help people dose photochemotherapy. So, it was originally about how easily you tanned, how easily you burned. [00:13:00] And then it got co-opted into being used to assign color from an image. So that's not how it's meant to be used. And also, when the skill was first

developed, it actually excluded Brown and Black skin tones that was added later.

[00:13:15] And there's six categories in it, which is obviously not inclusive of the full diversity of human skin tone. But that imperfect scale is what we used when we were trying to collect our images, and we tried to actually assign the skin tone based on what was recorded by the clinician seen in person, because in photography, like, the lighting can impact what the skin looks like.

[00:13:40] So that's the scale. And so, we, particularly the light skin tones are Fitzpatrick 1 and 2, 3 and 4 sort of in the middle, and then 5 and 6 usually represents Brown and Black skin. So we, in our paper, are looking at performance of Fitzpatrick skin tones 1 and 2 versus 5 and 6, [00:14:00] which I'll just call like white versus Brown and Black skin.

[00:14:04] Got it. Go ahead, Raj. Roxana, how are dermatologists trained to rate Fitzpatrick skin types? Are they trained to rate these skin types? So, it's definitely discussed clinically, like when you write your medical note, and the reason that people care again is to try to make a decision on how easily someone might hyperpigment, meaning that how easily their skin might turn brown from inflammation, or how easily they might burn with certain treatments that involve light.

[00:14:37] So, it is something that we actually go through in training, and actually in work with Matt Groh, we looked at the variability of labeling skin images, which is different, obviously, than the in-person labeling or asking people about how easily they burn. But for a machine learning purpose, we wanted to understand how much variability there was between people. [00:15:00]

[00:15:01] And it turns out there is some variability, but most of the time if you have dermatologists labeling images, they're within one point either below or above each other. Got it. And just before we move on real quick, just because I find this interesting, how does that impact treatment recommendation or a patient's clinical course?

[00:15:21] What Fitzpatrick bucket they're put into? The real way to get, clinically speaking, it shouldn't just be what you think they're Fitzpatrick, but you should be asking them questions about how photosensitive they are, how easily they burn, if they have a history of having their skin sort of change color after inflammation, because you can't always know that without actually asking.

[00:15:47] Someone could have brown skin and actually be very photosensitive because of a medication that they're on. So, it's, it's important not to make assumptions without actually asking your patients about their [00:16:00] experiences clinically. Right. You could be a 3 or 4, but if you're on a lot of antibiotics or something, then you could be photosensitive or something like that.

[00:16:06] Yeah. Okay. So, I think that was very helpful background. So, if I understand correctly, you saw that there was this unmet need in the literature that most of the datasets were not diverse. We've seen this in other areas too, like GWAS, where a lot of the studies were done in white Europeans. You kind of had a sense of that this was happening for AI for dermatology.

[00:16:24] So, you actually collected a big dataset that had a huge representation of different types of skin tones, right? Could you tell us how you curated that data? Yeah, so we curated and de-identified data from Stanford. It was an effort of many researchers to make sure that we were comparing the, we were using the pathology reports to make the diagnosis.

[00:16:49] We were not just looking at the image and saying, oh, we think it's this. We actually had the pathology reports. We wanted the labels to be as clean as possible. We had a dermatopathologist sitting with a [00:17:00] dermatologist reviewing what's in the report, what it looks like clinically to assign each of those labels.

[00:17:05] We were looking at the assigned Fitzpatrick skin tone label in the chart, and then having two dermatologists confirm that they thought that that label was correct. So, for us, this was meant to be a benchmark. It is not a large enough dataset to train an entire, you know, deep learning model on, but it is one that you can use to benchmark against.

[00:17:30] How many images did you end up with? I believe it is 656. Okay, nice. I mean, it sounds like a lot of manually intensive labor given all of the eyes that sort of looked at each image. Yes, it was. And I know that that is not tenable when you're talking about wanting something like 10,000, you know, 100,000 images for training.

[00:17:57] Everything doesn't have to be ImageNet. I think that having [00:18:00] very good, high quality benchmark datasets that have a good amount of TLC, which it sounds like this dataset definitely does, are super important things to have to understand how these models work. That's how we felt too. And I think that, you know, for us, the other thing that we did was that for

Fitzpatrick 1 and 2, or the white skin tones, and Fitzpatrick 5 and 6, the Brown and Black skin tones, we tried to match the patients by sex, age, the time that the photograph was taken, meaning like to try to match the quote camera technology and sort of the diagnostic category so that we could do sort of a head-to-head comparison of those two groups.

[00:18:44] And then we even looked at, for example, we had to look at photo quality between the two groups. We had dermatologists label photo quality to make sure that the photo quality was similar between the two groups so that when people are running algorithms against looking at [00:19:00] performance in Fitzpatrick 1 and 2 and 5 and 6, we could minimize as many confounders as possible in that comparison.

[00:19:08] It's almost like you have a counterfactual lesion for each person in the dataset. Like, the only thing that has been changed has been the skin tone. So, it's not quite perfect. I wish it were because, you know, in some cases, we had to match categories as like non-melanoma skin cancers had to be a bucket in and of itself because it becomes difficult to match lesions perfectly.

[00:19:32] But in general, like every benign lesion has a benign lesion that's in a similar diagnostic space. Yeah, makes sense. So, what did you find? How do the current dermatology classifiers do? Yeah, so I mean we, one thing is, I'm all about open science. I think that's why we need open science because there's a lot of commercial algorithms or algorithms that have been published about that we couldn't get access [00:20:00] to.

[00:20:00] We asked. We, we went and asked around. Hey, you've published on this algorithm. You're making claims that this might be used commercially someday or touch patients. Can we test it on the benchmark? And usually the answer is no. So, we ended up – Roxana, when it's access to the algorithm, you don't even necessarily need the weights for the model itself.

[00:20:22] You just need some type of API or ability to run an image through it. Exactly. And even, even access to that was difficult for some of these commercial models. Yes, that is correct. Thank you for clarifying. Yeah, we're not asking for access to the weights. We're just asking for access to the interface at the time, or an API, or something. And we ended up testing models that were open source.

[00:20:49] So, these are not necessarily models that are going to be used clinically, but they were ones that we could access to and had been previously published on and had really [00:21:00] good performance. And we looked at

three different models. And the interesting thing, I mean this is not a surprise to anyone who's kind of been following this space, is that all three models had performance drop offs in general when they test it on a new dataset. And we know this because sometimes these models overfit to features in the dataset that they were trained on, and when you introduce some kind of new external dataset to them, they'll have some performance drop off.

[00:21:30] It could be because of differences in the camera technology used to acquire the images, differences in lighting. Dermatology is kind of hard because we don't have standards for how we take our images. So that was the first thing we noticed, but the part that was like most concerning to us was that

[00:21:48] there were significant differences in how the algorithms performed on Fitzpatrick 1 and 2, the white skin tones, versus Fitzpatrick 5 and 6, the Brown and Black skin tones. [00:22:00] That sadly is not surprising. Yeah. Can I ask a philosophical question? Do humans, are they worse at diagnosing dark skin tones also? Has anyone looked at that?

[00:22:12] That's an excellent question. So, I mentioned that a lot of the education materials for humans has underrepresented Brown and Black skin tones. And this has been something many other dermatologists in there's a skin of color society that focuses on making sure that we have equitable care across skin tones.

[00:22:36] There are many dermatologists who have brought this up over and over again. About representation, and education, and training of dermatologists. And in terms, you know, there have been survey studies that have shown that dermatology residents, some portion of them don't feel as comfortable making diagnosis across diverse skin tones.

[00:22:57] There is an amazing TED [00:23:00] Talk by Dr. Jenna Lester exactly on this topic. In terms of has anyone sort of systematically looked at this, stay tuned. We have a paper coming out where we did this with just images, which is obviously different than clinical care, but we wanted to look at it in the sort of like the teledermatology where you just have some image and not that much history.

[00:23:25] So there are differences that we saw in that upcoming paper. Because I ask because I'm wondering, like, how we fix this. And if you trained the model on a more representative dataset, but the labels are coming from this very error-prone process on dark skin, would that fix it? Or is there something more structural that needs to happen?

[00:23:46] You did this very thoughtful curation where you sat down and intentionally looked at the darker skin tones. And I guess, like, what does the path forward look like? So, I feel confident about our labels because we looked at things that were [00:24:00] biopsies and we had path reports on them. Now of course there's variability in pathology and that exists, but one thing we actually did was have dermatologists label the images and we saw differences in labeling between the two skin tone groups.

[00:24:19] Of course that doesn't, it's not representative of what happens in clinical care. That's just giving people an image and asking them to label a disease, which is they don't have the opportunity to ask history or do their clinical exam on the lesion. Uh, but from a labeling standpoint, we think that actually if you are just relying on dermatologists for labeling and you don't have ground truth,

[00:24:46] that's a loaded word. If you don't have a histopathology label, you have more noise if you have the dermatologist labeling the data. To go back to what you're saying, like, how do we fix this, right? Because I'm diving [00:25:00] into the machine learning side of it, but I think that there's the AI realm. But we need systemic change in medicine, too.

[00:25:10] Like, this is not an AI-only problem. AI is reflecting the biases that exist in the human realm. And my feeling has always been that you cannot rely on AI to fix problems that you have to actually tackle in the human realm. Which is, we need to make a concerted effort to improve the training of dermatologists so that dermatologists do a better job.

[00:25:40] We need to improve access to care. A lot of the health disparities that exist in dermatology are not AI-related issues, but they spill into AI. And I don't think that even if you created the most perfect, fair AI algorithm, that it's going to necessarily be a [00:26:00] bandaid for the problems that exist systemically within the medical system.

[00:26:05] At the risk of ruining what would be a perfect transition to the next set of questions that we want to ask you, I just want to ask one quick follow up. Just since you're an AI researcher and a practicing dermatologist, what's your sense of the penetration of this technology in dermatology? So, I've seen in radiology, it just all of a sudden has happened where there are AI systems in reading rooms and triaging chest x-ray reads.

[00:26:26] What's your sense of how much penetration there's been for AI in dermatology. So, it's been interesting in the U.S., image based – there's other

forms – image-based AI, there's been no FDA approved algorithms as of us talking. There are people running trials to try to get FDA approval. So, and there've also been now, finally, there've been, for a while, there were no even prospective clinical trials of say, like a dermatologist using AI to see

[00:26:56] if it improved, like, their sensitivity or specificity [00:27:00] for finding skin cancers. So, now there's been some prospective trials. So, we in dermatology are, definitely, I like to joke that we're a decade behind radiology. I don't know if it's truly a decade. In other countries, there have been things that have received, for example, CE mark.

[00:27:21] And have been used in clinics. They're also direct to consumer apps, which are a little concerning because as I mentioned, some of them make diagnostic claims without actually having FDA approval or published trials. So we actually, one of the things that we've done is we've looked at some of the consumer apps that are available in app stores in the U.S. and I mean, a majority of this, like we don't know anything about

[00:27:48] how they actually perform. There's like no published material. But in terms of penetrations, like, is there AI in my clinic? There's not AI in my clinic yet. How long do you think it is [00:28:00] going to be until there's AI? In my clinic. In your clinic, Roxana. Yeah. You know, maybe 5 to 10 years. The first place that I, my, this is my guess.

[00:28:12] Well, I guess, what are we calling AI in my clinic? Because if we're talking about like, large language models in the EMR system, people are already trying to test that out now. So maybe before, like, you know, if we're talking about dermatology AI and like image-based, my guess is that the first thing that'll come out, and this is just me guessing, is something with like a dermatoscope, which is the, so a dermatoscope is like, basically a fancy magnifying glass that costs $900 for no reason.

[00:28:46] It's like this device that you put on the lesion, it magnifies it, it shines a special light on it. It's nothing more than that, but it's a little bit more standardized because you put the magnifying glass [00:29:00] and everything in that field of view can be captured in an image. And one of the largest public datasets of dermatology AI images for training models, the International Skin Imaging Collaboration, is largely dermoscopy images.

[00:29:15] And so that space has moved a lot faster than the clinical image-based models. So, my guess is the first thing is that there'll be some AI companion to the dermatoscope that comes out. And there've been some

clinical trials in that space. So we'll see. Got it. To be determined. Does the dermatoscope plug directly into the EMR, or how does it interface, or does it have its own output and then you have to enter something?

[00:29:40] No, it right now. It's low tech. It's a magnifying glass. You can it has, it has an attachment. You can attach an iPhone to it so you can take a picture. So, my guess is that basically based on some of the trials – There'll be some readout – Yeah, you'll actually have to use like your iPhone to take the picture with the hardware, [00:30:00] right?

[00:30:00] The hardware is just going to magnify the image. You'll use your iPhone to take the picture and you'll have something that runs and gives you some readout. That's my guess. Got it. Very interesting to see. And thank you for entertaining the question because it's a little unfair question too, which is predict the future.

[00:30:19] I could be completely wrong. Yeah. We appreciate the speculation and I think you're the top person to do it. So, I think this is a good transition point, actually. So, you mentioned large language models. We want to stay on the topic of algorithmic bias connecting to your work in dermatology, but now switching over to LLMs, large language models.

[00:30:43] You published a paper pretty recently. I think it's titled "Large language models propagate race-based medicine." And as I understand it, it also hits another theme that you referenced a few moments ago, which is AI picking up bias that is [00:31:00] in society. That is in the way we practice medicine and potentially amplifying it and propagating existing ways of practicing medicine, but not necessarily updating as these models have changed, as race has been removed from some of these equations. Not staying up-to-date with current best clinical practice and recommendations from clinical societies.

[00:31:25] So, could you maybe just set the stage for us by telling us about what motivated this paper, and what you did for the study, and maybe briefly what you found. Yeah. So, I am a dermatologist, but as I like to say, I like to do research throughout all of health care AI, having done like a year of internal medicine training, having done rotations as a medical student, I'm not just focused

[00:31:52] only on things that impact dermatology. And of course, when I remember I was at NeurIPS, as [00:32:00] many of us were. And Andrew, I think actually one of your students was the first person to say, hey, did you guys

know that there was an update, ChatGPT had just been announced. Yeah, I remember that we were at like a cocktail party or something.

[00:32:15] And my student was like, there's this thing called ChatGPT. I think we should probably pay attention to it. And we're like, meh. Yeah. Yeah. Yeah. No, because this is, this is. This is a year ago now, right? November 2022. It feels like 10 years ago at this point. It does feel like 10 years ago. Yeah. That's how I feel about having started my faculty job.

[00:32:35] It's been like a month and I'm like, that month feels like a decade, but yes, actually it's hilarious that we're talking now because I remember the stage. We were walking back from one of the hosted company parties that they have at NeurIPS in New Orleans, and we were walking back and he – your student – said, you need to look at this.

[00:32:56] And we were both like, oh, we've played with the GPT-3 and [00:33:00] like, eh. And then actually, I remember I went back to my hotel room. And instead of getting ready to go to bed, I logged on, made an account, and started talking to it. And then I said, whoa, this is very different than GPT-3. And I started asking it a lot of medical questions, which I had done with GPT-3 as well.

[00:33:22] And now GPT-3.5 was just like answering it in ways that shocked me. And, of course, I think for anyone who's ever interacted with these systems, like for the first time, there's that moment of being very impressed and then you start to dig in and try to think about, okay, where the problems are. And I think that over the last year, a lot of people, and of course, GPT-4 has come out now and people have seen where some of the issues are.

[00:33:53] And so one thought I had. The reason we did this paper is because medicine is very slow to [00:34:00] adopt things, and I had been working on image-based AI. We had even done a clinical trial of one of our models, and that model was not even a diagnostic model, and it was just very slow adoption. And then all of a sudden it felt like overnight there were hospital systems saying, hey, we're piloting these large language models

[00:34:21] into our medical systems. And that was a little surprise. I don't know if you guys have been surprised and shocked, but I've been kind of. Well, thankfully the hospitals we're affiliated with move even slower than the ones that you're affiliated with, so we haven't had to address that yet. So, I said, whoa, like people are actually talking about integrating this into the medical system.

[00:34:44] So I have so many, I have so many questions, but yeah, keep, keep going. Cause I, and I want your perspective also on that at Stanford too, but because I, it probably is faster than, as Andy is saying, it's probably faster than the way things are moving here, [00:35:00] but I think, yes, there is a extreme amount of energy to get these into, into practice.

[00:35:05] I will say that a lot of this stuff is still in the pilot study phase, but it's not just Stanford. It's many academic institutions. I just had never seen something go from it just came out to now we're doing clinical pilot studies in such a short period of time. And so, my research group, we've thought about AI fairness and bias for a while, mostly in computer vision.

[00:35:29] And so then we started thinking about like, how could this show up with large language models? And so, one thought we had is that we knew that physicians and patients were likely asking these models questions, clinical questions. And so, we pulled, so there's this 2016 PNAS paper that looks at some of the incorrect race-based misconceptions that medical trainees have.

[00:35:55] So we pulled some questions from that paper, and then we [00:36:00] also had a group of experts, who are all authors on the paper, sit together and think like what other questions we might ask. And so, there's been a lot of discussion around the use of race in medical algorithms because race is not biological, it's a social construct.

[00:36:17] And so studies have shown that there can be disparate outcomes from using race in algorithms. And so that's why, for example, we don't use race in kidney function calculation. And in fact, that change was made before, for example, the ChatGPT models came out. So, if you ask ChatGPT about it, it knows about that change.

[00:36:44] And yet, if you ask it, how do you calculate EGF? So it references the new National Kidney Foundation, American Society of Nephrology, recommended race-free equation, but then recommends using the race-[00:37:00] based previous equation. So, yeah, I mean, I'm just saying, like, if you ask it, do you know about this information?

[00:37:05] Okay. Okay. So it's aware of it. It's in its artificial brain somewhere. Yes. But when you ask it, this is my creatinine, this is my patient, or this is my age. What is my EGFR? It will use the race stratified, the race based. We actually made it very simple. We just said, how do you calculate EGFR? We just asked it.

[00:37:25] So if you ask it, if it knows about the newer way of doing things, it knows. So that's just to show that it's not, that it doesn't know. It does know, but if you ask it, how do you calculate EGFR? It will give you, and actually, you know, all the models we tested had problems. So it's, I'm not just picking on ChatGPT.

[00:37:46] What other models? Did you test the other proprietary models? Yeah. BARD, CLAWD, 3.5 and 4 for GPT. Were the models meaningfully different from one another? You know, the way that the answers were written [00:38:00] were certainly different. We have sort of a heat map in our paper that shows like which models were better than the other.

[00:38:07] In general, many of them failed similarly on certain questions and many of them passed. For example, most of them, except for one, passed the question on the genetic basis of race, which, again, race doesn't have a genetic basis. It's a social construct. So, it's just interesting to see what they failed at and what they, I'm sure there's that reinforcement learning with human feedback for some of the questions that we asked is my guess.

[00:38:36] But actually, one thing that was really interesting to me that I wanted to point out is Some of the models, not only did they give the incorrect race-based equation, they actually gave an argument for why you should use that equation using like a false race, racist trope, which is that there's difference in muscle mass between races.

[00:38:58] Which is a completely debunked [00:39:00] racist trope, and so that was even more interesting in a bad way. Not surprising, because that stuff is out in the world, but deeply concerning, because if somebody comes in with those biases, it just gets confirmed, right? I'm curious, did the sort of rationalization come unprompted?

[00:39:22] It's like, hey, I'm, trust me, I'm not racist, like, here's the reasoning behind why I'm doing this. It was unprompted. I'm saying that this appeared simply in response to just asking, like, how do you calculate EGFR? There's a, yeah, there's a, like, me thinks the lady doth protest too much component to that, it seems like.

[00:39:41] Right. I mean, we also. We also asked, like, how do you calculate EGFR and gave the race of the patient as well, but even if you don't give a race of the patient. So, it's interesting because when you talk about lung capacity, if you just ask about calculating lung capacity, it doesn't give a [00:40:00] race-based answer.

[00:40:00] But if you ask it about calculating lung capacity and give a patient race, then it starts talking about. Debunked differences, it starts claiming that there are differences in lung capacity between races that are not supposed to be used that don't exist. So, you know, Roxana, similar to Andy's question about dermatology and how dermatologists fare in rating, basically treating patients with different skin tones.

[00:40:30] Do you think that the models here, so the large language models, do you think they're also reflecting the societal bias, the way that race has been and currently still remains in many instances, embedded into physiological equations and is used as sort of an inductive bias right into your model for physiology for many, but I think reducing applications in medicine.

[00:40:58] Do you think these models are [00:41:00] essentially reflecting that, I guess is the first question. And then my second question is, how do you think large language models, the ones that you tested, compare to practicing physicians in terms of bringing bias into the clinical encounter? So, I think for the first question, the answer is yes, because obviously the large language models are first trained on large amounts of text data that reflects human thinking.

[00:41:29] I mean, it's not like the models magically learned bias by themselves. These models learn from human data. And so, the biases in human data get picked up and regurgitated with large language models. I think your second question's a really difficult one. I'm not sure how to answer. People will say humans are biased and, therefore, the models are going to be biased or that's okay. And I actually don't think that's okay.

[00:41:58] I think if we're going to build these [00:42:00] systems and put them into the health care system, we need to be building systems that are fair and are not going to worsen health disparities or continue to perpetuate them as they exist now. And as I mentioned before, even if you built the most perfect model, having these systemic issues on the human side you have to fix that, too.

[00:42:20] Technology is not going to save everything. Yeah, no. So I agree. And I think the time to do that is now before they are, you know, before they are widespread. But I think understanding human bias is maybe a path to thinking about designing. Right? A robust human AI collaboration. And so we have all these nice examples of GPT passing this or that.

[00:42:45] Right. And we're almost desensitized to these papers now, right? Because it's been done so much. But we have very few studies of how humans

and AI operate together. Yes. And I think we're gonna have more of those. But just as in the context of accuracy, I think thinking about [00:43:00] bias, if we understand human bias and we understand the way the algorithm is biased, I wonder if there is a path forward that you see where we could design large language models to actually inject some nuance into the way that humans are reasoning, because humans have a ton of bias too, right?

[00:43:17] And it's not to say that that excuses algorithmic bias. But it is to say that we should be aware of the status quo and design systems and improve systemic education, all that, but I think design systems that hopefully allow that collaboration of the human and the AI together to reduce bias. I agree with, I mean, you could imagine that

[00:43:36] you could have the system designed to always give the correct type of algorithm that's like the most up to date and explain to the human doctor why, like this is why. Yes, like the kidney function example that what that would have looked like would be in 2021 the National Kidney Foundation American Society of Nephrology recommended [00:44:00] a race-free creatinine-based equation.

[00:44:02] Here it is. And then if they wanted context, they could have context on the previous race-based equations and further information, but that's not what you found. And then an explanation why, right? Because when those decisions were made, there were explanations of why, like how the prior equation could cause disparate outcomes, why some of the assumptions that were made were harmful.

[00:44:27] You could build that in. I see what you're saying, like it could be actually a tool for when you do the human-AI collaboration. The AI helps sort of the human understand like the state-of-the-art now. Great. All right. I think that was, Andy, is this a good time? I think it's a good time for the lightning round.

[00:44:46] Are you ready? Oh boy. Roxana, we have a lightning round. I am aware. I've listened to your podcast. Okay. So, you don't even need to set up so we can just hop right into it. So, I think that this is a good first lightning round question just because of how nicely it [00:45:00] dovetails. Will LLMs be net positive for medicine over the next 5 years?

[00:45:07] I don't have a great answer. I, I don't. Five years, I'm not allowed to say I don't know. These have to be short answers. Is it going to be net positive 5 years? I'm a little concerned because we really need to do more research and

have frameworks for assessing when they're working, when they're not working, and how biased they are.

[00:45:29] I think in the long term they could be net positive, but I don't know if 5 years is long enough. So maybe I'll, I'll just say perhaps net neutral because they won't do much of anything over the next 5 years. Is that, I wouldn't say they aren't going to do anything. I just, yeah, I'm terrible at predicting the future.

[00:45:48] Now, if there are several companies working on it, helping it write medical notes, and if they get that to work well. That would really make my life better. But right, so there are some administrative things like [00:46:00] that that are probably not harmless, but the stakes are lower. That seem likely to come over the next 5 years.

[00:46:06] And so if those come in the next year, I think they'll cause improvements in quality of life on administrative stuff. I'm not sure in terms of medical care. I'm still wary about them being ready for actual medical care. Roxana, if you weren't in medicine, what job would you be doing? So, I'm going to say like, not medicine, STEM, or science related at all.

[00:46:28] Because my answer to this is that I would be a national park ranger, and I would be the park ranger that gets to hike all day and check on everyone. Because I love, I just love being in nature. I love hiking. And it just seems like it would seems like a cool job. You get to meet people from all over the world and hike all day.

[00:46:45] Follow on question. Uh, top one or two national parks. Okay. So, I mean, I live in California, so, you know, I've been to Yosemite. Um, yeah, I've, I've kind of lost, I've kind of lost count number of times. [00:47:00] Pacific Symposium on Biocomputing is in Big Island. So, Volcanoes National Park is another one I've been to many, many times.

[00:47:08] Which is an amazing park because it changes every time we go, and there's lava, and you know I know you guys have your competing conference, but we have volcanoes. And that conference was started completely out of jealousy for PSB just to be clear. So, I like that because I learned two things: one, is that you like the outdoors and want to be a park ranger, and two, I think we need to switch up our lightning round questions because that one seemed prepared. It seemed like you had thought about that before. Oh no, actually it's because it feel like every icebreaker that I do always has that question.

[00:47:42] So, literally we just had a retreat where that was the icebreaker question. Got it. Okay, so the next one. What is the best thing that you have read or watched in the last year? The *AI Grand Rounds* podcast. Actually, the podcast is really good. It's been very. No, no, no, [00:48:00] just get something other than the podcast.

[00:48:02] Thank you though. Your check's in the mail. It I, no, no, the podcast, the podcast is quite excellent. Uh, you know, I'm just trying to think because it's been hard with small kids to like, really watch much TV. What's your favorite episode of Bluey? My daughter loves Numberblocks. So, it's been a lot of, I've been listening to, um, I've been listening to a lot of audiobooks.

[00:48:30] That's how I get through. I'm just trying to see, sorry, give me a second. I know that it's supposed to be fast. That one. Oh, it doesn't have to be fast because we can make it seem fast in the post. Oh yeah, that's true. I'm like trying to even, I like literally can't even remember. I've listened to some audiobooks, but I'm just trying to remember, um, what I've even like listened to recently.

[00:48:54] This is very embarrassing. I was gonna say, all the, like Netflix [00:49:00] or like HBO, nothing? I mean, I, nothing, uh, oh, you know what? I, I did enjoy, I did enjoy, I just recently listened to the audiobook, *Range*, which I liked because similar to in that book, I have spent a lot of time in training and exploring different things and that's kind of helped me put stuff together.

[00:49:24] And so that book talks about essentially how people have come up problems in new ways by bringing in like past experiences from different fields. All right. So, I, I think I know the answer to this question, but I'm very, very curious if I'm right here, will AI and medicine be driven more by computer scientists or by clinicians?

[00:49:48] I think it has to be driven by clinicians in the sense. I actually think it really has to be driven by teams, not one or the other, and interdisciplinary people who understand both sides. [00:50:00] Because domains, I think domain-specific expertise is so important. I talk to computer scientists all the time who are going after a problem that's not even a real problem in medicine.

[00:50:12] And, you know, I'm like, don't spend all your time and effort on this. Or they make some sort of assumption about the data in medicine that's not true. And, so, then they built this whole model that's built on some assumption about what the data is like that ends up being untrue. So, I do think they need domain expertise help and obviously clinicians who don't have AI experience

think that AI can do things that it cannot do or think don't realize how the models are trained and don't realize what pitfalls or biases exist in the model.

[00:50:46] So, really, I think it has to be interdisciplinary. People or teams. I think that that's been a recurrent theme on the podcast, so I think there's not much to object with there. Right. So, if you could have dinner with one person alive or dead, who would [00:51:00] it be? See, you can see I really did not prepare for these questions.

[00:51:08] That's the point. We like your candid take. I know. Um, who would I have dinner with? Oh, Marie Curie. I mean, as a, as a female scientist, like. Oh, yeah. Yeah, yeah, yeah. You know, I would love to, I mean, she's clearly brilliant and would just love to talk to her and see how her brain works. Yeah, that's a good choice.

[00:51:33] Awesome answer. Alright, our last lightning round question. Do you think that things created by AI can be considered art? That is an excellent question. I think that things created by AI usually are not in isolation from humans. So, I love art and I think that art is about expression of emotion and feeling and [00:52:00] experience, and it makes you feel something.

[00:52:03] And from what I've seen from AI-generated art, there's usually a human still behind it, right? A human trying to express using AI to express. The prompt. The prompter. Can I, can I, can I change the setup then slightly? Let's say that we hook ChatGPT up to Dolly, which you can already and say "make art." Okay.

[00:52:28] So that, that doesn't feel like art to me because again, I feel like art is about a human conveying their experience in some way. So, I think humans and AI working together can make art because there's some emotion or experience that's being conveyed that can then be taken in by the observer, but maybe not, like, just randomly generated images from.

[00:52:54] If I could put on my, like, elbow patches for just a second. If a human looking at the thing created by the [00:53:00] AI has some type of, uh, real feelings or, you know, beauty being in the eye of the beholder. Right. I guess in that case, it could. I think this is a very nuanced, complicated question that you're trying to make me put my foot down.

[00:53:14] And I don't want to get canceled on Twitter. We'll leave it at that. Alright, Roxana, you survived the lightning round. That was excellent. We just

have a few concluding big picture questions for you. The first is what areas of medicine do you think will be the most resistant to change from AI?

[00:53:34] Resistant to change? I mean, I think some of it is, appropriate resistance to change. I think we, the human element is just so important in medical care. As someone who's received medical care, as someone who's had family, who's received bad news. Like there's something so important about, you know, the empathy of another human being in the room, conveying that information, [00:54:00] helping answer your questions.

[00:54:00] So, for example, we should not be using AI to deliver bad news. And I think that actually people keep trying to use AI to build diagnostic algorithms, particularly with imaging data. But I think the actual diagnostic process that happens in medicine is quite nuanced, more than people realize. And this is an example that I actually recently gave.

[00:54:27] Just a very easy, like, not even a complicated case. So, we're looking at a lesion. We're trying to label some dermatology data, right? We have images of the lesions. We have the biopsy results. Turns out the biopsy results are not very definitive. And then we're looking at follow up notes, like literally this dermatopathologist and I are trying to like label this image data.

[00:54:48] And we're reading the follow-up notes to see what treatments were tried, like what the dermatologist thought in order to come up with a diagnostic label, because I think people think that diagnosis is [00:55:00] always black and white. So, I actually think that diagnosis is a lot harder than people realize, and I think diagnostic tasks, there might be some resistance there compared to using AI to help with administrative triage, decision support, but like straight-up diagnosis is actually a lot harder than here's the diagnosis.

[00:55:25] In some cases, some things are straightforward, but many things are not. So, for the first example that you gave, which is appropriately resistant, and I like the way you phrased that, appropriately resistant to change, things like delivering bad news, right, or counseling a patient that you don't want to come from a computer, you want this to come from another human.

[00:55:45] A lot of AI leaders now, medical AI leaders, are arguing that AI might actually enable doctors to have more time with their patients because they absorb some of that administrative burden, right? Or because they're allowing the doctor to make [00:56:00] eye contact with the patient and not just be entering things into the computer.

[00:56:03] I'm just curious, your personal stance. Are you optimistic about, let's say, digital scribes or AI agents that are listening to the encounter that allow you to then make eye contact, focus on your patient, have time to be there? I would love a digital scribe. I would say two things about it. One, need to ensure patient privacy on any company that's listening in to the encounter and building such a model.

[00:56:29] Patient privacy is key. Two, there needs to be ways to verify. Like, in case you don't, hey, I don't remember actually, the patient saying that, like being able to look in. You need to know the provenance. Yeah, provenance, exactly. And there are companies who are doing it in exactly that way. So, I think if you have those two

[00:56:48] sort of pillars, um, and also like the third, making sure there's not any kind of hallucination that happens, I think an AI scribe, if you can meet those criteria, would [00:57:00] absolutely allow more empathetic care, more eye contact. I actually work with a human scribe. And when I started working with a human scribe, it was totally changed my life because I'm not at the computer at all now

[00:57:15] with my patients, which is what I prefer. I don't want to be looking at the computer. But when you have so many encounters back-to-back, if you don't write notes down, you're not going to, um, remember exactly what happened, which is important. And so, it makes such a huge difference to have somebody, whether that's human or AI,

[00:57:35] kind of document that encounter so that you can just focus in because the thing is is that when you're with patients it's really important to look at their face to make eye contact, but also to sort of read what's happening How is the person reacting to the information they're sharing? That's an important part of the art of medicine.

[00:57:55] Do they seem anxious? Do they seem concerned? Do they look like they might [00:58:00] cry? Like, you need to know that information. You can't do that if you're staring at a computer. Great. So, I'd like to ask a completely different kind of question. So, in addition to being AI nerds, the three of us also have something else in common.

[00:58:14] We're all junior faculty who have small kids. So, I have a 4-year-old, we have another one on the way. I started my faculty job in the same month as our daughter was born. So, that was a crazy time in my life. So, I was

wondering if you'd be willing to share what your experience of going on this sort of academic ride has been like,

[00:58:32] while also balancing the needs of small kids and like, what strategies do you have? What has worked well? What hasn't worked well? Does the word chaos feel relevant? It feels, um, exhaustive. Super relevant. It feels relevant and exhaustively descriptive. Yes, I see the little trampoline you have behind you. Um, yeah, so I have, she just turned 5 and I actually also have another [00:59:00] one on the way.

[00:59:00] So, I'm 33-weeks pregnant. Congratulations. And, uh, yeah, starting a faculty job while pregnant has been an interesting experience. But as I tell, I try to tell the trainees, I'm like, there's really no good time. Like also you cannot time it. Sometimes things don't work out the way you expect it to. You just have to

[00:59:20] do what's best for your family and not try to time things. Cause it, unfortunately it doesn't work out like that. I think that for me, like my family is very important to me. And, so, I prioritize, I mean, one thing that's nice about our job is flexibility because I'm a clinician half day a week and I cannot cancel clinic last minute.

[00:59:45] But, if I needed to work on some writing and something happens, like, I can write later that day, or later in the evening when my daughter's gone to sleep. So, I do appreciate some of the flexibility afforded by an academic research career. I think I try to talk about this because the training is so long, especially for M.D. Ph.D.'s, which is again why I say, like, you just have to do what's best for your family.

[01:00:13] We actually don't live near any family, which makes it extra difficult, but we do have family that has come in and supported as needed, and I think that's really huge. I think it's really hard. I think institutions need to do a much better job of supporting young trainees and faculty that have kids. Cause whether that's like affordable childcare, cause childcare is so expensive or just having like easy backup options..

[01:00:44] Um, I don't know. How, how, how do you guys feel? No, I, I, that all resonates with me. I love that you also share that with your trainees because I, I try and do the same thing. Like, um, for instance, Friday is Veteran's Day. And so that means daycare is closed. And so that means we usually have our lab meetings on [01:01:00] Friday.

[01:01:00] And so, uh, it means that there's no lab meetings, you know, we're recording this podcast interview later than we intended to because my daughter got a fever and couldn't go to daycare. So, we had to reschedule and thank you for understanding that. But I, I think that I, I try and be transparent like that with my trainees too. One, because then they know sort of what I'm up to.

[01:01:19] But I think the surprise for me was how little insight I had into what being a working parent would be before I became one myself. And, so, I think that there's a big information asymmetry and you don't really know what it's going to be like until you're there. So at least I'm trying to hand it down to the next generation so that they can know what the other side is going to look like.

[01:01:39] I also think it's very helpful when you have colleagues who are working parents who then understand that meetings at a certain time, which corresponds exactly to the preschool pickup or drop-off don't work so well or completely. When you emailed me that your child had a fever, my immediate reaction was like, yeah, I've [01:02:00] totally been there.

[01:02:02] I'm not frazzled at all by that. In fact, you had my deep empathy. Also, because many times they come home and then they feel better and they're running around the house and then they can't go back for 24 hours, but they're not actually well. I always say that they often have a virus but they're not sick.

[01:02:20] Yeah, exactly. Ineligible for daycare, but not really sick. Yeah, exactly. Exactly. Alright, Roxanna, this is our last question. What are you most optimistic about for the future of AI in medicine? What I'm most optimistic about is the future, like the people who are in training now, the medical students, the graduate students, the postdocs.

[01:02:45] I have had a chance to work with many of them, and to give lectures, and a lot of people are caring about fairness and equity in AI. And I think that in order for this to go correctly, it [01:03:00] can't be like, oh yeah, there's a subset of researchers who care about fairness and equity and AI in medicine, right? And everyone else just does whatever.

[01:03:09] It actually has to be, like, this has to be like, baked into the system. It has to be, everybody has to be an AI researcher who cares about fairness and equity, not just one portion. And as I've given talks, I have found that I think like the next generation of researchers who are coming down the pipeline think this is deeply important. Care about this. Bring it up. Ask me about it.

[01:03:35] Even before, even before I bring up the topic, they're asking me about it. And that, I think, gives me optimism for the future because these are going to be sort of our colleagues and who are going to be hopefully helping build better systems and helping implement them. That's great. And just to connect that to your large language models paper on race-based medicine, I think it was [01:04:00] medical students across the country who were a big part of pushing for change in the way that these equations are incorporating race over the past few years, too.

[01:04:08] So, I think it's, you're, you're totally spot on and this has been a really, really amazing. Thank you so much for joining us for *AI Grand Rounds*. Thank you so much for having me. This was fun.