AI Grand Round Podcast #7
06.28.23
Can AI Be Harmful? A Conversation with MIT's Dr. Marzyeh Ghassemi

Marzyeh Ghassemi:

And this is a very early paper, right? This is pre sort of the GPT rush. A lot of what we were trying to point out there was, when you pre-process medical notes into some embedding space, there's a lot being captured there that you might not want to be captured. And first in the paper we show this very visceral example of what it might be, and then we show across many tasks and many potential ways of trying to fix the gap that you're unable to address the performance gap in the papers case specifically between black and white patients. So it's something that we sort of closed on in that paper and said, if you're using contextual language models, if you're using word embeddings to summarise a patient's state or to process a patient's medical record that's very long, and then turn it into a more compact representation, it'll probably get these important clinical concepts that you think it will, right? Because they're pretty good, but it'll get other things too. And some of those will be undesirable and will lead to horror performance, this what we found.

Raj Manrai:

That was Marzyeh Ghassemi MIT discussing algorithmic bias. Welcome to any NEJM AI Grand Rounds. I'm Raj Manrai, and I'm here with my co-host Andy Beam. We're excited to bring you our conversation with Marzyeh Ghassemi. She's an assistant professor at MIT in Electrical Engineering and Computer Science and the Institute for Medical Engineering and Science. She's been at the forefront of medical machine learning for several years and is working on developing and applying machine learning to understand and improve health in ways that are robust, private and fair. Andy, I really enjoyed this conversation and I learned a lot from Marzyeh.

Andy Beam:

I agree Raj, and I've known Marzyeh for a long time and she's definitely one of my favourite high entropy personalities. I find her research really fascinating, especially when she reveals how biases in clinical data can significantly impact the clinical outcomes of AI algorithms. There are a few points in this conversation that really stuck with me. And what I found really fascinating was her investigation into AI's recognition of patient race and medical imaging and how that poses important ethical questions for the deployment of AI. Obviously, I'm partial to her paper on explainability, which I co-wrote with her. But what I left this conversation thinking was that Marzyeh is not just advancing the technical side of the field, but she's also shaping a future where technology serves patients with equity and fairness, and I found that to be a very important message. And now we're excited to bring you our conversation with Dr. Marzyeh Ghassemi on AI Grand Rounds. The NEJM AI Grand Rounds podcast is sponsored by Microsoft and VIS AI. We thank them for their support. All right, welcome to AI Grand Rounds, Marzyeh, we're excited to have you here today.

Marzyeh Ghassemi:

Thanks for chatting with me.

Raj Manrai:

So Marcia, this is a question we'd like to get started with. Could you please tell us about the training procedure for your own neural network? How did you get interested in AI? What data and experiences led you to where you are today?

Marzyeh Ghassemi:

Wow. I think I got interested in AI when I was a master student at Oxford. I was there as a Marshall Scholar. And it was reasonably early days where Hinton's lab at University of Toronto was still demonstrating that there were reasonable results, but nothing that beat other systems and state-of-the-art benchmark problems in vision, for example. And so I did a master's where one of the things we looked at was could you predict acute hypotensive episodes in the intensive care unit, maybe as a precursor to sepsis? And we benchmarked neural networks as one of the methods you could use. But they did very poorly as we all know now because that was the pre-GPU, pre-millions of examples or even tens of thousands at that point of examples of demonstrations or episodes or data points. And so when I started my PhD at MIT one of the things that I had looked at was, are there ways to scale up some of these procedures?

And it wasn't until the very last year of my PhD where a lot of the methods that are now popularised and work really well were being released out of the U of T labs. And one of the final papers of my PhD that I was a last author on and a master student was a first author on, we demonstrated that you could actually do state-of-the-art prediction of in-hospital episodes of different kinds of tasks using recurrent and convolutional neural networks. And so there was this really big journey from when I did my master's and it was still sort of a toy to a set of techniques and methods, and really just hardware and data that enabled these methods to work the way they do.

Andy Beam:

Marzyeh, I've known you for a while but I actually don't think I know the answer to this question. How did you get interested in computer science and then how did you get interested in medicine once you had decided on computer science?

Marzyeh Ghassemi:

Well, I hope my family's not listening to this. I started college early and I come from a family of engineers. So my father is a chemical engineer and my uncle is a mechanical engineer. They're both professors, and I have another cousin that was doing another kind of engineering, they're all engineers. And when I started in undergrad, I was still living with my parents and I remember you had to select a major. And the only kind of engineering or technical field that seemed foreign to me, which meant that my uncle or father couldn't do my homework or look at my homework for me was computer science. And so I thought, "Oh, I don't know what that is. This will be great. This is going to be something that nobody else will be able to look over my work on." Which sounds very crazy but seems to have worked out. So I really enjoyed computer science. I think it's also, I'm one of those kids who really liked puzzles and I think a lot of computer scientists are kids who tended to like puzzles or thinking about problems in different ways.

Medicine happened because after I got my bachelor's, I actually went to work for Intel in Portland for a couple of years before I started my master's. And while I was there I was working for their health group

and they were trying to look at, can you take health data from some of their mobile platforms that were collecting different kinds of signals and then make intelligent predictions? And so it was very difficult to do at that time. This is even pre my master's, right? But we had some interesting projects that I thought were just really fascinating. And so I actually applied to the programme I did in Oxford for biomedical engineering, because I had had this really great experience working in the Intel labs on using different kinds of algorithms to predict with health data.

Raj Manrai:

That's great. So Marzyeh, I want to transition now to a few of your key research papers. So you direct research group at MIT at CSAIL. This is the Computer Science and Artificial Intelligence Laboratory. This is a storied place that has had AI in its name long before the current phase of excitement. You work broadly on medical machine learning, and one of the areas that I've really appreciated your work has been on algorithmic bias. I don't know if you like that term in particular, but let's say algorithmic bias is this sort of subfield of very important machine learning. We're all talking now about ChatGPT, GPT4. But you were working on language models and trying to understand what types of biases are latent or embedded in these models for many years before ChatGPT first came online. So maybe we could start with your paper titled Hurtful Words, Quantifying Biases in Clinical Contextual Word Embeddings. Could you maybe first tell our listeners what word embeddings are and then maybe tell us how this project came about and what your team's main findings were?

Marzyeh Ghassemi:

A maybe good way of thinking about contextual language models and word embeddings is if you had lots and lots of pictures of something, right? And you were trying to remember what order the pictures came in, or people sometimes do this for passwords. You need some sort of pneumonic, you need some sort of mapping to remember, first I saw a golden retriever, then I saw a chihuahua, then I saw a bulldog. And you essentially find ways that these things are similar maybe in some space. And then you tell a story to yourself. And that's how people often remember really long sequences of things. Language models aren't too far from that. If you want an analogy, you take lots and lots of examples in some observed space of sentences or words or images. And then you find some latent space, some sort of mapping where similar features are grouped similarly.

And just like the mnemonic allows us to relate these things that are similar in feature space, in the observe space to each other in our own memory, but by thinking of this story we've constructed in our mind, you can think about this latent space in the model as being a way for it to understand if I needed to substitute out a word in a sentence, what would be an appropriate word? Well, it's seen lots of examples and just in the same way that we construct this sort of internal model of what words are similar and how I might relate different images or concepts to each other by looking at many, many, many examples, this latent space often penalises similar things being far apart. And so you can think about word embeddings or contextual language models as having some sort of space in which things that are semantically similar or meaningful being near one another.

For this specific research project, what the lead author did, my student did is they took one of the contextual language models that's publicly available. This is the cyber model. And it's one that has been trained on scientific abstracts. And so we thought it was a good method to look at for this setting. And then we took a real medical note and we removed parts of the note that specifically had a patient's race. And so the snippet of the note said, "Blank patient was belligerent and violent. Patient was sent to, fill in the blank." And it's the first figure of the paper because we were really surprised by the outcome. I think we had expected it for maybe other language models but not for cyber. We found that if you said that a

white or Caucasian patient was belligerent and violent, then the model would fill in the rest of the medical note with they were sent to the hospital. But if you put African, African-American or Black patient was belligerent and violent, then the model would auto complete the note with the patient was sent to prison.

And we found that there was a performance gap when you use these contextual embeddings to process notes between different kinds of patients. So here we specifically looked at patients of different self-reported race. And this is a very early paper, right? This is pre sort of the GVT rush. A lot of what we were trying to point out there was, when you pre-process medical notes into some embedding space, there's a lot being captured there that you might not want to be captured. And first in the paper we show this very visceral example of what it might be, and then we show across many tasks and many potential ways of trying to fix the gap that you're unable to address the performance gap, in the paper case specifically between Black and White patients.

So it's something that we sort of closed on in that paper and said if you're using contextual language models, if you're using word embeddings to summarise a patient's state or to process a patient's medical record that's very long and then turn it into a more compact representation, it'll probably get these important clinical concepts that you think it will, right? Because they're pretty good. But it'll get other things too and some of those will be undesirable and will lead to poorer performance is what we found.

Raj Manrai:

And how do you think the field has sort of changed over the last few years since you've published this? We now have very large models, right? It is being ubiquitously used by high school students to professors to likely medical professionals, as well in many ways on a daily basis. Are the lessons that you found from that paper applicable to the models today? Or are you optimistic that the field is sort of addressing this in a way that they were not addressing this a few years ago?

Marzyeh Ghassemi:

I don't know if you saw the reasonably public release of the GPT4 paper on arXiv, which is the platform that many people share pre-prints of their content before it's peer reviewed. There was a really unfortunate incident where they had posted the LaTeX, the underlying code for generating this paper. And in the LaTeX, it's screenshotted and on Twitter now. You can see that there was an entire section on the toxic language generation of GPT4 unprompted. That the authors had very, I think reasonable content about and were discussing in a very intelligent way. So they have this whole section. And then in the actual paper that whole section is commented out. And so I think that the answer to your question is, this is definitely still a problem as acknowledged in this pre-print by the people who released the GPT4 paper.

I don't think anybody who works with large language models would argue that we often see toxic content being generated by all forms, all capacities of language models, and that this is an active area of research and it is extremely unlikely that we are going to remove all human biases from large language models. So I think that's a reasonably well agreed upon point right now. Something that is maybe more contentious or that is being actively discussed is what to do about it, right? So there's both sides of the spectrum. Some people say, "Well, kill all the language models. Why do we even have them if we know that they do these bad things?" Some people say, "Well, it's imperfect just like every single other thing you use and really is it so much worse than humans? It's just parroting the bad things that humans say." And then I also think those are different perspectives maybe from researchers.

There are many regulatory levers you could imagine pulling about large language models and these are being actively discussed right now in Congress, right? Where they're having discussions about if we have bias content that's generated, is that in and of itself something that needs to be regulated? Is it the use of that bias content in a commercial system that is then handled by a federal consumer protection agency? Does it need to be handled by all of the different offices that handle violations of civil rights? So like HHS in a healthcare setting for example? So I think understanding what the implications of having a flawed tool like this are, is something that's still actively debated.

Raj Manrai:

So that's great. And I think this area is so important. I think you're hitting the nail on the head here that, there's a lot of contentious debate right now around how to address the problem and where the sort of regulatory oversight can be most useful for moving forward. Maybe I can ask you a somewhat related question, but this is more about your lab and the way you select problems for and design approaches for both auditing these models but also designing mechanisms to lead to less biassed and better models that perform well across populations. We've had these debates in the field both in the sort of general machine learning community and in medical machine learning, about whether the problem and therefore the time that's spent on diagnosing sources of bias is best on the data and the lack of data, representation of data or what the data's encoding about, the healthcare system or access the care or things like this.

Or whether it's sort of best spent on looking at the model or designing new models that are better able to tackle problems in a fair more robust way that works better across populations. Is this a meaningful distinction for you when you are selecting projects in the group advising students, how do you think about where your group can have the most impact on improving large language models and removing these very pernicious sources of bias?

Marzyeh Ghassemi:

Well, I would say my group in general focuses on what we call healthy machine learning. And so this isn't limited to large language models because there are many classes of models that are used in healthcare settings. We have had several projects that have focused on understanding where bias enters models on the data side, and whether state-of-the-art models that are available today that are standard within the wider machine learning community function well in the specific types of spurious correlation or attribute shift or class imbalance that we tend to see in health. And the answer is often no, right? So many of the benchmark data sets that you see that are available for model training, they are collected in specific ways or simulated in specific ways to have specific attributes or specific kinds of shifts, specific kinds of correlations. And those don't tend to match what we see in practise in medical data sets.

And so some of what we do is try to understand what are these attributes of medical data sets and are there ways to maybe identify better ways of either collecting data or robust find methods to very biassed samples? I will say some of these problems are hard to address because the samples are small and cannot be made larger. So you can imagine if I complained that the gestational diabetic prediction model they use doesn't work well on Western Asian women, my doctor might say, "Well, there aren't that many, not just in this sample, but generally it would be quite hard to increase this sample." But when we talk about models just not working on women, which they don't often, that's maybe a more difficult, less defensible error in a model because women are not a minority. And so maybe you would imagine that when I see specific kinds of biases in data, it's easy to attribute some of them to very defensible gaps that we see in healthcare systems because populations just often are not present in the kind of volume needed to do machine learning well.

And in some cases it's completely indefensible on this is because the systems themselves do not work well for certain patients. A lot of what we do on the non-data side, thinking about how models themselves can become more robust when you have a certain baseline of data available, is thinking about what makes this model perform poorly. And often it's something that is a flaw in model learning that is not a flaw in non-healthcare settings. So think about how differential privacy works. If you haven't heard of it, differential privacy is a state-of-the-art technique to guarantee that if an adversary has access to your model or outputs from your model, they can't recover underlying information, underlying data in the model. And so it's a very popular technique if you need sort of bulletproof privacy against attackers with unreasonable means. But the problem is the way that differential privacy works, the technical mechanism that makes it bulletproof is anytime there's a point during the machine learning process that pulls the model's weights around too much, it's influencing the gradient too much in this batch of data you selected, you noise and clip the gradient.

And so you're saying if there's a point that's kind of an outlier and it's like pulling things around that'll make it really recognisable at the end for the model. And we don't want that, that violates privacy and so let's noise and clip that, right? And that's how you address it and that's how you get these amazing learning guarantees. That works for image data and many other kinds of data, right? Because you have all these outliers that you maybe don't want to identifiable. In healthcare data, outliers are minority patients. Which means when you apply differential privacy in a vanilla way to learning in healthcare data, you noise and clip in our studies black patients. And so you're losing predictive performance and influence of minority patients by default. And so I think some of what we try to address on the model side is that the default settings of machine learning models are often not things that are desirable in healthcare settings. And then trying to resolve, "Okay, what is desirable in this setting?" Often it's robustness.

What kind of robustness to what kind of perturbations, what are the kinds of noise we might expect to see in these settings? So there are some really interesting technical challenges to be overcome and often they're really just inspired by on the ground problems that we encounter.

Andy Beam:

Marzyeh, maybe I could hop in here and ask a question sort of a follow up question. I'd like to get your thoughts on the following conjecture that I just made up. It seems like one of the problems here is that we have a model, we have the model. And we are operating under the assumption that a single model will be able to serve the preferences of everyone. And what we've seen recently is a movement towards aligning a single model towards different tasks, either using reinforcement learning or instruction following in the case of language models. So is it possible that some of these... I think obviously some of the things that you've pointed out in your papers are by any objective definition, horrible and we'd want to get rid of those.

But do you think that there's a potential to reduce some of these problems by having everyone have their own sort of personally aligned version of these models where you can fill out a questionnaire or a set of preferences and then the model will be adapted to you, versus trying to think about how do we create one model that maximally serves all possible populations?

Marzyeh Ghassemi:

I think having one model that maximally serves all populations is definitely impossible. I think the issue with having a set of models that are customised to different settings is who gets to decide on those settings? Right? So I think ideally you would want every individual to get to decide on the settings of the models that, for example, determine what care their insurance will pay for or determine what

medications they have an option to choose or any of these potential questions. And I think it's very unlikely that patients themselves or even maybe providers will be the people who are allowed to tune or provide alignment to models when we have variation. So I think that one of the things that will be important as a community as we move away from this understanding that there can be one model that does everything, which obviously does not work well. I think it's going to be really important to make a decision about who gets to decide what the alignments are of any potential model that is deployed.

A toy example of this is you can put GPT4 into sort of very strict factual generation mode, where it doesn't move off of its evidence supported manifold very much. Or you can tune it so that it imagines, that's a human word that people have used to describe this, but it interpolate between portions of its latent space where there aren't a lot of samples, there's not a lot of evidence. And you can imagine as a human that in some cases one or the other might be more desirable. And so I think having more customizable models and just more models in general for different settings makes a lot of sense. One thing I will say is, if you tell me that there is a model that cannot predict cancer risks or thrombosis risk after chemotherapy for all patients equally, well, I would say that makes sense. People have different physiologies, there are different kinds of cancers, people have different sensitivities to these chemotherapy drugs. It might make sense to have multiple models that are each learning what makes sense for different kinds of patients.

If you tell me that you're not able to build a risk score model that predicts postpartum morbidity and mortality that works for both black and white women, I would be a lot more sceptical of the reasons why you need two different models. Because if you can't use all of the same data to predict the risk between these two groups, there's probably something wrong with the data you're collecting and more models are not helpful. So I think sometimes when people say we might need more models, you can maybe push back and say, "Well, maybe that in and of itself is telling you something and this process needs to change."

Andy Beam:

Got it.

Raj Manrai:

Just add that I think the desire to sort of customise or tailor the model to aspects of an individual has really been a very contentious debate with certain aspects of let's say demographic identity. And so race for example, has been the subject of a lot of focus in estimated glomerular filtration rate, eGFR, indexing for kidney function, but many other areas of medicine. And so I think it's pretty closely aligned with what Marzyeh is saying here, which is that the kind of inductive bias or the set of attributes that are determined to be those that we can index on or not index on are not always going to be aligned with what's in the best interest of the patient. And so there, there's a movement now across medicine with common physiological equations to move away from indexing on race. Because of reification of race as biology, the delayed access to care and not understanding the actual sort of causal factors that race is proxying for, why we're observing differences in eGFR pulmonary function testing equations across different groups.

Marzyeh Ghassemi:

So this is interesting because, well, for two reasons. Number one, because I moved here recently. So before I was a professor at MIT, I was at University of Toronto for two years. And they do not collect self-reported race in the Canadian healthcare data. Which makes it impossible to verify how many, for example, black or white women die in childbirth because they don't actually know who is black and

white. And when I spoke to one clinician about this, they said, "It's because we have no biases here so you don't, that's an American problem." I'm here to tell you it is not an American problem. This is a problem everywhere. It may not be that you have the same biases against the same groups. I promise you there are biases in society. And so I think it's very important that we collect this information no matter what, right? This is why we collect self-reported race for education or any of these other settings where equity is important, it's because of the Civil Rights Act. We need to know what is happening and data is powerful.

We do have some results from a paper that will appear in ICML in a couple of months now. So it's a new paper where we show that you can, in many cases find settings where using group attributes and prediction will improve your overall performance for the entire group. So you're trying to, for example, predict sleep apnea. If you include sex and age, then you can train a logistic regression model and your overall performance will go up. But your subgroup performance could go down, for example, for older female patients. And so there are ways that you can try to constrain learning such that no subgroup is adversely affected by inclusion of group attributes or demographic attributes. But as it stands right now, one of the issues is we often assume that more data is better. And that's true on average, right? That's true overall.

But there are some subgroups that for example, would have benefited from lying to you and saying, "Well, I don't know what the gender or the age or the self-reported race of this individual is." Because they would have gotten better performance from a model that was not allowed to use those attributes. And so I think balancing this overall performance of a model with the loss of performance in subgroups is another thing that we should consider as we relook, as we reexamine all of these medical equations and risk scores that have traditionally used demographic factors.

Raj Manrai:

Yeah, I think it's a super important point. It's a very critical distinction between monitoring and auditing bias across different race groups and embedding race is sort of a predictive feature of a given physiological equation. So you mentioned Canada, but I think France also takes this approach and maybe a few other places too, of not collecting or actually being illegal to collect race and ethnicity information. Super interesting. So I want to transition to a different data modality, but stay on this topic before we get into a great and totally non-contentious discussion on explainability.

Marzyeh Ghassemi:

Oh no.

Raj Manrai:

And then Andy is going to lead in a moment. But before we dive into that wonderful topic, I want to just stay on this topic of bias and latent information just for a few more minutes that's encoded in medical data and go from text to imaging. So you published this paper in Lancet Digital Health last year. And this paper was called AI Recognition of Patient Race in Medical Imaging, A Modelling Study. I'm not a radiologist, I read the paper, I looked at the images and some of these images you were doing these kind of sensitivity analyses or additional experiments where you're adding noise. And it really just looked like a white noise to me that you're turning on the TV and you're not getting any reception on your antenna and you're taking these types of images that are really noised, feeding them into a machine learning model and you're able to infer the patient's race for who that image belongs to.

And so this was a wild result and I can't say I have any hypotheses for why this would make sense, why this would be a signal that you could infer. And in some ways it relates to this kind of theme that we

discussed in the context of language models, right? Around what's baked in, hidden, not in sight, but hidden into these often otherwise inscrutable or under scrutinised models. So could you first maybe just tell us about why you started looking at... It's a very interesting question, but why you started looking at this question and then tell us about what you found in the paper? And maybe as much as you can share an update on where that paper, where this field has moved or where you've moved since this paper was published last year.

Marzyeh Ghassemi:

I have to admit, this paper involved a little bit of gaslighting of me to a student. So sometimes you tell students to run an experiment and you tell them to do it as a falsification hypothesis. So you say, take this data and predict this variable. You'll be able to predict it basically perfectly and then try to predict this other variable and you won't be able to predict it perfectly. And you should not be able to predict it. It should be, the AUC should be no better than noise.

Raj Manrai:

It's a negative control.

Marzyeh Ghassemi:

It's a negative control, right? And then you tell them, right? And if you can't, then your student have made a mistake and mixed up... There's leakage from your training set to your test set. So go make sure that your code is actually correctly written. I'm sure you've done that with a student, right? I will never do that again to a student. So this paper largely was a result of several of these back and forth with some students. And so there's a group of radiologists who are collaborators on the paper. And they had asked this question, "Can you do this prediction?" And they had some preliminary results. And I had told my student to go reproduce it and I said, "This is not something you can do. This is a falsification hypothesis. There's no way this is possible." And they did it. And then I said, "Well, you've miswritten your code." And I added another student and said, "Go check their code." And then two students did it and then we started doing these weekly calls where I felt like I was losing my mind, just a...

One of the radiologists also felt like she was losing her mind and started just searching for papers and came up with a couple of really strange papers, one about X-rays of bird feathers and how you could tell the colour of the bird feather, the level of melanination in the X-ray. And then she also found this other... These are very niche papers by the way. They're like older papers, not in journals I recognise. And she found another paper, because we were all just grasping at straws at this point. She found another paper that suggested that when you fed mice... So it was this really random result. They had all these white mice and some of the mice, they were using the mice for some experiment, they had fed them mushrooms and then they euthanized the mice and took X-rays of them. And they had this weird footnote where they said, you can tell that some of them, the mice, in the X-ray ate brown mushrooms because the brown mushrooms are more melanated and it shows up somehow in the x-ray.

And so we thought, okay, some of this initial evidence plus a lot of what we've seen in commercial camera calibration error on darker skinned people, might indicate that this is because melanination levels in darker skinned patients who are more likely to suffer report African-American or Black race, are probably being detected by this imaging modality that is using different parts of the frequency spectrum. And it probably influences bounce back, right? There is no way to verify this unless you have photographs of people's skin because we're using self-reported race. And we don't know, it would be ridiculous if this was true, but it could be that self-reported race is not correlated with melanination

level, right? And so we also tested... We asked the radiologist name every crazy thing you've ever heard somebody say, could be a way you could detect somebody's self-reported race from a chest X-ray.

And they said, well, maybe it's body mass index. We checked that it's not that. Maybe it's breast density, it's not that. Maybe it's bone density. We check that it's not that. Maybe it's the disease distribution. We check that it's not that. And so based on some of the images you saw that look like white noise, they've been banned past or high pass filtered. So you only allow certain parts of the frequency spectrum through. And so I had one reviewer for this paper say like, "Why are you even making this point? Is it important that maybe some of this high frequency information and chest X-rays conveys something about melanination level, which is correlated with self-reported race?" On its own it's actually not... It's like fantastic, and in a technical sense just really amazing that we figured this out. I'm so honoured to have been part of this project and part of these calls where people just were shocked and did not understand, and we feel like we were doing real science as opposed to engineering, which a lot of machine learning can converge on.

But on its own, it is not a bad thing. I am going to say, just by itself. The fact that we have these different levels of melanination that impact in small ways, chest X-ray bounce back in these imaging modalities and machine learning model can detect that and humans can't, that's not a problem. What is a problem is that you could have a model that perfectly detects self-reported race when humans cannot. Radiologists cannot detect self-reported race from chest x-rays. We tested them and they can't. And that model could be perfectly wrong on every black patient it sees. Radiologists don't often get a patient's self-reported race when they get a chest X-ray. But they get maybe some indication, 32-year old male check for fracture, but they don't get all of the patient's medical records or demographics. And you might not be able to tell that a model was consistently misperforming on only one kind of person.

And so it's not bad in and of itself, it's just something where we need to recognize it's this perfect illustration of recognizing that self-reported race is both proxied in a lot of medical data in ways we will not be able to understand or detect. And it is a proxy for many things in medical data, again, in ways that we won't be able to understand or detect.

Raj Manrai:

Is it fair to say that the one, I think this is what you're getting at too. But one of the implications of this paper is that you might think that a convolution neural net or a new age vision transformer that is not using race to predict some particular outcome is actually using race that's embedded in chest X-rays-

Marzyeh Ghassemi:

Definitely.

Raj Manrai:

... or some other modality. And so if you are making some assumptions or conclusions about the lack of use of race from these images, those are immediately called into question. So are you still working on this sort of threat of inquiry or is this prompted kind of future studies or?

Marzyeh Ghassemi:

Yeah, so we have a paper we're working on right now. Looking at the best ways to try to remove the self-reported race of a patient while not influencing your ability to use medical imaging for different clinical tasks. And it's not always easy to do depending on the label you're trying to predict and the imaging modality you're trying to use. One of the things that's really strange is, we found that across all the models that we've evaluated, let's say that you train a model to predict some medical outcome,

right? And then you just take that representation that has been used to predict some medical outcomes. So you take the latent space. And then you try to predict a patient's self-reported race with it. That model hasn't been trained to predict race, right? But we found that many of them can at a very, very high capacity just as a sidebar.

And then we looked at amongst all those models we've trained, which ones have the highest rate of disparity. So the highest difference in true positives between black and white patients. And we found that the more obvious it was from the model who was black and white, the worse the disparity was in the model's predictions. And so what we think is that sometimes this side information that's being learned by models is really harmful to improving overall clinical accuracy of models. And so if you can balance those two trying to remove this side information that's being learned and proxied into medical images, for example, and get good performance on medical tasks, that's going to improve everything and lead to a much more robust model ultimately.

Raj Manrai:

That's really great and I think that's a great moment to transition to understanding what is in models and what models are using and whether we need to be able to explain those in healthcare. So I'll turn it over to Andy.

Andy Beam:

Yeah. So that's a great segue. So, we're going to enter the explainability portion of the conversation here, Marzyeh.

Marzyeh Ghassemi:

No.

Andy Beam:

So we're going to talk about your paper, The False Hope of Current Approaches to Explainable Artificial Intelligence in Healthcare, that you wrote with Lauren Oakden-Rayner, and then some second rate scholar.

Marzyeh Ghassemi:

Some random guy. Like I met him on the street-

Andy Beam:

Who doesn't-

Marzyeh Ghassemi:

And he just seemed nice.

Andy Beam:

So I think the context for this paper was that Lauren and I had been at a meeting and had been asked about, should explainability be requirement for medical AI? You and I had had many conversations expressing our concern that this was going to be a requirement for medical AI, that it was going to have to be "explainable." And I'll give you a chance to attempt a definition at what Explainable AI is. I certainly, Raj has heard me talk about this topic long before I think this paper came out. So, maybe you

could just give us the context for this paper sort of where we were with Explainable AI. I have my own two cents on this paper, but first I'd love to give you the floor to talk about what you hear when you say Explainable AI and what are the pros and cons for making AI "explainable"?

Marzyeh Ghassemi:

I think the problem with explainability and Explainable AI is actually exactly that it is a technically squishy definition. There is no real way of verifying that something is explainable. And it seems to me that many of the most popular explainability methods are often post-hoc simplification methods. Right? And so these are methods and some are very popular, I will say. So lime or shop for local explanation methods, sparse decision trees or generalized additive models are these local explanation methods. But all they're doing is adding a simpler model, in many cases, a linear projection locally or globally to this big snakey decision boundary that your black box model has created. And again, my first problem with this is that this definition in and of itself is squishy. The second problem is anytime you explain something, and I'm using explain now to mean using these post-hoc explanation methods that are often simplifications with local or global linear projections, you make it slightly worse, right? By default, right? We are simplifying, we are removing degrees of freedom and complexity from a model so that we can understand it in some lower dimensional space perhaps.

And so other work that we've done recently, so this is a paper that a student of mine wrote called, you'll Love this, Andy, The Road to Explainability is Paved with Bias. We found that when you use these state-of-the-art, very popular local and global explanation methods, they make your results less fair. And they make them less fair for patients we found of different self-reported sexes, but then also for many other settings. So this is for adult income estimation across different sexes for an educational setting of, do we think a patient could pass an exam for people of self-reported races? And also for recidivism risk prediction, which is sort of this famous example where models tend to be unfair. And so we found that when you explain a model, you make it less fair. And then the level of explanation, the more you explain it, the less fair it gets because you're making it simpler and simpler and simpler. And so my first problem is that you're demanding this thing that is not well defined. My second problem is when you demand it, you make the model worse according to this metric that I really care about.

And my third problem is we know explanations make people more likely to follow bad advice. And we've known this for a long time in robotic systems. There's really fantastic work by several prominent robotics demonstrating this, especially in settings where people are under stress or believe that a system can mitigate some risk or has access to information that they don't. I would argue that medicine checks a lot of these boxes, and so it's not a setting where we want explainability that will turn off critical decision-making skills or engage automation bias. I think the parting shot for this is medicine has many, many black boxes. Some are true black boxes. We do not know how acetaminophen works or lithium. And others are not real black boxes like an MRI machine, we know how those work. But most people who use them do not really know how they work and they don't have to know exactly how they work. They just have to know that they are well calibrated, regularly serviced, and that the output should be used in a specific way for decision making. So that's my whole case.

Andy Beam:

So I think that was a good summary. If I could just maybe put a finer point on it. I think the question that I hear a lot from clinicians primarily is about trust. And they will say something like, "How can I use this black box AI tool if it can't explain its reasoning to me?" And I think that wish comes from a well motivated place. They have had to explain their reasoning as part of their own training on rounds. They usually have to give some type of systems based interpretation of what's going on with the patient that

would then recommended a treatment. So, how would you respond to a clinician who says, well, I can't trust this if it can't explain how it works to me.

Marzyeh Ghassemi:

I would say you do trust many systems already that cannot explain how they work to you, first. And then the second thing that I would say is, you would be more poorly served by a model that could explain itself to you. Because human psychology has been well probed and anything that can explain itself is far more likely to fool you. And we have very, very good evidence that in human decision making, models with more perceived transparency or explainability hamper people's ability to detect when a model is making a serious mistake. And the same people are able to detect those mistakes in, "black boxes" where there is no explanation. And so I understand the desire, but I think as people who work with evidence, we have to recognize that maybe this desire is something that is not actually what's best for both patients and providers.

Andy Beam:

I also think that we often want to use these systems that don't admit a simple causal explanation. So if there's some simple monogenic disease, and we understand the biology behind that, you really don't need these systems. But when the goal is adding up lots of very small statistical contributions to disease, again, it's hard to reduce that to a simple explanation. And that's precisely the area where we want to use tools like this. So I think that there's also kind of a use case mismatch.

Marzyeh Ghassemi:

Agreed.

Andy Beam:

That we want to use these in situations where we want some type of well calibrated probability that integrates a ton of information. If we're not in that scenario, we probably don't need these tools in the first place.

Marzyeh Ghassemi:

It's true.

Andy Beam:

Raj, was that sufficiently well balanced? Do we need to push back?

Marzyeh Ghassemi:

I feel like you got the wrong people.

Raj Manrai:

As a non author of this paper, I could ask one question to both of you. So this is piggybacking off of Andy's great question on trust. Do you think this is sort of fundamentally an empirical question, where we should ask doctors what they need and patients what they need for trusting a model and that we need more survey research, we need more human subjects, kind of what makes you trust a model or not trust a model type work?

Marzyeh Ghassemi:

I don't think we should ask people what they want. I think that there's lots of very good evidence demonstrating that both people are very poor judges of their own performance on things, and also very poor judges of what actually helps them be better at operational processes. And this is not just in health, this is across all aspects of humans performing jobs. I think we need more human subjects research. And the analogy that I think is really powerful here is all the research that was done in the aviation and space industry, to try to figure out how pilots would best be served by having lots and lots of automated systems that sometimes give you information and ask you to make a decision, and sometimes just tell you what to do. And how best to integrate those into a spaceship or into an airplane so that we have more safe processes overall.

And that process of studying how best people work with automation is something that requires that we actually put, in that case it was pilots, but in our case it's doctors. That you put doctors into these settings where maybe there's a requirement that you have to undergo extensive training, that includes hundreds or thousands of hours in simulation so that you understand how best to interact with automated systems. And that's what we do, again with pilots right now and that's managed by these federal agencies. So if we want to get to a place where we're comfortable and confident that these systems help us be better, we need to study that instead of serving people about how they feel about technology, that's also important. It's important that we feel empowered, that we are happy in our jobs, that we have tools that we enjoy using. But if you have a tool you love using that leads to more patient death or more physician burnout. I don't think that's the ultimate goal.

Raj Manrai:

So studying human machine collaboration and building that trust via a sort of natural understanding of how humans and machines can work best together. Andy, what's your view on [inaudible 00:51:37]-

Andy Beam:

Yeah. No, I agree. I think that user preference surveys are not the way to go. But looking at some type of outcome which includes physician satisfaction and things like that. I totally agree with that. And I just know on Marzyeh's point, I'm sure that our listeners who are young physicians coming out of training who are in their 10th year of post-graduate education and training are thrilled to hear that they need to spend a hundred to a thousand hours in a simulator to be able to use these tools.

Marzyeh Ghassemi:

Does nobody play Switch or Xbox anymore? I mean.

Andy Beam:

I mean, I do. Tears of the Kingdom just came out so obviously I'm all over that. Okay. So I think that's a good transition to the lightning round. Marzyeh, are you ready?

Marzyeh Ghassemi:

No, I'll never be ready.

Andy Beam:

So these are all over the board. Some of them are silly, some of them are serious. It's up to you to figure out which is which. But the only rule is that you have to be concise in your answer.

Marzyeh Ghassemi:

Oh no.

Andy Beam:

Okay. Question number one, what is a core principle that informs your life?

Marzyeh Ghassemi:

There's only a plan A, there's no plan B.

Andy Beam:

Nice. I like it.

Raj Manrai:

If you weren't a professor, what job would you be doing?

Marzyeh Ghassemi:

I think my second career would be a midwife. I really appreciated having a strong midwife who also was clinically credentialed here in Massachusetts for the birth of my children. It was very empowering.

Andy Beam:

I always say that you're one of my most favorite high entropy personalities, and your first two answers have not disappointed on that promise. So question number three, what was the most important thing you learned during the year that you worked at Verily?

Marzyeh Ghassemi:

I think the most important thing I learned there was that collaborations with larger health systems can be very hard to extend beyond limited settings of study. You often really need a very well scoped agreement, and that's not true in all cases when we're working with smaller entities or you have a personal relationship?

Raj Manrai:

Marzyeh, do you believe in the scale hypothesis?

Marzyeh Ghassemi:

This scale hypothesis-

Raj Manrai:

I can define scale... I can define what I mean if it's helpful.

Marzyeh Ghassemi:

Yes, please.

Raj Manrai:

All right. So I'll define the scale hypothesis as the fundamental driver in the performance of machine learning, deep learning in particular has been the size of compute and data used in training models. More provocatively, let's say we can achieve human level AGI by continuing to scale larger models with more compute.

Marzyeh Ghassemi:

Okay, so this is a contentious Moore's law applied to AI, I see. I don't think so. I don't think that we can make it there. And I think part of that is because our way of sampling data is so much more diverse than the systems that we have designed to create and sample and feed data, even the limited modalities that we give systems now. It's hard to imagine ever approximating what an infant gets in a year of life.

Andy Beam:

We'll come back to that later. If you could have dinner with one person dead or alive, who would it be?

Marzyeh Ghassemi:

That's a really hard one. Wow. It might be with a favorite author actually.

Andy Beam:

And who would that be?

Marzyeh Ghassemi:

I have different favorite authors now, but when I was a teenager I actually loved... Before it was in vogue to do historical fiction or retellings of myths. There is a book by C.S. Lewis called Till We Have Faces, and it's a retelling of the story of Psyche. And it's retold in this sort of feminist lens by her sister saying, "I loved her and I wanted her to be with me. I didn't want to sabotage her." So I would love to, if I could get to, I would love to take that perspective and then also couple it with some of maybe the more modern retellings and some of the authors that have also started to reimagine stories now. So I really like Kamila Shamsie's book Home Fire. And this is a retelling of the story of Antigone, which is about the sister whose brother goes to this war and it's branded as a traitor. And so the state will not let her bury his body and it leads to these really tragic outcomes.

But it's told through a modern Muslim family in post 9/11 where a brother again joins a foreign army. And so this is treasonous and the sister can't bury his body. And so I love this idea of taking an old story and making it yours and relevant to your context.

Andy Beam:

So that was an excellent twofer because often we also ask what your favorite book is, and I feel like we got a little bit of that in there also.

Marzyeh Ghassemi:

I also recommend Exhalations by Ted Chiang. I think it's such an accessible set of short stories, to introduce even much younger readers to a lot of these concepts in computer science and AI that we think about philosophically. But if you lose sight of it, sometimes you can lose sight of why it's so important to understand the impact that technology can have on humanity.

Raj Manrai:

All right. Our last lightning round question, do you think things created by AI can be considered art?

Marzyeh Ghassemi:

I think so. I think art is so subjective, right? I think art is something that you experience and sometimes your experience of art is not what the creator intended. Right? You can enjoy something in a different way than somebody had created it to be enjoyed. And I think it's totally fine. I think that one of the exciting things about AI generated art, there's lots of things that are perhaps negative about it that I will not go into as a non-expert in art and copyright law. But I will say an exciting thing is it brings a lot of accessibility. And I think often art, especially as a child of immigrants where maybe art was not something that my family had experience with or access to, I think it brings a lot of access to people, and I think that can be used in a really good way in educational settings.

Andy Beam:

Awesome. Great. Okay. So now we're going to zoom out a little bit and ask you some more big picture questions that are slightly less focused. I mean, feel free to take these in any direction that you want. So I've heard you talk a lot about the importance of open data and having academia being first class participants in research, for a lot of good reasons. So I wonder what you think about big tech companies being the primary drivers of innovation for a certain kind of machine learning research, and especially how that affects machine learning research and healthcare.

Marzyeh Ghassemi:

I think for those of you who maybe remember what life was like before Alexa and Siri, it used to be that at academic labs at universities did all of this state-of-the-art language modelling, right? Or spoken language modelling, I should say. So this is voice work. And there was a posting every year at most major universities where they were hiring people in these roles. That's not true anymore. If you look at two different communities, both the speech community, the machine learning speech community, and the machine learning vision community. The machine learning speech community has been eaten by industry. All those data sets are private, they're owned by Amazon or Microsoft or Google. And so it's very, very difficult now to do state-of-the-art work in spoken language machine learning, without having an intimate connection to industry. And that guides what work you are able to do in a lot of cases, you're not completely independent.

If we look at vision, those data sets are often public. They're openly available, they are used widely, and it's an industry standard to release your data to the point where it would be very strange if your paper was accepted in many mainstream machine learning conferences, or even in many clinical plus machine learning conferences if you didn't release the vision data that you were using. Now you can say, well there's a natural privacy risk from one, and maybe not as much depending on exactly what you're photographing from the other. But I think there's also a difference in ethos here. I think if we in health allow these sort of existing bias, and I mean here human bias, statistical not like a statistical bias but human bias towards wanting to keep things to ourselves and to profit off of a thing that maybe we think is ours to dominate in health data, we're going to see a very similar dynamic where all of the really good research, the state of the art research, the cutting edge research, it has to be done in industry because nobody else is able to get data at the scale that is necessary and perhaps the diversity that is necessary to do machine learning research.

And so while I really am a big advocate for better forms of patient consent and compensation for data and privacy, a lot of the time when people in health say they cannot give you data due to patient privacy concerns, what they mean is they will de-identify that data so it is legal, HIPAA de-identified to sell it to a

company, they can and they will. But they don't want to give it to you for free. And so I think one of the things that's really important is that as we take larger looks at spaces where regulators could perhaps improve performance of models and increase fairness, you could imagine that some agencies could require that data sets are made public or put into archives like the NIH archive so that models can be audited against them for verifying performance in different subgroups.

Andy Beam:

I guess this is a question that I think all three of us ask ourselves at various times, but given those trends, those like macro trends, what is it that keeps you in academia and what is it that makes you think that academia is the best place to do the kind of research that you want to do?

Marzyeh Ghassemi:

Academia is a hundred percent the best place to do the kind of work I think we're all doing. I think the draw of academia for me is that you are your own boss. There are obviously constraints on what you do, both practical and very real constraints on a budget or research directions or things like this. But in the end, you make a decision about what's important to you and what you value and then you're able to execute that. And I think that's extremely attractive. I think also it's very important that somebody who feels beholden to a public of some sort maintains the same or a similar capacity to do this kind of research. In the end, companies are beholden to their shareholders. That's what they're supposed to do. They're supposed to serve those people. They're supposed to increase their profits. We on the other hand, unless we are motivated by something very different than most academics, we're motivated to get research out there to get science out there.

And maybe we have different individual biases as individual researchers about exactly what we study, but taken together as a community, I think we can increase the amount of knowledge about how these systems work, when they work, how best to use them and how best to improve them.

Andy Beam:

Yeah, I think that that probably resonates with me and Raj pretty strongly too. One more follow up question. How do you feel about the dominance of large language models right now? Is this a shiny object that's distracting us from other important problems? Or is this something that deserves the attention that it's getting?

Marzyeh Ghassemi:

I think large language models are a technical innovation that deserve the attention that they're getting. In the same way that neural networks back when they beat the ImageNet competitions for the first time deserved the attention they were getting. It's not that I think the technical innovation is not fantastic. They're very impressive and I think we all agree that they are very impressive. It's that I'm concerned at the level of hammer we are ascribing to large language models. There was this rush when neural networks first worked extremely well in some of the vision tasks to say, and they'll work on everything just as they are. Let's use them for everything. And then that sort of calmed down and variations and innovations in neural network design, architecture design training, that all happened. Right? And now we have, I think a very robust community that does this research. And we're still in that sort of nascent excitement phase for large language models where people are saying, "Just as they are, these large language models are such a pretty hammer and every nail I can imagine they should hit it."

What I'm concerned about is that in the 2000s when neural networks were hot as opposed to now in 2023, nobody suggested initially that those first neural networks that could recognize dogs really well

should be deployed as dog catchers. That wasn't on the table. There weren't startups everywhere saying, "In my dog catching business is this little robot that uses this system." But that's what's happening with large language models in this crazy moment of excitement over this really impressive piece of technology. People are saying, "And let's use it right now. I'm deploying it as we speak. I'm going to press the button." And I think that is the really strange thing for me. the thing that feels like we've gone too far, not that we admire the technology.

Andy Beam:

Right. The technology is good and valuable, but there's perhaps greater scope creep for what it can do relative to previous areas of excitement.

Marzyeh Ghassemi:

Right.

Raj Manrai:

So I think that goes well with our next question, which is maybe now winnowing into medicine specifically, extrapolating from the current trends that you see in AI in medicine, what worries you the most?

Marzyeh Ghassemi:

Do you want a ranked list? What is the-

Raj Manrai:

Two things. Two things.

Marzyeh Ghassemi:

Two things.

Raj Manrai:

Yeah.

Marzyeh Ghassemi:

Two things. The things that worry me about AI in medicine I think primarily are, number one, we do not have well established industry-wide processes for performing audits of models, in ways that evaluate along several different metrics, performance against different subgroups of patients. Because it's fine if a model, for example doesn't work as well on some people. You should just know that, and then perhaps not use it in that subgroup of patients. Right? And we don't do that right now, I'm sorry I think that that's a big concern. It's something I'm really worried about. And something that we can address I think both as a community and then also with different sort of regulatory arms and legislation. The second thing I'm really worried about is all of these startups using technology in ways that they think is going to be perhaps helpful, but ultimately is just going to entrench either poor healthcare or tiers of access to different kinds of healthcare.

Because if you're saying there's some unserved population and we'll just use a chat bot to serve them because they don't have access, that's probably not the best solution. Right? I think part of the concern in health is there are some problems that technology is a very good partial or complete solution to, and

there are many where it is not. And I am seeing now recently many examples of technology being used for something that is a social problem or a human problem, and it will not add. It will only make worse.

Andy Beam:

All right, so final question, and this is one that I'm excited to ask you. I'll ask our sound engineer, Mike to maybe get his hand over the bleep button just in case.

Marzyeh Ghassemi:

What are you going to ask?

Andy Beam:

What is your most controversial opinion? And feel free to let loose.

Marzyeh Ghassemi:

Wait. My most controversial opinion on machine learning, on health, in life?

Andy Beam:

We'll give you the friends and family response. You can say for whatever you like.

Marzyeh Ghassemi:

Oh, wow. That's very expansive. My most controversial opinion.

Andy Beam:

It can be on machine learning, if you prefer though.

Marzyeh Ghassemi:

It can be on machine learning. Wow, I have so many. I feel really bad. Andy, you're not supposed to agree with me. This is one of those things that you tell your friends and they're supposed to say, "No, you? Never." Wow. I am so stumped here. I don't know about most, I think maybe a controversial opinion that I have about machine learning is that, so I think the way that we currently teach machine learning is really stupid. I think that it over emphasizes toy problems and kaggle competitions and performance metrics or it over emphasizes proving theories and bounds. And I think we don't in general or I haven't seen courses or ways of instructing machine learning yet to new students, that demonstrates, look at this field. It is a powerful amalgam of optimization, statistics, more classical AI techniques. And also hardware speedups, vast amounts of data that have ethical considerations. I think we often teach machine learning as siloed communities, but it's something that everybody wants to learn.

The undergraduate machine learning class at MIT gets 400, 500 students per semester, right? And so lots, lots of students are learning about this field, and I think we as a community have not come to a consensus about how to represent ourselves and we tend to still be very siloed within our communities, and I think that that's stupid and a disservice to people. Another controversial opinion. What are yours, Andy? I feel like Andy, Raj can-

Andy Beam:

Hey, I'm the host here. You don't get to ask those questions.

Marzyeh Ghassemi:

I feel like both of you so.... You're such nice people, I can't imagine that you have controversial opinions.

Andy Beam:

Maybe we'll have a bonus hour when we open a bottle of scotch and we can have our controversial opinions said.

Marzyeh Ghassemi:

You have your actual control... I don't know if it's controversial, but I will say, I think that Marvel movies really suck now. They're just terrible. Why are we watching them? Do we care? Do we care about any of these characters anymore? Probably not. It doesn't make sense.

Andy Beam:

No, I agree with that controversial opinion too. I'm with you there.

Marzyeh Ghassemi:

We'll get killed later for it. It's fine.

Andy Beam:

Yeah.

Marzyeh Ghassemi:

The problem is, have you ever had a controversial opinion that you don't think should be controversial? Like women should have the same access to healthcare as men. Or moms should have paid time off, or there should be universal healthcare in the United States. Or universal basic income isn't a terrible idea. Sometimes I feel like I have these opinions and I'll say them and other people will say, "Well, that's crazy. You can't really believe that." And I'll think, I thought this was, I thought we agreed as a community that this was cool. Also, I think we're now at an age, Andy, I'm not sure about you, Raj. But we're at an age where as older millennials, our opinions are controversial for Gen Z. My controversial opinion, Gen Z is that it is okay to require exams in courses. I know that is now controversial, but I think it is okay. I know it's not ideal, but it is the only way we can evaluate short of doing individual vivas for every person.

Raj Manrai:

And this is without GPT4 being at the fingertips of the student who's being evaluated.

Marzyeh Ghassemi:

So far.

Raj Manrai:

Okay.

Marzyeh Ghassemi:

So far actually, on all of the problems that we've designed for our machine learning course, GPT4 does really poorly.

Raj Manrai:

That is a testament to your course because I think it's a pretty unique feature

Marzyeh Ghassemi:

Maybe.

Raj Manrai:

All right. This was wonderful, Marzyeh. Thank you so much for being on AI Grand Rounds. I learned a lot and we really appreciate your time and you sharing your expertise with us.

Marzyeh Ghassemi:

Thanks for having me.