AI GR P19 06.12.24 Adam Rodman

[00:00:00] You can see that we have this weird medical system where it can do more than any medical system can in the past, but people trust their doctors a whole lot less than like the 1950s. I, as a historian, did not see large language models. This is a surprise to me, right? I was very bleak about what the future was going to look like, that it was going to be more reductive, more breaking people into individual pieces.

[00:00:28] And now I see a technology that, again, it's not human, I don't want to anthropomorphize it, but that seemingly understands a lot of these contextual factors and the things that make us human and give us meaning in our own sense of disease and our sense in the world and our sense of community and what disease and suffering means.

[00:00:46] In a technology that could really change the trajectory of where I thought medicine was going, which was a place that, like, in many ways I'm very old fashioned, right? Like, I reflect the values of medieval physicians and [00:01:00] ancient physicians and many modern physicians too, right? There are things that are standard over time, and I see these technologies as a way to get us back to some of those core things of what a physician has done while not losing many of those advantages that come with big data. That come with collecting information.

[00:01:28] Welcome to another episode of *NEJM AI Grand Rounds*. I'm Raj Manrai and I'm here with my co-host and good friend, Andy Beam. Today, we are really excited to bring you our conversation with Dr. Adam Rodman. Adam is an assistant professor of medicine at Harvard Medical School and a practicing physician at the Beth Israel Deaconess Medical Center.

[00:01:49] Andy, I think this was a really fun conversation. You know, Adam is in general pretty fun to talk to. He's a both historian and futurist who has this really interesting perspective from [00:02:00] looking and studying the history of medicine and medical decision making. But now really focusing on AI and what large language models can do for diagnosis

[00:02:08] and in treatment recommendations. He published this great paper that really caught my attention about a year ago on, using GPT-4 to diagnose cases from the *NEJM* Clinicopathologic Conferences. And we dig into that one, but really just get his perspective generally on where this is all going.

[00:02:26] This is really, really fun and full of interesting insights from Adam. Yeah. I love this conversation just because Adam is so hard to put in a box. So, he studied economics as an undergrad, as you mentioned, is a connoisseur of one of my favorite subjects, which is the history of medical AI. So, we got to nerd out about expert systems, uh, in the history of medical AI.

[00:02:45] I'll forgive him for the fact that he went to UNC. I'll give him a pass on that. But for a lot of reasons, it was such a fun conversation. He's so high energy, so full of creativity. And I think that he's so high energy, he hasn't let the health care system kind of grind the energy out of him. He still comes [00:03:00] to all of his research questions and to all of his clinical duties with this sort of sense of wonder and sense of energy and creativity.

[00:03:06] And optimism. Yeah. And it's hard to maintain that in today's health care system. So, it was just a breath of fresh air. Yeah, we tried our best to keep up with him, and it was just such a fun conversation. I think you'll hear that come through in the conversation. I totally agree. And just, would it have been better if he went to Duke or was UNC the, can you explain, I think, give some Raleigh, North Carolina context for our listeners.

[00:03:28] Yeah, so here's a little North Carolinian context here. So, I went to NC State. My wife went to UNC and, uh, for a lot of reasons, NC State folks tend to have a one-sided rivalry with UNC. I'm not sure they really think much about us. So, I always bristle a little bit when I'm talking to a Tar Heel. But even given that, Adam's a great guy and I had a great time talking about the history of medicine with him.

[00:03:49] Awesome.

[00:03:53] The *NEJM AI Grand Rounds* podcast is brought to you by Microsoft, Viz.ai, Lyric, [00:04:00] and Elevance Health. We thank them for their support. And with that, we bring you our conversation with Adam Rodman. Well, Adam, thank you so much for joining us on *AI Grand Rounds*. We're excited to have you. Thank you very much, Andy and Raj.

[00:04:16] Adam, great to see you. So, this is a question that we always get started with. Could you please tell us about the training procedure for Adam Rodman's neural network? How did you get into training? Yes, exactly. Wait, I have two, I have two more, two more sub-questions here. How did you get interested in AI?

[00:04:32] And what data and experiences led you to where you are today? So, we're really talking about pre-training and then maybe a little reinforcement learning. Yeah, pre training and the fine tuning. I'm, I've been fine tuned. I guess everyone is sort of fine-tuned by their harsh reality with the world, right?

[00:04:49] Yes, I think that's fair. I have a very, I think, unusual path towards this field and this research that we're doing, which is that I approach this as a [00:05:00] historian. So, for almost a decade, I think, one of my big focuses of my research has been, and I'm sorry, I know you're gonna make fun of me for saying these words, but on epistemology.

[00:05:11] So, the approach, like, physician's approach to knowledge, especially as it pertains to diagnosis. And then also, secondarily, the concept of nosology. Again, you're gonna make fun of me for this, but how we, like, what structures we use to define diseases. And my focus, my research focus has been really with the era you'd call modern medicine, but in particular from the late 19th century through all the way through the 20th.

[00:05:35] So there was a huge focus on artificial intelligence, right? We've been talking, the oldest quote that I've ever found referencing artificial intelligence comes from 1918 from Bernard Shaw, where he's like talking about what a mechanical physician artificial intelligence would look like. And actually, when he published that in the 30s, he used the word robot doctor, which is funny because the word robot was only invented two years before.

[00:05:57] So these are very, very old ideas. And [00:06:00] if you go back to like the 1940s and 50s, there was this, I mean, I think similar to today, there was this expectation that doctors were over, right? Medicine was soon going to be revolutionized by electronic computers. And that generation of physicians truly thought that, right?

[00:06:17] They were out there proselytizing that computers are coming. So I was, you know this, but I was, I had written a book and I was like working on the proposal for my second book, which is basically about what we're talking about, about the structure of knowledge and how it pertains to diagnosis over thousands of years.

[00:06:32] And then GPT, I had played with language models before, if you guys recall some of the horrific chatbots that were released to the world in the early 2020s. And I mean, in my field, no one thought that these were going to amount to a whole lot at all. So, I remember I got ChatGPT 3.5 very shortly

after OpenAI released it, and I said, well, okay, I have hundreds of manuscripts of historical tests of artificial intelligence.

[00:06:59] [00:07:00] How does 3.5 measure up? And shocker, it didn't do well at all. And then, 4 comes out, and I get it the day it comes out. And I just run a benchmark on one of my own patients through it. And I, my, my jaw hits the ground. Because I realize, right, like, by the historical standards that physicians have traditionally used like stretching back to Internist-1, or even further right you can go back to Ledley and Lusted in the 50s.

[00:07:25] It, with no specific medical training, was outperforming anything that had come before. And, basically, I, at that point, was like, I can't be the only person that realizes this. Hey, Adam, can I hop in? Yeah, yeah, yeah. So, I'm gonna, I try and one up you here on vocabulary and appeal to your historian philosopher sensibilities.

[00:07:46] I think you've started us a little bit in medias res here in that we actually want to hear about how you got interested in medicine. What was the young Adam Rodman like? And what specific things led you to this intersection of medicine and AI? Oh, young Adam Rodman was a pain [00:08:00] in the butt. Old Adam Rodman is still a pain in the butt.

[00:08:03] Yeah. So, when I was a young physician, even a medical student, I was obsessed with the idea of why we did things, right? I was unwilling to accept the world as something that just exists, right? There must be justifications or reasons. And this is really what led me into a path. I studied history in college, but this is what led me into the path of being a historian.

[00:08:23] And initially I was interested in therapeutics, right? Yeah. As an intern, you're like, well. Well, why is this dose of Lasix that's a loop diuretic? Why is this what we do? Is this better than what had come before? And of course, when you pull at those threads, you often learn that a lot of the things that we hold dear in medicine are built on a foundation of sand.

[00:08:42] And the more that you pull at that thread, you discover that there's just so much fundamental uncertainty in our field, which is some of that's inherent to the practice of medicine, but the field of medicine does not acknowledge that, right? There's a lot of what I would call pseudo confidence. In the same breath, we'll speak with the same level of confidence about something that we know really well, as to [00:09:00] something that's basically a coin flip.

[00:09:01] And, again, I am fully accepting that, especially now that I'm getting old, there are just limitations to our evidence, and that's fine. But medicine does a bad job of discussing that uncertainty. So I, we were talking about podcasting earlier, actually started a podcast when I was a resident called *Bedside Rounds* that later grew into an academic history project.

[00:09:22] It's a very good podcast. Just gonna, just gonna plug it a little bit. It's a very, very good podcast for our listeners. Okay, Adam, keep going. Oh, I mean, you know, if you can't get me talking, I'm just going to keep going. Yes. Yeah. And as Raj pointed out, if you listen to the podcast, you can sort of get an evolution of my thoughts about the nature of medicine and how like really our approach to knowledge is baked into that epistemology, right?

[00:09:47] How do we know what we know? How has that changed over time? And really with these interesting threads that stretch back to the beginning of modern medicine about how does the way we define disease, how [00:10:00] does our technology affect what we know and our certainty about medicine? And this, this touches so many fields, like diagnostics, like randomized controlled trials, and of course informatics.

[00:10:10] Yeah, so I was just gonna say, it's interesting that personality type seems to be questioning. You don't take things for granted. At least in my experience, that's not the classical phenotype of someone who's interested in going into medicine. So, what was it about being a doctor? Like, why did you get into medicine in the first place?

[00:10:26] We were just talking about this. I was not a premed. I studied history. After college, I joined an economic policy think tank. And, I hated it. I just sat at a computer all day. I wanted to be around people. So, I made a really rash decision to go into medicine based on knowing very little about the field.

[00:10:43] I also should have done a post bac in retrospect. What I did instead is I was working full time at a think tank and took organic chemistry while working full time. I got a C. Apparently Tulane School of Medicine made a terrible mistake by admitting me, and I've been a pain in the butt ever since. So, no, [00:11:00] that's, I wasn't a pain in the butt when I was a medical student.

[00:11:02] I had to get to, like, residency to really be a pain in the butt. Adam, it's fascinating, because I, I knew a lot about that. So, you know, for our listeners, Adam and I are close collaborators. We've been working together for some time, and I knew a lot about that, but I don't think I appreciated exactly that moment that you decided to go to medical school and what triggered it.

[00:11:21] But a lot makes sense now because I think because of that arc, you have a very unique take on problems in medicine. And I think you question the status quo while also respecting a lot of what's come before and we'll dive into that in the episode. So, I think I want to transition to some of your work and I want to start with

[00:11:41] your recent series of papers on evaluating large language models for clinical decision making. So, the paper that first caught my attention that you published back, I think in June 2023 in *JAMA*. So, this feels like 100 years ago in AI time. This is less than a year ago now, was on [00:12:00] evaluating GPT-4 on the *NEJM* CPCs, the Clinicopathologic Conferences, or also known as the Case Records of the Massachusetts General Hospital.

[00:12:10] And so, I had read a lot of these cases as a quasi-med student, you know, Ph.D. student in the HST program, where we were learning about diagnostic reasoning, how experts think about these cases. And so, I knew the sort of historical significance and also how important they were didactically and how hard they were for many physicians.

[00:12:30] And so I, this really caught my attention because the model performed pretty well, relatively out of the box on these cases. But I was wondering maybe you could start with that paper when you got the idea, what the backstory was for how you came up with the idea and maybe also just give our listeners a physician's take on the meaning of a CPC and the formula and what a CPC is all about.

[00:12:54] I'm going to start by going back almost a century. So, there's two historical trends that intersect in in that paper [00:13:00] and the idea of CPCs. Then the CPC, the clinical pathological conference dates to the early 20th century. It's actually started at, at Harvard. One of the many things we're very proud of.

[00:13:10] And it was adopted actually from the legal tradition. So, this comes from a time when doctors are really starting to get interested in what we today would call clinical reasoning or even metacognition, which is not just what is the right answer, but how do we think? How do we reason? How does new information change that?

[00:13:29] And what are reasonable disagreements? Not necessarily new ideas like the phrase differential diagnosis dates to the early 19th century, but really formalizing that in a process. So, the idea of the CPC, I think it's Richard Kabat, was that the doctor who had a case, and a case that was a mystery at the

time, would discuss that with discussants who would think out loud for the audience.

[00:13:52] And the idea of all of these cases is that it's like a mystery. It's like a Sherlock Holmes novel. There's a whodunit at the end. So, we have a pathologic [00:14:00] diagnosis in the original ones. Obviously now in the 21st century, they don't all have pathologic diagnoses. But in the old days, we would walk through like a mystery case.

[00:14:07] People would say their thoughts, they'd list their suspects, exonerate, rule them in, disagree, come to an answer. Meant to be an explicitly didactic exercise in reasoning, right? The point of it is not so much to teach you about a disease, but to teach you how to think about a disease. So fast forward to the 1950s.

[00:14:25] At this point, the New England Journal CPCs, or the Mass General CPCs, I guess that's what it's called, the Case Records of the Massachusetts General Hospital, but the CPCs had become an institution. Some other journals had versions, but the New England Journal's was the biggest. And you have these two brilliant proto informaticists, Ledley and Lusted, and they are imagining how you would teach a computer to reason, right?

[00:14:49] Because prior to this, let's say you look at the works of Keeve Brodman. He didn't believe that doctors really knew what they were doing when they were reasoning. He was like, oh, it's all intuitive. It's all hand waving. But [00:15:00] Ledley and Lusted were like, well, that can't possibly be true. I mean, look at these CPCs.

[00:15:04] And they sat down, and they tried to use what they called at the time von Neumann's Theory of Games. So, game theory, right? Because game theory is only a couple of years old to model out in these individual cases, how people reason and what they came up with is what you and I today would call probabilistic or Bayesian reasoning, right?

[00:15:22] The idea that when they analyze these cases, doctors, didn't say it out loud, but they intuitively had a pretest probability, and then they were looking at each piece of data that came in, and trying to figure out whether that increased or decreased the posterior probability. And they didn't use the words sensitivity and specificity yet, those had only been adapted to medicine very recently, but that's what they're getting at, right?

[00:15:44] That you could have a sort of Bayesian system, and they also imagined an iterative system where new information would train the model on punch cards, right? They actually have a bunch of cool pictures in the *Science* paper with punch cards. And what they proposed is that because this is how they modeled reasoning. [00:16:00]

[00:16:00] If you built a computerized system that could make medical decisions. What is the best way to test it? Using these formalized CPCs. Because they're so well structured. And basically, since Ledley and Lusted wrote that paper, every single AI system since has used CPCs as the gold standard. So, the most famous one is Internist-1,

[00:16:23] which was developed in the late 1970s. Blackjack Myers, Jack Myers, he was a chair of medicine at Pitt and the president of the AMA, and he had an eidetic memory, right, a photographic memory, and he set about basically recreating his mind at a computer. So, when Internist-1 was evaluated in the *New England Journal of Medicine*,

[00:16:41] it did great, right? It probably outperformed doctors in the early 1980s, but they used the CPCs to evaluate it. And then, really, since that time, if you look at all of the other commercial products, so later QMR, Isabel, DXplain, every single one of the studies used CPCs. So, you got a nerdy historian [00:17:00] who's like, this seems really impressive.

[00:17:01] It seems to, grok medicine in a way that, I mean, I've used both Isabel and DXplain, but they don't. They seem far more like statistical engines. This thing seems to have some sort of, I'm doing air quotes, insight. Just to be clear, this thing right now you're talking about is GPT-4. GPT-4. Yeah.

[00:17:18] So I designed an experiment that would have made Ledley and Lusted happy, right? So, I used some of these old measures. I tried to use the best methodology that I could. And unlike old studies. So, if you look at like the explain Isabel studies or even the original Internist-1 study, someone was manually inputting all of that.

[00:17:37] It took like an hour to put a case and manually inputting all this data. In this case, we just put the acontextual information in the context window of, in this case, we didn't use the API, we used the chatbot. And by the standards that have been agreed upon in the like differential diagnosis generator field, right out of the box, GPT-4 performed, it tied the best existing system, which of [00:18:00] course had been like trained over a decade and had tons of information.

[00:18:03] And this thing could just do it like that. Can I hop in here and ask a question? Because you just made me realize something that I, I've never thought about before. So, uh, you know, Dxplain, Internist, all of these things were essentially like an idealized Bayesian version of clinical reasoning. And, when you talk about what should the ideal physician do?

[00:18:22] That's kind of put up as they should be perfect Bayesians. They should integrate and condition on any information available. But like, actually what I hear you saying is that mimicking the messy process of human reasoning, where we're being intuitive and we're grokking actually works better in the real world.

[00:18:38] So should we give up on that ideal Bayesian doctor? Like, like what should we take away from the fact that we did this bottom up, instead of top down. Yes, that's exactly it. Um, no, we shouldn't, I think, well, clearly there's a role for Bayesian reasoning in certain domains, or even as my friend Shani Herzig likes to say, Bayesian reasoning is the cherry on top of the sundae of clinical reasoning, but it's not the sundae.

[00:18:59] And I [00:19:00] mean, we've known this for a long time. So, a lot of these early informaticists knew this. So, one of the famous examples is, um, Elstein is doing all of these wonderful studies, like interview studies at Harvard in the late 70s, and he's looking at medical students and he's looking at attendings and having them talk out loud through cases.

[00:19:16] And what he discovers is that both the med students and the attendings show the same thought process, right? They're all little Sherlock Holmeses, they're asking questions. Going after different differential diagnoses, but then a weird thing happens, which is that the attendings ask like five to 10 questions and get the answer.

[00:19:32] And the medical students go on for like half an hour and still don't have a very good differential. So, it really, it's at that point in the late 70s, and this is like the era of Kahneman and Tversky, that Elstein realizes there's something else going on. And nowadays we know, like it's an organizational

[00:19:48] idea of reasoning, which is that there are ways that human beings organize information in their head. We call it script theory, and that basically from the second we start hearing about a patient, we go [00:20:00] down certain routes. And it starts to develop these schema, so these comparisons of different scripts to give us a an idea, oh, this could be a pulmonary embolism or this could be metastatic cancer in the lungs that further drives our questioning.

[00:20:13] So yes, I think what is special, and this is cutting to the chase, what is special about language models is that they're a bottom-up model of reasoning that really mimics the expert clinician system 1. What they do not do well is any of these sort of deliberate metacognitive strategies, most notably Bayesian.

[00:20:30] Like, they can't do Bayesian reasoning. Of course, me and Raj had a study that showed that their implicit understanding of Bayesian reasoning is better than humans. So, they can't understand Bayesian reasoning, but they can't understand it better than we can't understand it. Um, but yeah, so that's what's new about this technology.

[00:20:46] Also, what's new is that its inputs are textual, like contextual text information, meaning that they can be scaled. One of the things that I always like to tell people is, look, guys, in the early 1980s, we had computers that if you [00:21:00] inputted it appropriately, could reason better than humans in difficult cases.

[00:21:04] Ah, Raj, we talked about problem knowledge couplers, like what Larry Weed developed, and in the domains in which it worked, it was incredibly impressive, but you would just sit in front of a computer doing like, yes, no. If you could structure the information in a certain way. But you know, you what's so fascinating about the literature is if you go back to, I dunno what year this was, maybe it's 70s, maybe 60s, Pete Solovitz and Steve Palka, right.

[00:21:28] Are starting to question systematically looking at those systems and enumerate the limitations of the sort of purely probabilistic Bayesian approach to automated diagnosis. So, they have this really good paper, categorical and probabilistic reasoning and medical diagnosis, where they outline the data hungriness, the combinatorial explosion of probabilities that you actually need to estimate to have this

[00:21:51] purely Bayesian approach to the world. And then I think also sketch what would be probably, pretty intellectually aligned with what you just said about script theory, [00:22:00] which is how categorical, or this kind of fuzzy, this fuzzy set of rules or winnowing of the space of possibilities is a critical first step before we then launch into the purely Bayesian approach to update decision making.

[00:22:14] I think just to defend the Bayesians, right? Like, you know, there is a, in the decision theorists, right? There is a lot that we learn about, I think, utilities and how to extract information from the patient and how to extract

information from groups of individuals that we don't really talk about in even the current wave of excitement about large language models.

[00:22:32] So what is a patient utility? What is the risk preference? What is a tolerance for the patient? What are their goals? I think there's some of our historians and pioneers of medical decision making have really thought very carefully about that. And what I personally would love to see is us, maybe not necessarily operationalizing it the same way, but injecting that wisdom and that patient preference, maybe via large language models, systematically into decision making and into maybe even the medical record, right, Adam? [00:23:00]

[00:23:00] Maybe there's a path to that. Yes, you're trying to, you know exactly how I feel about that. Yes. Okay, so I think this is a good transition. So, one thing you did really special in that paper in *JAMA*, the CPC's paper, where you evaluated GPT-4 is, and you alluded to this, is that you did, something nuanced in your evaluation, right?

[00:23:21] So you evaluated the correctness of the differential diagnosis, not by a multiple-choice type accuracy. There is no such thing for those differentials, right? For the CPCs. But you use this validated, psychometric. And again, I've started talking like this because I'm spending a lot of time with Adam these days, but well validated psychometric to grade essentially how accurate this model was with respect to previous incarnations.

[00:23:46] So that one was the bond scale, right? Then just to transition to another one of your papers, you used another psychometric, which maybe you can tell us about and also how you see this field evolving. This is a recent paper in *JAMA Internal Medicine* led by Steph [00:24:00] Cabral and you. And you guys used what I understand is the R-IDEA scale to evaluate the correctness and the diagnostic reasoning abilities of GPT-4 with respect to human physicians on a common set of cases.

[00:24:14] So maybe you could tell us what it takes to build and create a well validated psychometric and how you see this field evolving for evaluating large language models and human reasoning as well. I was gonna say what a psychometric is. Yeah, a psychometric sounds very intense and scary. It sounds like from the 1960s where you'd, what was it the CIA was doing experiments on people with LSD?

[00:24:36] What is that? MKUltra, right? So, it sounds like something from that. And I guess it sort of is, right? A psychometric is an evaluation that is

meant to evaluate something that goes on in the human mind. Of course, we, Raj, you and I have joked about this. We have no way to see into the human mind. It's very Cartesian.

[00:24:53] Maybe I'm the only, sentient being here, and you guys are hallucinations that the demons have put in my head. However, because we can't [00:25:00] see into people's minds, we have to rely on external scales. So, a psychometric is a way of evaluating that, right? Classic psychometrics may or may not be valid, like the Myers Briggs test, right?

[00:25:11] Or IQ tests. All of these are purportedly psychometrics to tell us something about what happens on the inside of the head. So, psychometrics in reasoning have been used for 20, 30, let's say 30 years in medicine, um, and they've been used actually for a fairly important, though not exciting from an AI perspective, which is the teaching of clinical reasoning, so medical education.

[00:25:39] Clinical reasoning curricula really entered medical schools in the 1980s, right? Prior to the 1980s, the focus of medical school had really been on knowledge transfer. You just need to know everything there is to know about diseases. There's actually a famous ethnographic study from the 60s where they actually go over this, the boys in white, and they don't teach them anything about how to think, like you just learn by modeling other people.

[00:25:58] In the [00:26:00] 1980s, first, the advent of the computer, right? People are like, well, you don't need to, or personal computer, you don't need to know everything, you can look it up. So, these are people who are way ahead of themselves, right? This is like the Apple II and the Commodore 64 era, but they're seeing a future where you can look everything up.

[00:26:15] And also knowledge generation and subspecialization had taken off that no medical student could know at all. So, a focus in medical school started to switch to how to learn and how to think. In order to teach people how to think, we needed to have ways to grade people on how to think. So, a lot of very smart doctors and really led by cognitive psychologists started to evaluate expert clinicians as well as medical students on what good thinking actually looked like.

[00:26:42] A lot of these are psychological theories that would sound very familiar to anyone who reads like Malcolm Gladwell or all the pop psychology books these days. It wasn't pop psychology back in the 80s. It was really boring experimental psychology, but that's how it goes. And the psychometrics began

to get used in medical education, [00:27:00] and depending on which psychometric we're looking at, got more and more evidence as time went by.

[00:27:06] So when Steph and I wanted to evaluate both the reasoning of humans compared to the reasoning of computers, we hit the literature. I already knew the literature, but we looked like to see what would be the best tool to do this, and we selected a psychometric called R-IDEA, which is a psychometric The R is revised, but IDEA, which is a psychometric that looks at the presentation of reasoning as it can be gleaned from a written document.

[00:27:34] And we chose this in particular because Verity Schaye and her team at NYU had done a wonderful study validating this just a couple of years ago, where they used a BERT model to read something like 30, 40,000 resident notes and correlate the strength of the psychometric with the cognitive load and also the quality of the resonance.

[00:27:55] So a psychometric for reasoning currently and probably [00:28:00] from here on out is the best way that we have to evaluate human reasoning in order to establish validity. You have to have both face validity, right? So, you need people who are experts say, well, this makes sense. But then you have to establish validity in populations, which means the population you want to study. You need to run it on a lot of people. You need to make sure that, for example, there's a dose response, like as people gain expertise, it goes up. That you can see different subgroups, and that there's some degree of both consistency in grading it and internal reliability.

[00:28:30] And psychometrics, this is an old field. The classic one is a Likert scale, right? The 1 to 5. Likert scales go back to the 1950s, and it's one of these things that we take for granted, right? I walk into a bathroom in the airport, and it wants me to read it on a scale of 1 to 5 on how my experience was. And what's really funny is you can look at arguments against Likert scales in the 1950s, and they're still the case today, right?

[00:28:50] We haven't, we use these things commonly. That doesn't mean that there's big problems in using them to rate things. These more validated psychometrics are an old older tool, but one [00:29:00] that, well, as you know, or as I hope you believe, is increasingly important as we now have language models that can at least mimic the thought processes of humans.

[00:29:10] So I don't want to detract because I want to hear about the results of the study, but I'm struck by the fact that it sounds like we're doing machine psychology. And I wonder in what ways human psychometrics are informative

or misleading for machines. So, like this guy, Eliezer Yudkowsky. Talks about this all the time from the perspective of existential risk.

[00:29:30] But he's like, what they're actually doing is they're able to cosplay or put on masks and they can simulate any kind of human in any kind of scenario. And so, while you may be getting insight into the simulation of the particular context that you're using the large language model, you're getting very little insight into the actual underlying

[00:29:47] large language model itself, which may be important when the model is used in cases it wasn't tested for. So, I wonder how you think about that dichotomy. So that's great. So, both of those things are simultaneously [00:30:00] true. And this is what's really crazy. They're true in humans too. So, I'm going to give you the human example.

[00:30:06] So, back in the early 90s, we had this brilliant cognitive psychologist, who's still alive, named Bordage, who did research on how expert humans reason. And what he discovered is that when expert doctors reason, they, and this is right when script theory is starting to be developed, but they look at these things called semantic qualifiers.

[00:30:26] So, for example, and I had a patient like this last week, I was on service. If one knee is swollen versus multiple joints are swollen, so monoarticular versus polyarticular, it really changes how I think about that problem. If it came on in 12 hours versus two months, it changes how I think about that problem.

[00:30:42] We now call these things semantic qualifiers. So Bordage, when he described this, was looking at the difference between experts and novices. Now, in medical school, we teach people to use these words from the first day of medical school to describe their problems, right? So Bordage was looking at a world [00:31:00] where they weren't taught this, right?

[00:31:01] These are just emergent abilities. Now we are teaching people to display the presentation of reasoning. And what's really cool is Bordage's study, the smart guy, also was practicing for like 40 years, so lots of time to see things change. And what's interesting is we can see that using the presentation of reasoning, especially in novice learners, does not actually mean they are reasoning better.

[00:31:21] So it's very similar to the problem with the language model, right? They are cosplaying an expert clinician, but does that mean that they are

actually reasoning? No. They're not alive. They're not humans. What's interesting, and the reason I think it's important is for two reasons. One, I think that, well, humans are going to be interacting with these, right?

[00:31:41] So the question becomes not, is the model actually thinking in this way, but how does the model cosplaying an expert clinician affect the human that's using it? And then number two for scalable oversight. So, you use these psychometrics to train the models to evaluate human beings. How does that work, right?

[00:31:59] Can the [00:32:00] models actually use this understanding, this cosplay of reasoning to evaluate actual human beings? At the end of the day though, we end up in an ouroboros, right? Is the, all of us interact with the world via text or speech. And they're just estimations of what's going on in our head. So, this is a problem that, this is why I love these technologies, right?

[00:32:20] We're talking about robot psychology. Obviously not the same as human psychology, but these problems exist for humans as well. Do you identify with Dr. Susan Calvin from *I, Robot*, robopsychologist and Isaac Asimov series? Robopsy- I, I was gonna say I'd like to become a robot. Would you be Chief? Chief? Yeah. Chief.

[00:32:39] Chief robopsychologist of, of a- For a second I thought you were asking if I identified with the Roombas. Yeah. Oh, yes. Yes. Do you identify with the vacuum cleaner that moves around your house? You know, my, my kids when they draw pictures of our family, they draw me, my wife, the two of them, our dog, and also Roomba.

[00:32:59] [00:33:00] Amazing. Amazing. Alright, Andy, should we go on to the lightning round? Let's do it.

[00:33:15] So you've listened to the podcast before, Adam, so you know how this goes. So, I'll spare you the introduction. The first question, which superpower would you rather have for a day? Invisibility or the ability to fly? Um, well, first of all, I think anybody who says invisibility is going to be up to no good.

[00:33:32] So I'm going to go with the ability to fly. Well, so that's funny. Cause that was my answer because the ability to fly exists already as a technology. So, I would want something that gives me a differential advantage, but I guess that means that I'm, that that's my villain origin story. Okay. Yeah, see, that's the difference.

[00:33:44] I mean, it would give you a comparative advantage, but I think it's difficult to use that ethically. Whereas flying, I mean, I don't think that, what's the unethical use of flying? Adam, if you weren't in medicine- It's a good question. It's a good psychometric. It's a good psychometric. [00:34:00] Oh yeah, the lightning round of psychometrics, I like it.

[00:34:02] Adam, if you weren't in medicine, what job would you be doing? Oof. Either a cognitive psychologist or a journalist. Science journalist. Yeah, on brand, on brand. Yeah, I'm pretty predictable. Okay, so here's your chance to be controversial. Me? Which medical specialty could be most easily replaced by a large language model?

[00:34:23] Which medical specialty? Yeah, so that's the, I think I have a unique perspective on who and what can be replaced because I think it's less of a specialty and more of the cognitive tasks of that specialty. So, what we're looking for is a specialty that relies mostly on textual interpretation, with less formal reasoning and certainly no procedures.

[00:34:46] So, I think a lot of people want to say things like radiology, well, radiology might be a language model, but radiology or dermatology, but I think some of these specialties, so like say oncology, right, where it's [00:35:00] effectively looking at large amounts of data from clinical trials and then individualizing that to patients.

[00:35:07] I don't think that oncologists are going to go away, but I think that a lot of those treatment decisions that they now make are going to be taken over by language models. So other things like nephrology, maybe even rheumatology in my field in medicine. So, anything that relies on a large amount of text input, text output. Any of the messy fields, right, where there's more epistemic uncertainty.

[00:35:28] So like my job, a general internist, the primary care doctor and, uh, inpatient medicine doctor, those are really, really messy. So, I think it'll take longer to replace us. Is that a good answer? It's interesting. Yeah. I wondered if telemedicine plus psychiatry was going to be where you would go with that, but I think.

[00:35:45] Oh, well, this is great. Maybe so psychiatry is fascinating, right? Because psychiatry is this field that has a different nosological model than everybody else, right? So, in psychiatry, diseases are defined purely on symptom basis. Whereas in the rest of the field that [00:36:00] based on pathological anatomy, there's a pathophysiologic change.

[00:36:02] And there's been this really weird trend, right? Like tertiary syphilis, general paresis of the insane used to be considered a psychiatric disease until we had actually malaria therapy. So that Nobel Prize in medicine in 27 was giving malaria to people to cure syphilis, but then penicillin and then it became a medical disease.

[00:36:20] So like you peel off things from psychiatry when they have a pathologic basis, which is really interesting, right? Because you could say, well, maybe psychiatry will be easier to replace because it's purely a patient based nosological model. But also, maybe the squishiness of those borders, the natural grayness of those borders are going to more inherently require a human being to navigate that rather than a language model, because you can't tell the language model where bipolar one and just hypomania begin, because they're human creations.

[00:36:52] So I don't know. Adam, I think I know the answer to this question, but I'm probably, I wouldn't be surprised if I'm totally [00:37:00] wrong. Who is your favorite historical doctor? Who do you think my answer is? Osler? No. No. Okay, I'm wrong. I'm wrong. I got a surprising answer. Okay. Do you know who - you're way wrong.

[00:37:16] Do you know who Pierre Louis is? You do, of course. Tell us about Pierre Louis. Pierre Louis. Yeah, so, so Pierre Louis is, well, for you guys, he's like the first proto informatician, proto epidemiologist. So, Pierre Louis is, I'd say, in the second generation of modern medicine. So, he's practicing in the 1820s and 1830s.

[00:37:37] And he is, uh, friends with Laplace. So presumably, we don't know, there's a paucity of primary sources from this, but presumably he's hanging out with these nerdy, you know, maths guys, and is like, well, there's no reason that I can't do this in people. So, in the 1820s, he does some studies on, like, the distribution of yellow fever to, like, basically saying, like, look, if [00:38:00] I look at a lot of different patients and I collect objective data, I can discover ways,

[00:38:05] and you have to keep in mind, he has no modern statistical methods, but I can discover ways just using simple arithmetic that I can help my current patients. And then in the late 1820s and early 1830s, he decides to be a pain in the butt. So again, I love anybody who wants to be a pain in the butt. In Paris at the time, there was an obsession with leeching.

[00:38:24] So there's this famous doctor, Victor Broussais, the vampire of Paris, that was his nickname. And people were getting like 50 leeches at a time, having liters of their blood removed. And women had dresses in the style of leeches. It was like an obsession, especially in Paris. And Pierre Louis was like okay,

[00:38:42] so we're saying this is a great therapy, especially for pneumonia. Yeah. What if I actually look at the data, and he had his little notebook where he kept all this data on patients. And he said, look, if bloodletting is so great, then in my patients with pneumonia, if they got bloodletting early, they should live a lot longer than the people who got the lifesaving [00:39:00] therapy late.

[00:39:00] And he just sat down, and you can see his journal still exists. The simple arithmetic that he went through, and he was, like, shocked at the end. He was like actually, the people with the bloodletting are dying at a much higher rate than the people who get late bloodletting. This doesn't make any sense, but Pierre Louis, being a wonderful pain in the butt, just went out and published this and accepted the large amount of blowback.

[00:39:21] And it's a great example. He didn't really have that much impact at the time, but people were not ready. Modern statistics didn't exist. They were like, that's cool that you added up some numbers, man. Who cares? Again, it's the 1830s, but he had this small core of dedicated followers who would go on to invent, like, regular epidemiology.

[00:39:42] So, Pierre Louis the prototypical doctor, who's a pain in the butt, combined great patient care with a knowledge of population medicine and statistics, and really brought about the modern world. Nice. No need to apologize. I think the Mark Cuban episode already got us our TVMA rating. I think that [00:40:00] was a great episode by the way.

[00:40:01] Andy, I want to ask just one wrinkle on that question. So that is not the answer that I was expecting, and I love it. And it's a great story in so many ways. Maybe I could just ask a follow-up lightning round question. I don't think we've ever done this, but who is your favorite historical fictional doctor?

[00:40:17] My favorite historical fictional doctor. Oof, that's a good one. Or it doesn't have to be historical. It could be contemporary fictional doctor. From Dr. House to, whoever. Oh, can I do a Star Trek doctor? Can I do a Star Trek doctor? Yes. And why. Yeah. My favorite fictional doctor is, oh God, I'm such a nerd.

[00:40:36] So I'm going to apologize in advance, but that would be The Doctor. So, the hologram doctor from Star Trek Voyager, because he is a computer doctor who learns what it means to be human. Classic Pinocchio story. Nice, nice. To show you that I can also do in context learning, I'm gonna do a, I'm gonna call an audible here and not ask you the question that I had planned to ask you.

[00:40:57] Given that you're a hardcore medical [00:41:00] history nerd, I want to ask this overrated, underrated question. So, Hippocrates. Overrated, underrated as the figure in the history of medicine. Overrated. Can you say a little bit more why? Oh, I mean, so there's a move. He's got good press. So, there's a, first, Hippocrates wrote only a tiny fraction of the documents.

[00:41:20] There was a whole school called the Hippocratic School. That stretched like hundreds of years. Many of which he was not alive, so. And a lot of the Hippocratic ideas are actually ideas that were popular generally around the Mediterranean. So, Hippocrates just got really good press, right? Great ideas in terms of naturalistic medicine, but we have papyri from ancient Egypt that show that they were practicing naturalistic medicine like 1,500 years before.

[00:41:45] So overrated for sure. Awesome. I knew you'd have a great answer for that one. Glad I asked it. Alright, Adam. You got more? I'm ready. We have one, one last one, and I love that I get to ask you this. What would Michel Foucault think about ChatGPT? [00:42:00] Oh, yeah. What would Michel Foucault think about ChatGPT?

[00:42:03] Foucault would, all the post structuralists, like Derrida and Baudrillard would all look at ChatGPT and say this is what we were talking about, right? This is the final stage of human cultural evolution, which is computers mimicking humans that are being consumed by computers mimicking humans and being evaluated by computers mimicking humans.

[00:42:25] So I think the post structuralists would have, I mean, Foucault is a very pessimistic man, so been very pessimistic about it, but also not surprised. Excellent. I think, Adam, congratulations. You survived the lightning round so we have just a final few questions here. And these are more big picture concluding questions.

[00:42:44] So thoughts to leave us with and to leave the listeners with. So, I've seen you do this personally and I've been very impressed by it and so I was hoping you could really give us your philosophy and your approach here which is there's a lot of medical students [00:43:00] and residents now including many

of whom you're training and you interact with who are very interested in artificial intelligence.

[00:43:06] They want to get involved and they're wondering what the best way is for them to learn about AI and to start pursuing research for example in AI. So maybe for the med students and the residents, what do you think is the best thing that they can do right now, if they're interested in, and getting involved in medical AI?

[00:43:25] So that's a great question. And I mean, to hedge a little bit, it depends on their skillset, right? If this is a, if you have like a Ph.D. in computer science before this, it's going to be like basic computer science research, but generally we are in a time where what we really need in this field is collaborations between the people building the machines and the people who understand what it is to practice medicine.

[00:43:50] And in particular, the invisible parts that you can't see, like what's going on inside the human mind and inside patient minds. So, my advice for any resident who [00:44:00] wants to get into this would be to focus on what you really bring to this, which is your medical knowledge and your medical expertise and partner, right?

[00:44:09] Partner with computer scientists, labs that are doing this research to bring your perspective. I generally disagree with this idea that we need doctors to learn like the principles of machine learning. Like I use CT scans all the time. From my perspective, a CT scan is a magic doughnut that a patient comes in and I get a cool image out of.

[00:44:25] I don't need to understand the physics of a CT scan in order to better take care of my patients. I do not think that people need to understand like the complex architecture of neural networks in order to understand how to use a language model to take care of their patients. And from a research perspective what, if anyone is listening, what you are bringing is your understanding of the context of medical care.

[00:44:45] Is there anything specifically you would change about medical education to better prepare students for this AI future? Well, uh, I just so happen to be leading the task force at Harvard to figure this out. It's a great question because the fairest answer is no one knows. [00:45:00] At this point, a lot of the, and I don't know if you guys would agree or disagree with this, but a lot of the things that have come out of language models are surprising and, like, this word is overused, but emergent, right?

[00:45:09] Things that are not necessarily programmed in. And I have an old person brain that was trained, even, like, you know, the Internet, I had computers pretty early, but I was trained, my brain was trained in really an analog world. And the new generations that are coming up used language models in college. My children had ChatGPT tell them stories this morning.

[00:45:31] So they're using language models from the ages of three and five. So, I think that right now, from a medical education perspective, the best thing that us old people can do is keep an open mind, look at how our learners are using it. I personally think a lot of the benefits of language models, and Raj knows that I think this, are like, unexpected, right?

[00:45:51] I don't think the benefits of language models are going to be, oh, look, I have a scribe that can listen to me. Oh, it can write my notes. I think a lot of the benefits are going to be, oh, look, all this [00:46:00] highly contextual text that we produce can now be understood in mass and give us feedback. And from a learner perspective, like, oh, look, I can have a personal tutor who can tell me what I'm doing well and not, and give me personal teaching.

[00:46:11] So, just keep an open mind. Generally, I think we should focus on those things that make us most human. So, I think communication skills navigating difficult situations and reasoning, especially other metacognitive strategies, because like, I think the future with language models is not going to be like what the techno futurists think, oh, what is it?

[00:46:33] Martin Shkreli created Dr. Gupta, right? Like it's anyone who uses that. I assume people are just using it to make fun of it, but it's a terrible idea. But language models are going to have a huge impact. It's just going to be a lot weirder than any of us predict right now. So let me ask the compliment to that question.

[00:46:48] For a long time, there was this annual contest called the Loebner Prize, where people would write a chatbot that would try and kind of pass the Turing test. Every year, they would actually give out another thing called the [00:47:00] Most Human Human Award, which is there were human contestants in this, and whoever seemed the most human would win this.

[00:47:06] So. What should human doctors focus on as AI comes more and more into the clinic and is more involved in clinical practice? What would be the equivalent of the most human human prize for a physician? Yeah, and this is a great question to complement the journal that we all work for, *NEJM AI*.

[00:47:25] Everyone read John Chen's recent piece on the, uh, so, my gut feeling is to say communication skills through difficult situations. But then I read that John Chen piece, right, where he had a really difficult conversation about a patient with advanced dementia who was going to get a PEG tube, and he thought it was a bad idea. And the conversation with the wife went very poorly, so he decided to recreate that conversation role playing. And then he discovered that, I believe that ChatGPT said was like something like, I understand that your husband's a fighter, but there's many different ways to fight. That might be doing everything, but fighting might also be the [00:48:00] bravery to go for comfort at the end of life or something like that.

[00:48:03] And I feel the same way that Jonathan did when he had that experience, which is like, wow, this is more human than human. So, Andy, that's to answer your question, like, I don't know, right? We've taken for granted that like sitting at the bedside holding someone, obviously only a human being can do that right now because we don't have robots, but some of this empathetic communication, even if it is just a simulacra of human caring, seems to be effective.

[00:48:27] So I don't know. And this is from the guy who's trying to do this at Harvard. Um, but before we head to the last question, I would just like to say that you have won the vocabulary award for *AIGR* guests so far. I feel like the level of Latin has been higher and the number of SAT words has been fantastic.

[00:48:41] Is it simulacra? Yeah, that and lots of great SAT words in this episode. So, so I guess like just zooming out, why are you doing this? What about this whole AI endeavor gives you cause for optimism, and where do you hope it's going? Well, it's funny that you think I'm optimistic. Raj, am I [00:49:00] optimistic or pessimistic?

[00:49:01] Uh, I think you're optimistic. I kind of agree. I think you're an optimist.

[00:49:09] Yeah, so why am I doing this? If you look at the grand trends of history, and let's just stretch 200-something years, right? There has been a trend in medicine since Pierre Louis on breaking human beings down into data, right? The word data itself in medicine really only comes from the 1980s. We were using things like the facts of disease before this.

[00:49:32] But to understand humans by breaking them down, which, to be clear, has allowed for great advances in medical care, because we can run advanced clinical trials, because we understand lots of sciences better, we can

do amazing things. We can knock out genes and put a pig kidney in another human being, for God's sake.

[00:49:51] Like, that's amazing. But there have been real costs to this approach as well, in the dehumanization of patients, in the isolation of [00:50:00] humans from their medical care alienation, you can see that we have this weird medical system where it can do more than any medical system can in the past, but people trust their doctors a whole lot less than like the 1950s.

[00:50:12] So people truly feel like isolated and put apart from their medical system. I, as a historian, did not see large language models. This is a surprise to me, right? I was very bleak about what the future was going to look like, that it was going to be more reductive. More breaking people into individual pieces, and now I see a technology that, again, it's not human, I don't want to anthropomorphize it, but that understands, understands, air quotes, seemingly understands a lot of these contextual factors and the things that make us human and give us meaning.

[00:50:45] In our own sense of disease and our sense in the world and our sense of community and what, what disease and suffering means in a technology that could really change the trajectory of where I thought medicine was going, which was a place that like, [00:51:00] in many ways I'm, I'm very old fashioned, right? Like I reflect the values of medieval physicians and ancient physicians and many modern physicians too, right?

[00:51:08] There are things that are standard over time. And I see these technologies as a way to get us back to some of those core things of what a physician has done, while not losing many of those advantages that come with big data, that come with collecting information. So, AI putting the humanity back in medicine, if I was gonna do some summarization there.

[00:51:29] Yeah. And I'm also deeply pessimistic that people are going to use these technologies the wrong way and make our lives much worse. So that's what, that's why I'm such a crazy person doing all this work because they're powerful and what I see happening right now with LLMs being rolled out, why are we using LLMs to draft the first message to patients?

[00:51:47] Isn't like, communicating with patients one of the most fundamental things that humans do? And this is what we're using these powerful technologies for? So, I'm also motivated by, if I can swear, like a deep fear that we're gonna fuck this up. I mean, [00:52:00] but isn't that interesting? Is the reason for that because that's the part of the job that physicians enjoy the least?

[00:52:04] Or why is that the point of entry there? Well, it's not that it's talking to patients is not inherently what physicians enjoy the least. It's because we've built a system in which physicians have been converted into a data entry clerk. So, because of that, now communicating with patients has been reduced to like this.

[00:52:20] You know, sitting in your pajamas at the computer doing that. So, of course, this is America in the 21st century. So instead of rethinking the system that makes us miserable, we're like, hey, let's make a computer do this important part. Got it. Adam, this has been fascinating and thanks so much for coming on.

[00:52:36] It's been like a pretty broad ranging discussion, and I know I certainly learned a ton. So, thanks again. Well, thank you guys for having me. And to think I was trying to like limit my vocabulary usage. Yeah. Thanks for taking it easy on us. Yeah. Yeah. Oh, Raj knows what I'm like. Adam, that was, that was [00:53:00] amazing.