AI-GR Podcast 16 03.19.24.Eric Horvitz

[00:00:00] Microsoft Research came out of a discussion between Bill Gates and Nathan Myhrvold. He was the Chief Science Advisor to Bill Gates. Rick Rashid was pursued as the Founding Director. I think that was 1991. We were brought in as the first AI group. There was a specific call-out tool, gains and contributions we had made with probabilistic or Bayesian inference.

And that approach resonated deeply with Nathan and with Bill, per what Nathan was telling us at the time. When we came up, we were probably within the first 10 people, three of us, at the lab. There was a group that came over from the IBM Research Center in Natural Language Processing. There was a small compilers team at the time.

Microsoft Research grew very quickly in its first few years. We spent a tremendous amount of time recruiting, and [00:01:00] it was interesting, I remember back then, to even think about. What would it be like to build a lab? We established certain principles. For example, this amazing principle that there would never be any controls on a researcher's decision to publish anything they thought was important to publish.

That would not be, there would not be a tower of reviews on that process, and that has stood the test of time. That really has given the labs quite a different feel than you'd have in maybe a traditional corporate R&D environment. The lab grew to comprise all areas of important areas of computer science over time.

And people at the lab would do a mix of applications as well as foundational work.

Welcome to another episode of *NEJM AI Grand Rounds*. I'm Raj Munrai, and I'm here with my co-host, Andy Beam. We have an amazing conversation to share today, and it's with Dr. Eric Horvitz, who is the Chief Scientific Officer of Microsoft. [00:02:00] Eric told us about his work at Microsoft, which goes back several decades, including how he got started at the company all the way to his work on AI today.

Andy, I know this word is overused, but I think Eric is truly unique in being both a card-carrying M.D. and a Ph.D. computer scientist who works not just at the intersection of the two fields, but who's made fundamental advances to both

fields on their own. As one example, he wrote the classic machine learning paper on a Bayesian approach for identifying spam email.

And he's leading efforts at Microsoft now to understand the strengths and limitations of large language models like GPT-4 and medical diagnosis. We had a really wide-ranging conversation spanning medical AI, decision theory, and computer science, and it was a lot of fun. He's also done really interesting things like the 100-year study on AI, where he gave a donation to Stanford to help them study AI over a century time horizon.

That's the kind of long-term forward-thinking projects that have been the hallmark of Eric's career. He's also on something called PCAST, which [00:03:00] is a White House level organization that advises the president. And one of their key mandates is advising the White House on sensible AI policy and sensible AI regulation.

The number of areas that he has touched in computer science, medicine, and public policy has been very impressive.

The *NEJM AI Grand Rounds* podcast is brought to you by Microsoft, VisAI, Lyric, and Elevance Health. We thank them for their support. And with that, we bring you our conversation with Eric Horvitz. So, Eric, thanks for joining us on *AI Grand Rounds* today. It's great to be here. Thanks for having me. Eric, great to have you on *AI Grand Rounds*.

So, this is a question that we ask all of our guests. Could you walk us through the training procedure for your own neural network? How did you get interested in AI? And what data and experiences led you to where you are today? It's an interesting reflection for me to go back and think about all that.

I have always been interested in mechanism. [00:04:00] I've sought explanations for things in the world, how do they work. And that led me to doing a biophysics degree as an undergrad. I was excited about learning about physics and biology, chemistry, how they come together in various ways. And during that time, I ended up getting more and more curious about this interesting black box we call.

Nervous systems, in particular human nervous systems, intrigued me, what the heck was going on and who were we? And that led me to a neurobiology lab and I did a deep dive and you know, when you're an undergrad looking at neurobiology, you get placed into very specific segments of research and questions.

And I was becoming expert at looking at single neurons, unit activity. building microelectrodes, pulling them, as they say, and sticking individual neurons. And in dark rooms, I'd watch the clicks and clacks of single neurons. And that raised my [00:05:00] curiosity even higher. How on earth did even a trillion or a 100 trillion of these little units generate this fluid conscious experience that people have?

And then when I was applying to grad schools, I pursued a Ph.D. in neurobiology and thought I would add an M.D. that would give me more depth and the future possibility of working more with the grand challenge of understanding human minds. And by the time I got to Stanford to do an M.D. Ph.D., I was already so intrigued at the modeling approach, the AI approach to getting around the mysteries and complexities of dealing with many, many neurons and trying to build a larger understandings from the activity there.

And so, I was taking lots of courses on main campus, grad courses in AI. And before long, I moved my Ph.D. over into AI at the foundations at the principal's level. And then later came back to Ted Shurtleff's lab doing AI and medicine. That seemed like a nice [00:06:00] unification, even though I stayed on the, more on the principal side, probability theory, decision theory, and so on, of principles of bounded rationality.

But I found an interesting direction, which was how do you apply formal methods, bringing decision theory and probability theory into what was then the dominant paradigm of AI in medicine, rule based expert systems, as they were called. Those are such a lot of excitement about those systems, but they didn't seem to handle uncertainty very well.

And by pushing into probability theory and trying to bring that back, I would say, to AI, I mean, AI research ended pushing me into understanding better what we meant by bounded rationality, systems that can't possibly do it all. They can't possibly finish thinking when it's a hard problem, like a traa care problem under time pressure.

What do they do? And coming up with a formal method for how they would back off and become bounded rational in a way that was justifiable became kind of a passion of mine. And that led me into broader areas of AI. Can I, can I hop in here and [00:07:00] ask a question? Cause I'm super intrigued by the sort of scientific origin of your interest in AI.

So, you were trying to understand from a neuroscience perspective, how intelligence arises, I wonder if you see any parallels. in what's happening now

with trying to do interpretability of large language models or other big complex, there's a movement right now called mechanistic interpretability, which is essentially applying like neuroscience principles to large language models.

How promising do you think that is in light of your experience in neurobiology? Well, popping up a level, it's interesting. So, here I was making, in some ways, a crisp decision at a point in my career that I would not be pursuing the complexity per the black box of nervous systems, heading into the world of Bayesian networks and more generally probabilistic graphical models where you knew the semantics, you knew the procedures, you knew every arc and node.

Getting into machine learning, it got a little bit more complex and black boxy, but still [00:08:00] understandable. And as we got into the, the, the world of neural models, neural network models. And now today it's interesting to come full circle and say, well, Eric, you couldn't escape, you're back in this world again, where you're grappling with complexity of the form we can't understand and explain very well.

Some of the most interesting behaviors that we're seeing now, some people refer to them as the emergent behaviors of, of abstraction, generalization, and composition, that we see, for example, with GPT-4. So here we are back again. And now I find myself. And recently working on a paper that's coming out at iClear, you know, where we're thinking, looking at what happens, looking at the actual activations of single neurons and patterns of neurons when a system can't answer a problem versus when it can.

And I had to like, giggle to myself saying, you know, Eric, there's probably going to be no easy way into the world of understanding deeply mind, or even our small versions of. clickers of minds we're [00:09:00] seeing in these large scale language models now. And so, tools like that you referred to just now as mechanistic analyses and explainability, understandability I think will be very, would be more and more important.

It's almost like here we are in. 2024, if you ask me back in 1989 during grad school years, what would I even be doing in 2024, let alone what it looked like. *The Jetsons* in that futuristic world. This idea that we'd be back again, and I didn't escape my pursuit of clarity and crisp understandings with the complexity we've now encountered, and also the incredible capabilities we're seeing at the same time.

I would be shocked probably to hear to learn about this back in 1989. Yeah. I thought that hearing about your early days, the symmetry just popped out to me from where you started to where you ended up, everything old is new again. So I don't know if that is exciting to you or frustrating that this [00:10:00] problem has followed you over the many years, but it seems to be like you, you are kind of back to where I guess I would say that as we pursued the path that I was on, we'll call it for now, probabilistic and decision theoretic systems, where we would, you know, handle the representations.

Bayesian networks, influence diagrams, foundations of utility theory in medical AI systems, for example. Not to geek out the listeners here, but you know, the idea of doing cost benefit analysis under uncertainty, doing diagnosis with methods we understand quite well. And by the way, those methods have reached a point where they're quite mature and we should not forego looking at them very carefully in their applications to health care while we're in the superheated time about neural language models.

But that to the side, and there is a sense of excitement, to answer your question, in that it didn't seem we were going to get to anywhere deeply close to what we see in, like, vertebrate [00:11:00] nervous systems on the path that I was on. We have methods that will do tremendously well in being competent collaborators when it comes to health care, sort of more precision in our diagnosis and our therapies with those methods.

But, now with what we're seeing now, I'm not going to say that we're necessarily getting closer to how minds work. But let me just say, I see a path to being surprised about large scale systems and interactions with tremendous amounts of training data and certain principles by which we train models to discover new, surprising capabilities that I didn't see on the path that I was on.

So, I'd say I'm surprised in that way. Yeah, it's certainly easier to experiment on a virtual brain than it is on a real one when you're trying to understand how it works. I guess too, I'd like to understand a little bit more about the M.D. side of your training. So, I think clearly you have an interest in intelligence and [00:12:00] AI.

It sounded like the medical training was in service of those scientific goals. But did you do residency, for example, and sort of how has the clinician side of you informed your career? Yeah, when I first came to My, my, doctoral work, my M.D., Ph.D. program, like other medical, first- and second-year medical students, I dove in with my class, you know, a full vibrant member of my four-person cadaver team in anatomy, neuro, you know, physiology classes.

But at the same time, I was like staying up late at night and spending all my day also going from my anatomy class, smelling of phenol into my AI class work. And people looking at me like, what's that odor on your body? Where are you coming from, man? In my experience, that odor is not that uncommon in computer science classes.

So maybe you weren't that obvious. Maybe a different odor, but anyway, but back, back to the question. I stopped out after like a couple clerkships to really go full bore my doctoral work. I got so into it, like so [00:13:00] into it. Most people in the world still know me as a researcher in AI, and many are surprised with it. Wait a minute, you have an M.D.?

And the story with that is that I was, got more and more interested in clinical medicine through my AI applications and talking to physicians, the experts on teams and so on, getting more immersed into that world from the point of view of trying to build systems to do, to be competent and to be helpful, on those challenge problems.

I finished my Ph.D. work. In the meantime, a colleague and I started a startup company that was actually taking some of our Ph.D. work and actually making clinical tools on this new thing that was getting more powerful called a PC. That company evolved into a larger scale company that we started taking our medical principles of Bayesian networks and decision making to other areas, to United Airlines for jet engine diagnosis and to NASA for space shuttle monitoring and so on.

And [00:14:00] right around this time, Microsoft This strange company to the north of the Bay Area reached out with Nathan Myhrvold, a close associate with Bill Gates. And we heard news that we're trying to start a research team and we're really interested in the research you're doing and in your company and so on.

Of course, we were stunned that, you know, why would this company that makes Windows 3.1 and a word processor and a spreadsheet be interested in even a research group and let alone AI technologies. That's a whole other story about where Microsoft Research came from and our work there when we came, came up, but we ended up being acquired and I promised my two colleagues I would stay long, no longer than six months.

That was 31 years ago. And when we made that deal, I hadn't finished my clinical clerkships. I was so passionate about my Ph.D. work. And I told the folks at, at my two colleagues, Jack Brees and David Herkerman, [00:15:00]

hey, look, you go up immediately. I really want to just, I I've gotten more interested in clinical medicine.

I want to just experience the whole, the whole, I want to do a deep dive into, into this area. I had like 10 months to go to finish up everything, given my earlier clerkships. And I just stopped, dropped everything, got completely focused. And I remember even like the tension between like writing, finishing up writing journal articles based on your dissertation, with like holding a clamp during surgery, during pediatric surgery, and thinking, oh, do I really want to be here?

But I started really getting into it as an, as maybe a more mature person. Like I started to love clinical. I wanted to excel in every rotation that I did, you know, get those top-notch recommendation letters and, you know, recommendation for house staff when you come back and all that. But in the end, I did a very rare thing for most M.D. training programs.

On my last day of my last rotation, for my commitment, it was, I remember it was September and the sun was setting on a [00:16:00] kind of cool fall day. I got into my car, went to my locker first, I mean, got my, all my stuff and I drove away without doing a residency. And I always felt like, oh, like it's really difficult to do that kind of thing.

But I made my decisions where I'd focus my attention. But per your comment, the clinical experience has proved to be extremely valuable, especially when it comes to, and I don't mean to. I know both of you are, didn't have your M.D., but you're doing lots of work in AI right now and AI health care. When it comes to talking to AI scientists who are largely in their labs, like they have no concept of some of the actual real-world problems.

And, you know, why is it so hard to like to translate this really cool technology into practice? To having been there and understanding daily life and the workflows, it gives you a different appreciation for, like, how hard it is to really innovate and to bring [00:17:00] innovations into the clinical realm to make a big difference.

I totally agree. And I was fortunate. So, I did a Ph.D., not an M.D., but during the Ph.D., it was kind of a medically themed Ph.D. in this sort of unique program between Harvard and MIT called HST, where we take a good chunk of the first two years of preclinical coursework. And then. Even more importantly, spend two summers in the clinic, taking histories and physicals, rounding with the teams, presenting cases.

And I think I really, now I'm looking back on this, if and when we are clinically relevant, I attribute so much of it to sort of that time that was spent just learning about how decisions are made and working with clinicians, building real collaborations. And so, you know, you have this, you know, I think, as Andy said, very unique background.

You're a card-carrying M.D. and a serious computer scientist. I think it's a pretty rare breed. I think it's, it's interesting to think, especially in GPT 4 era, post GPT [00:18:00] 4 era, what education for quants and AI researchers to help them be clinically relevant will look like. I think that's a good point to actually transition to some of your work, Eric.

So, we want to dig into your work at Microsoft and your research there. And I think you mentioned this a few moments ago, you, as I understand it, you helped to start Microsoft Research and you've been at Microsoft for a few decades now. So maybe we could start with telling us about the founding of Microsoft Research and some of your work on foundations and applications of AI and maybe human-AI interaction as well as some of the themes that you work on.

Yeah, so Microsoft Research came out of a discussion From my understanding, between Bill Gates and Nathan Myhrvold, who was, you might call him the CTO of the time, I don't think he was given that title at the time, but he was, you know, Chief Science Advisor to Bill Gates, and they have a beautiful [00:19:00] set of slides, which I think are available, which was like, why is a research lab important?

What would be the nature of Microsoft's research center? And so on. That was remarkably interesting. they, they were remarkably important slides and thoughts for how to, for framing Microsoft Research. Rick Rashid was pursued as the Founding Director. I think that was 1991 when we, our team was approached by Nathan Myhrvold, who was doing, you know, passionate recruiting of core initial teams.

We were bought in as the, the first AI group. There was a specific call-out tool, gains and contributions we had made with, with probabilistic or Bayesian inference, and that approach resonated deeply with Nathan and with Bill. Probably what Nathan was telling us at the time, when we came up, we were probably within the first 10 people, three of us, at the lab.

There was a group that had been, came over from [00:20:00] the IBM Research Center in Natural Language Processing. There was a small compilers team at

the time. Microsoft Research grew very quickly in its first few years. We spent a tremendous amount of time recruiting, and it was interesting, I remember back then, to even think about. What would it be like to build a lab?

We established certain principles. For example, this amazing principle that there would never be any controls on a researcher's decision to publish anything they thought was important to publish. That that would not be, there would not be a tower of reviews on that process. And that has stood the test of time.

That's really has given the center of quite a, of the labs, quite a different feel than you'd have in, you know, maybe a traditional corporate R&D environment. The lab grew to, to, to, to comprise all areas of important areas of computer science over time. And people at the lab would do a mix of [00:21:00] applications as well as foundational work, combinations thereof.

On our team, we started out being called the Decision Theory Group because we were so excited about that at the time, as a foundation, including probability. Eric, can I love to dig really into decision theory, but I suspect that many of our listeners actually don't have too much familiarity with decision theory.

So could I just interrupt you for just a moment to give maybe our clinician listeners, other folks who aren't as familiar with decision theory, an overview of what it is and just your definition of decision theory. So, there's decision theory and there's decision analysis, which is the engineering real world incarnation of how you apply decision theory on real world problems.

It goes back to the idea that the foundations of modern approaches to what are ideal decisions goes back to probability theory. [00:22:00] The axioms of probability, and on top of which, the axioms of utility or utility theory, and that's another set of assertions about preferences, about what's good and bad in the world, like what are desires under uncertainty, but an important test area or an application area for decision science more generally has been medicine, hard medical decision problems where you face a set of outcomes under uncertainty.

You have a set of actions that one might take. There are different likelihoods of different outcomes happening following each of the actions that you might commit to. These are irrevocable, commitments, decisions in the world. And the idea is, how do you discover the best action to take when there are cost benefit tradeoffs, or the outcomes are quite different, potentially great uncertainty?

And so, there are processes by where you frame the decision problem, [00:23:00] which is, what am I trying to do here exactly? What are the, you

know, what's the, what are the goals? What are the key possibilities? What are the actions possible? What's the disease process in this case, in a particular case, for example?

What are all possible outcomes of each action that might be taken? And then for each outcome, even before you get there, you can sort of try to really push on a patient's preferences. Like what does this mean for the patient? If I do this kind of prostate therapy, surgery, radiation. What are the tradeoffs?

And then how do I choose the best action? Especially given that there'll be uncertainties in what happens because you can't know for sure. So, decision analysis is the engineering approach to taking those principles of decision science or decision theory and bringing them to life. And they typically involve notions of estimating probabilities, computing expected values [00:24:00] to come up with the best decision, which is typically an expectation given the uncertainties.

A lot of us do this kind of thing qualitatively on the, in the clinic. just by looking at literature, thinking through the key actions and outcomes we care about and then talking with the patient to understand preferences and looking at best practices, and protocols and making a call. Typically, it should be the patient's decision, of course.

But sometimes you have hard problems that you really are hard to, you need to have work on paper, paper and pencil. And when it comes to bringing AI to bear or to leverage harder. AI technologies, we want to have systems that can give us estimates of the probabilities of the outcomes and then understand how to encode preferences of patients, for example, and then propagate them through to tell us what's the expected value, expected utility of each action that I might take, and let's pick the [00:25:00] best one, but make sure that that really resonates with what the patient has in mind for preferences.

As we get into talking about large scale language models, one of the interesting challenges is what's the role of these models when it comes to working in a decision theoretic or decision analytic way, which is the classic best practice or approach for hard decisions and making calls on the best action to take.

And that's going to get, that gets into these questions about can these large scale language models really give us probabilities, for example, that are credible, that are well calibrated. So, questions are coming up at the, you might say, at the intersection of traditional best approaches for how you harness AI to do hard

decision problems and what these systems now can offer us in terms of their powers we're seeing, which are remarkable in themselves.

So, I think when I first seriously got into decision analysis and decision theory in grad school, I was just blown [00:26:00] away that there was so much thought and that it was several decades old when I was starting to go through this and how to formally Reason, over diagnoses, and over patient utilities, formally elicit utilities, and then come up with rational decisions and also collect information in a very principled way, right?

Even what tests to order, what, what utilities need to be measured, things like this. And then I was struck, so I remember discussing this with Zak, who's my Ph.D. advisor and, you know, our Editor-in-Chief of, of *NEJM AI*. I remember discussing with him, I was like, you know, like, why, why isn't all of medicine like this right now?

Like, why aren't clinicians formally eliciting utilities and estimating probabilities and applying Bayes rule and doing this and that? And you know, it started this sort of multiyear discussion that has of course, not resolved, but is extremely entertaining. I think also informative about the sort of difference between some of the theoretical frameworks and [00:27:00] then the practical demands on the clinician who has a very limited amount of time and is maybe informally doing some of these things, but not approaching with all this machinery.

And so, I have to, you know, I, I think when we think about the sort of population health level guidelines that are set by some of the national organizations, I think they do take very formal decision theoretic approaches to making clinical recommendations. But for this sort of individual clinician seeing the patient, it seems that large language models that GPT-4 and its cousins may actually solve some of these fundamental problems that have seemed very elusive, right.

Or at least offer a path to solving some of these problems around both the provision of information, although of course there are many limitations, as well as the kind of interaction with an individual to collect utilities in a formal way. Do you see decision theory potentially having some type of revival with the advent of large language models?

You [00:28:00] know, we have work to do. I should say that my team got access to GPT-4, an early raw version of it, it's part of our responsibility to do safety studies for Microsoft. And we started, that's when this work that led to this

paper called “Sparks of AGI” came from. We started looking at this very, these models.

And of course, you know, you can imagine I would be diving in with all sorts of hard challenge problems in health care. And I even pushed the systems back in those days. Just on the topic of your question, can a system be told, you know, the chain of, I’ll use these fancy terms now, chain of thought reasoning, or other methods, other prompting methods, to think with Bayes rule, to explicitly, to use probabilistic reasoning and tell me what it was doing on a worksheet.

Could it then do decision theory? Could it do what we call hypothetical deductive reasoning in a loop, where we look at the symptoms? It’s a phrase that’s been used, that’s used over the years for, look at some [00:29:00] initial signs and symptoms, formulate a differential diagnosis, list of diseases by likelihood, based on that, compute, you mentioned this earlier, you know, the expected value of new information.

What’s the value of, and sort by the expected return of collecting new information, given the cost of making that collection and then getting the new information and going in a loop, hypothetical deductive loop. Could I drive GPT-4 to do this kind of thing? And did some early experiments in this. It was so interesting to even push hard on getting this, the systems to openly show their knowledge of probabilistic reasoning because they’ve, they’ve read about it.

And they’ve learned about that kind of thing. And to, you know, keep track of prior probabilities and posterior probabilities, you can imagine these classic chest pain examples. Forty-five-year-old white male, no history of cardiac illness, comes in clutching his chest, pain that he’s never had before in his [00:30:00] chest.

I’d like to also try to put it in her chest, because we’re going to look at gender issues and, statistics and so on, but having the system write down its reasoning from the point of view of Bayesian updating, Bayes rule and so on. And the system, you can see it struggled. You can see steam huffing and puffing, but it was making progress in this space.

It wasn’t always correct in some of the calculations. We know that’s not the strength of these systems, but to me, I’ll use the word again. Spark. There was the spark of the prospect that someday these systems could explicitly be tuned and trained and call other tools, for example, and they needed to do calculations that couldn’t do themselves.

Preference assessments to become fabulous companions as decision analytic consults. And so, I'm hopeful you can tell that our team, we're still exploring. I'm personally still exploring the possibilities. [00:31:00] I'll ask questions at times, you know, to even open AI. Hey, can we have these systems better calibrate their probabilities about X or Y?

And can we get into the log probs and see how we can sort of do some research to make these systems more competent at Bayesian diagnosis. I think, let's listen to this podcast 15 years from now. I'm guessing that we'll see an incredible synthesis of what we call traditional Bayesian inference and decision theory, decision analytic consultation, and large scale language models.

I mean, even now there are some, some basic things we can do that are just. no brainers, like framing your decision problem. What am I not thinking about? What more options might I have? People often say in a decision analytic problem, whether it be in health care or finance or in public policy, that a new option or a new consideration of an outcome can [00:32:00] dominate the whole analysis.

So on that front, if we believe that And I think we're seeing signs that large scale language models really can be, let's use the phrase, mind expanding for humans, for what they can do with their, I'll use another word, with their polymathic skills, their ability to compose and synthesize, and bring in real information and ideas and distinctions that one was not thinking about when they first started sketching out even a traditional decision tree.

So yes, I think there'll be lots of touch points. So, you have a term that I think I've seen you use before instead of the history of present illness, right? It's the history of future illness, right? Yes. Which is this sort of simulated possible future states, which is both a wonderful potential educational tool. But I think also a new way for a physician to consider maybe things that they haven't considered, that don't necessarily [00:33:00] align with their differential so far.

So that's super interesting. So, you know, you mentioned also that your group has done work on evaluating GPT-4 on medical challenge problems. I think you had a preprint. Maybe it's amazing also just how fast this field moves, right? How the progress has moved from January, of last year to now, but I think it was a preprint in the first half of last year on GPT-4 on medical challenge problems.

You had found that GPT-4 performed very well on USMLE-style questions. I think you had exceeded the sort of state of the art at the time. And then perhaps,

you know, I thought that paper was very interesting and perhaps even more interesting to me was your group's work just a couple months ago on prompting and how effective something which seems simple, but is increasingly, you know, we have this very rich evidence base is extremely important for eliciting the behavior out of these models that we want for a given task of how chaining together a few [00:34:00] different techniques into this kind of meta technique that you call Medprompt really unlocked certain capabilities from this generalist model, GPT-4 on challenging medical questions.

And so, anyone, you know, I've also noticed that. GPT-4 users and other language model users, they sort of seem very strong kind of bubble effects where some people are using the model every day for hours a day. They almost have a relationship with it, with the AI. And other folks have used it once or twice, maybe tried to use it as kind of a Google device to look themselves up, were not impressed and didn't really use it.

But anyone who's used it a lot knows that prompting really matters and how you sort of craft your interaction of the model really matters. So, from that standpoint, it's not surprising, but I think what's surprising. When you start to learn about it is what you actually do.

And so, it's some things like saying, think step-by-step, right? Which you call chain of thought, or I think for another one of the models, it's take a deep breath. Take a deep breath. That's my favorite. Yeah, so I, I literally, you know, Andy and I [00:35:00] are both fathers of, of young, young children. And it's literally the same advice that we give on a new task that our daughters are learning.

I think my six-year-old was jumping in the pool and I was like, take a deep breath and think about how you're going to do this. And it's literally what seems to work with some of these models. So maybe we could just take just a couple minutes and then I think we want to jump into sort of your efforts on responsible AI, but maybe, just on that Medprompt paper, could you just tell us about, maybe briefly about the background of the paper and then what your findings were, and then maybe also just like signal where you think this field is moving.

So, what those results mean for what's coming next. Yeah, so the, the background on that paper is, of course, teams, the team under my office, the Microsoft Research teams, we're very close with, have been looking. You know, since we've been playing with large language models at the power of, in some ways, [00:36:00] the power of, of how we communicate with these

models, it's so interesting, given how they work, to understand how powerful it is to set them up to be generalizing or synthesizing or abstracting.

Composing based on subtleties in language of how you describe who you are, your role and what it is that you want and how you characterize the nature of what would be a good, answer or output in this whole dialogue as you get, as the dialogue continues. I think Greg Brockman from OpenAI has put it in a very pithy way.

He said, a surprising amount of AI research is getting large language models to be in the right mood. And saying the right prompt and coaxing them to sort of be in the right headspace is where a lot of the engineering effort actually gets applied. It's interesting. English is the hottest new programming language.

And it's, it's amazing how powerful language is. We're [00:37:00] discovering the incredible foundations of how concepts are embedded in how language is used and the meaning of words and how they're strung together and then encoded and then represented and then reasoned with by these models. It's, I'd say, yeah, the right mood or the right mode.

Even saying things like, one way that I've used GPT-4 lately is as an expert editor. So, I like to write my own stuff. I don't like to have the system generates like the content for me, but I don't mind having the system and actually, you know, sort of enjoy, and I'm very thankful when something's an important docent that I've, I've written or a few paragraphs that I want to do a post to take a look, you know, to have GPT-4 take a, take a look at what I've done.

And the way I do this kind of human-AI collaboration session is to tell GPT-4, you are just such a talented, insightful editor in how you can take my material and really find places, precision places where you might want to [00:38:00] refine it and it can even be better than it is. Put those comments into angle brackets.

Don't touch my text directly but give me expert editorial remarks because you're just so great at this. Really, really thanks. Thanks so much. You know, the expression of, this camaraderie and appreciation for the kind of capabilities the system has, I haven't done a formal study of this, but I find myself not doing that just because I'm anthropomorphizing, because I want to squeeze top notch performance out of these systems, and in some ways, per Greg Brockman's comment.

It's putting the system in the right mode slash mood. and I think we'll learn more about this over time, but what's interesting in the Medprompt work as we call Medprompt is, and you said stringing together or composing. We're at a point now where we have notions of few shot, prompting or learning, different ways of doing that random, based on similarity metrics.

Then there's the idea of using a [00:39:00] chain of thought, like reason about your steps. Then there's the notions of ensembling and shuffling and ways to combine different answers all in a single prompt to come up with the best answer, looking for consistency. And so, what we did with Medprompt is we took several of these known methods and did a very careful layering of them and then going backwards, ablation, ablating them to see what was each component now adding to the power or the accuracy of the results we were getting, for example, on the MedQA, very challenging medical challenge problem benchmark.

The Medprompt Prompt work itself was kind of a fun experience in that it came out of what we called the Medprompt Marathon. I think I talked a little bit about this publicly where we just said, hey, we have a bunch of smart people. Let's get together. And we're going to have teams and when it's going to really go for it with some dedicated resources, we can get fast answers, sometimes getting, having a cluster so you can really [00:40:00] cycle fast, helps you think and be creative.

It was interesting to see how far we got with that to be the top scores on not just the medical challenge problems, but we went to this interesting large scale set of benchmarks called MMLU, which has challenged problems in philosophy and law and Electrical engineering, psychology and accounting.

And we said, wow, this Medprompt thing with these layers seems to be pretty general purpose. Well, we'll still call it Medprompt cause that's its roots, but what it's kind of learning about, you know, what's the, how in 2024, what's the. What are some best practice approaches to talking to these models? Some of the magic that we discovered with the Medprompt work was to actually use the model itself to do chain of thought and to generate the few shot reasoning, you know, write your own chains of thought, as examples.

We seem to be as, [00:41:00] as good or better than humans could do with creating sample chains of thought, as examples. That was pretty exciting to us. And it suggests that, which is coming to the fore now in multiple research projects, that these models can play multiple roles at different phases of problem solving and prompting.

You know, and this in some ways frames work now on these multi agent solutions like AutoGen, where you have, it's really the same basic language model, but you're giving it different roles as a programming construct and telling you're the critiquer, you're the generator, you're, you're going to, you know, work, check in with the human, and so on.

And building sets of agents to take on various aspects of the problem solving. But one more thing about the Medprompt effort, you know, we were seeing a lot of special case expert driven models going on, where we explicitly in our initial So, in our paper in the first half of last year, our original work [00:42:00] on the USMLE challenge problems, we explicitly said, we're not going to try hard.

We want to basically show you how powerful these language models are by just, just talking generally and asking for answers to these questions the way anybody might ask without doing any hard work. And that was the magic of what we were demonstrating. You know, then we saw some competing groups saying, hey, we're doing a lot better than you, we worked really hard at this, we brought in 30, 35 experts.

And we said, you know, we were like sitting on a chaise lounge, smoking a cigar. And that was the point, but you know, let's, we'll lean in now. We'll, we'll do a little marathon, and we'll just explore the space of what you can squeeze out of these models by talking to them properly. And that's what some of the background there.

What's your intuition for how much farther we can go, Eric? So, if you, I think you got up to 90% there. Do you think just prompting on GPT-4 can push us [00:43:00] another, another 5%? There's something to unlock in the model that we haven't unlocked in, in prompt space. You know, I, I'm, I'm sure there is, but the question is what's the, what, what are the margins of return, with effort right now?

Yeah. I think when it comes to these large benchmarks, like MMLU, all the medical portions of it, or MidQA, at some point, some of that remaining headroom is not going to be better thinking, but we discover, oh, that's really a bad question, or those answers were inconsistent in the benchmark. So, we're going to, we'd find some froth that's really not going to get better by being more brilliant, but by debugging the actual benchmark itself.

Nice. So, I'd like to transition to your work on the responsible AI. So, you've, you've had both a private sector hat and a government sector hat in both of

these, in this effort. I'd like to sort of touch on both of those. [00:44:00] So continuing with your work at Microsoft, I don't know whether or not this counts as Microsoft or not, but you've stood up this 100-year study of AI.

So, I'd love to one, hear about the genesis of that. What we hope to learn from a 100-year study of AI and what you hope comes out of that. Yeah, the 100-year study on AI, by the way, it's, we call it that just to It's kind of catchy, but the endowment to Stanford that our family did to stand up the 100-year study on AI is to have a report written with proactive guidance to government, civil society, the public, academia, every five years for as long as Stanford exists.

So, as John Hennessy told me, we can guarantee this will go on as long as Stanford exists. At the time, he thought that would be a quick fix. A pretty long time. So, it's probably, hopefully it's more than a 100-years. The background, the genesis of that study was that when I was president [00:45:00] of the Association for the Advancement of AI, this is the largest society of professional and scientific researchers in the world.

The triple AI, we're having the upcoming main conference coming up in a couple of weeks that I'll be at in Vancouver. But when I was president in 2008, I made the theme of my presidency, because it was just coming, AI is coming of age. Even with applications like the readmissions work we were doing at the time, for example, in real hospitals and testing things out.

I made the theme of my presidency, AI in the open world. And when I gave my presidential lecture, which is still online, it was all about, like, we have to sort of think through the principles and mechanisms. It gets back to bounded rationality, but how can we design our systems to do well in the open, in the scruffy, uncertain open world and still be robust and reliable?

And I talked about different ways to do that and so on. For the last part of my talk, I said we also, we have to also start thinking about AI people and society and its influences in the open [00:46:00] world. And some people thought that was like way over the top in 2008, but I called together and I announced at that lecture a study that would go on for several months that led to a meeting at Asilomar, for symbolic reasons, we went to Asilomar, called the Presidential Panel on Long Term AI Futures.

It was just this great study with top-notch people. You can go online and read about who was there and so on back in, in 2008 and 2009 because it spanned those two years. And in 2014, it was five years later, and I said, we should do

this again. That was so useful, but things are changing so quickly, even back then.

How can we do this again? And of course, as a computer scientist, we think about induction and N gets N plus one, and how can we do this forever? Every five years, you just have to establish the base case and exactly just watch out for recursion. But anyway, so we, we basically went to Stanford. The development office [00:47:00] thought we were a little bit out of the box and called the president.

And they were all kind of wishy washy. I don't know if you can do this or not, and we can guarantee this will go forever. John Hennessey, who I've known for many years back when he was teaching, when I was at grad school, said to me, Eric, this is a great idea. Let's just do this. You know, it's funny cause then you just maybe seven or eight years later when we were standing up, and I was helping on the advisory board for the human-centered AI, HAI program at Stanford, we had a big opening dinner and John was at my table and he looked at me and he said, remember back a few years ago we all thought this was crazy?

Well, I guess it was just a few years ahead of its time. But anyway, that study is now going into its third report, I recently mentioned, God, five years seemed like an appropriate, you know, base cycle back in 2013. When we stood this up, maybe we should speed this thing up. I mean, in terms of the recurrence on that.

If you go online and go to the 100-Year Study's site, you can go to my initial docent, what's called the [00:48:00] Framing Memo, where I listed 18 issues, challenges, and opportunities that I thought would stand the test of time. Like 100s of years. So maybe folks, your listeners can go out there and take a look at those.

We'll go in and check on those annually. And see where we are with these things, right? Cause they're all kind of like, I thought really raised questions about possible futures, you know, in 2013, and I think they still do anyway. In part, it turns out that that study, and it's the news about that, came back to Microsoft.

And people started talking about it at Microsoft. I mean, discussion in email, Satya and Brad Smith mentioned we were doing this. And we also talked about what's Microsoft's point of view on responsibility in AI. Satya called together a meeting and we started thinking together about what were our principles about AI as a company.

And this work led to what we now [00:49:00] call Microsoft's AI Principles. There's six of them and the standing up of what was called the Ether, it is called the Ether Committee, which stands for AI Ethics and Effects in Engineering and Research, which ended up defining and scoping out our earliest approaches to what, what the company would be doing when it to our responsibilities when it came to the development and fielding of AI technologies.

Before we get you to the lightning round, I want to ask about your role on the President's Council of Advisors of Science and Technology or PCAST for short. So, this is a not specific to AI council, if I understand correctly, but a large focus of PCAST recently has been on AI and you've helped advise the Biden administration on the executive order they released.

So, I guess to the extent that you're able to answer this, how is the federal government thinking about AI and AI regulation? So, the President's Council of Advisors in Science and Technology, PCAST, is, is, it just has been a fabulous [00:50:00] experience for me in terms of the colleagues and collegiality and the various projects and working groups that I've been involved with or contributed to.

One of the projects people out in, I'll call it NEJM podcast space, would find very interesting is we finished. And made public after about a year and a half of work, our report on patient safety, calling for a transformational effort, effort in patient safety with set of recommendations to the president.

So, we do a variety of projects. I'm just coming off of co-chairing a project called Cyber Physical Resilience for the Nation. That just was posted just last week. And you'll see that there's quite a bit going on with opportunity in that space. Yes, AI lit up the White House with interest. It lit up the Office of Science Technology Policy, which, which is the housing organization for PCAST.

There's been a great deal of engagement. The [00:51:00] PCAST group briefed the president directly more than once on AI. I've been very much involved with those briefings, and we were engaged at multiple levels, including levels of collaboration that, had influence on the president's executive order on AI.

There's both concern for the need to regulate and to guide, but perhaps even more so. Interest in making sure that America and the world more globally harnesses AI for great benefit. So those are coming together right now. And in some ways, there's a sense that if we could appropriately regulate the

technology when it comes to safety, for example, in rights and some specific concerns like let's say biosecurity, related issues.

We're freer to innovate and to get the most on the positive side of AI innovation at the frontiers. I'm going to try and, be a real podcast host [00:52:00] here and tie this back together to something you mentioned earlier, which is decision theory. So, you will see people on both sides saying that there's infinite benefit from AI, but also infinite harm.

And so, it's almost kind of this weird Pascal's wager situation where doing a sensible cost benefit calculation is almost impossible by definition. So how do you actually reason about cost benefit of AI, if not in the short term, then over the long term, when you have these kind of infinities that crop up according to people on both sides of this debate?

Well, I think many people who are working in this space aren't necessarily. pinning their needles to infinities. There's somewhere more in the middle and looking at, for example, the downsides as potential rough edges versus the end of the world, it comes down to, you know, our democracy works in that we, and our elected officials, and advisors get to work together in a [00:53:00] collaborative manner to think through the various inputs and possibilities to study them.

It takes wise committees. And I say wise, because you can't necessarily converge on the ideal answer, but you can do kind of a decision analysis and say, look, we're seeing these kinds of benefits, here are some concerns, maybe we can monitor them over time and learn about them. Maybe we could check back.

You can do some limited previews. We can do red teaming. What are some techniques that would provide some fail-safe dimensions to AI technologies and how they're being used? What are the key concerns from the point of view of different stakeholders? For example, if you ask me, what are the biggest concerns with AI going forward?

The two that come to mind for me are the two challenges are the flooding of the use of AI to flood the world with, with synthetic media, including nefarious uses of the technology to persuade, and to [00:54:00] impersonate, to generate propaganda. These are, these are I think are significant threats to democracies.

There was a recent example of a fake Biden robo call in the primary in New Hampshire, where they cloned his voice and had a very convincing robo call of

telling people not to vote in the primary. Right. That kind of thing. And of course, we've tracked that very closely, and we're, don't just, folks aren't just sitting on their hands.

Our teams and other teams have been very interested in, for example, media provenance technologies, watermarking technologies. These were mentioned. Front and central in the executive order, some of these technologies were based, you know, on my team and came out of our groups at Microsoft and are now being shared in larger coalitions like the Coalition for Content, Providence and Authenticity, C2PA, C2PA.org if you want to read about that work.

So, in the context of the elections coming up around the world over the next 12 to 24 months, [00:55:00] there's lots of interesting work going on right now among, among multiple organizations as to how to grapple with these issues. The prospect that AI could be used to disrupt the other area, the other area I'm interested in, and concerned about is biosecurity.

I mean, with look, the AI powered protein design and bioscience more generally is going to be game changing for health care and for just even understanding the foundations of biological systems. It also can put new powers in the hands of malevolent actors to generate new toxins, for example, or gain a function research.

We need to basically stay on top of that and come up with the right kinds of approaches to, for example, screening DNA synthesis in new ways and updating those screens over time. So, there are mitigations and directions for research on mitigations that are possible on all fronts of concern that we need to make investments in.

At the same time, we have to [00:56:00] really lean into, maybe with some courage, exploring the upside to not be paralyzed. Clearly, this technology is going to change so many aspects of life. I often say I don't want to scare people, but to me it makes me excited that, you know, 500 years from now, the next 25 years will be recognizable as a named period of time because of AI advances.

It's up to us to guide that technology. We can't shut it down or stop it. It's part of our natural curiosity, our science, but so are the guardrails, and so is our democratic world's approach to grappling with the tradeoffs and doing, back to your comment, even qualitative decision analyses, like what we want to try, what makes the most sense, what will be the best thing for populations of people, for different stakeholders.

Alright. So, I think that's a good transition to the lightning ground. Lightning [00:57:00] round is where we're going to ask you a series of questions, maybe serious, maybe goofy, maybe funny. We'll let you decide whether or not you think they're serious or goofy. and the goal for the lightning round. Okay. So, the first question is Microsoft has a lot of big AI for science initiatives.

And as you just mentioned, drug development is one of the most important applications for AI. So, the question is, can you see a day where Microsoft ever makes a drug? We would make drugs, but not produce them. In other words, our folks would be pursuing new approaches to antibiotics and pharmacology as part of their AI for science work.

But the idea of making the drug and distributing it would come through partners. We're kind of a platform and, and, you know, research, organization. And, I see us not competing with pharma, but, supercharging, helping pharma to be supercharged. We should have a drug, you know. Yeah. Yeah. [00:58:00]

Anyway. Yeah. Good question though. Alright. Here's our next lightning round question. Eric, if you could roll the clock forward five years, what are Microsoft Research's major contributions to the field of biomedicine? I would say that our contributions will be new tools that enable drugs to be discovered and tested in simulation.

And the other direction will be, we will see advances that will ideally bridge wet labs and the in-silico world by giving in silico techniques the ability to call and design experiments that they need to push forward on advancing. So, is that like automated lab? Is that sort of like a full stack AI automated lab?

Full stack automated science with humans in the loop, of course, but this whole, this whole, there's been a really interesting long-term prospect going back [00:59:00] to our decision science conversation, expected value of information. Can AI systems really use their curiosity to drive experimentation, to collapse uncertainty by looking for the information they need through experimentation and guiding that process?

To me, that would be just a beautiful supercharging of science more generally. So, you've obviously been thinking about AI for a long time, and I think what we've learned today, too, is that you're nothing if not a true Bayesian. So, I think going back 20 years from now, or going back 20 years from today, what about your world model of AI has been updated the most in light of evidence that you've gotten over the last 20 years?

So, what beliefs that you held 20 years ago needed the most updating today? My answer to that question would be that the biggest updates I've received happened in the last year and a half over large language models and what I consider the magic that they're showing, which [01:00:00] suggests to me that what's going on with unexpected surprises and capabilities might somehow be related to the surprises we see in the magic that comes from large scale tangles of neurons.

Will doctors still be responsible for documentation in five years or will generative models like ChatGPT have taken over that task? I certainly hope the latter will be true. I want doctors, I still want physicians to be in the driver's seat. I want to celebrate the primacy of their agency and I believe that it is.

Human touch and human connection will be even more important as with the rising sea of automation. So, I think, you have survived the lightning round, Eric. Congratulations. So, I don't know if you've noticed, but I've been trying to dance around the "Sparks of AGI" paper, cause I wanted to save it for the end.

So, now I'd like to revisit that. So, you, you, you and, you know, Sebastian Bubek and other collaborators at Microsoft wrote [01:01:00] this paper with a provocative title called "Sparks of AGI." Where it was really like the first in depth look at GPT-4 and its capabilities. So, I guess AGI has a certain connotation and comes with certain implications for better or worse.

I've always like said that if you say generalist, artificial intelligence, no one cares. But if you say AGI, people start to really want to debate you. So, why did you feel that GPT-4 had hit or was worthy of this, you know, somewhat hallowed term in the history of AI, we used to talk about narrow AGI versus AGI.

Like why, like what about your interactions with GPT-4 made you thought that it was ready to be called AGI? So many people in AI research, serious researchers have looked at the term with raised eyebrow AGI, which came kind of late 90s, early 2000s, because we always thought we were doing AGI.

In [01:02:00] other words, we always thought we were pursuing principles of general intelligence. That's the idea with AI research. In fact, Herb Simon's 19, I think it was 1959, project with Alan Newell and others was called General Problem Solver. That was the original GPS in my readings. And so we said, okay, okay, we get it.

Some people, many of whom came from outside AI, research community said, you know what? I see successes with narrow AI? Like that, Eric, that readmission system or that system that can predict which patients will get C. difficile, you know, in 48 hours. That's narrow. Look at humans, what they can do, all these abilities.

We call that more general intelligence and we should be pursuing that. Now it's not that AI research wasn't pursuing that deep down. It was just that we were having some successes on the narrow front. So, we were kind of like, yeah, we get it. Okay. Now, AGI also became associated with kind of [01:03:00] doomsday scenarios of, you know, visions of *Terminator*.

And this is what the rise of AGI would mean. But if you go back to the early definitions of AGI by people that were using that term, they were calling out as general a set of abilities as people have. And they would, and we actually searched around as we were playing with GPT-4 as part of our studies, Sebastian and I in particular were trying to figure out how to frame the work.

We wrote that first section together about like, how do you frame this work? What do you bring AGI in as a concept? And we talk in the first section of that paper about the history of the use of the term, just like you're asking me now, why we thought, you know what, let's, let's use that term. because it really does mesh with the initial intentions of that phrase.

And we don't need to go down the path of changing it to generalist AI or something like that, because it has been defined quite nicely in different ways that are quite similar to what we're getting at now. Now, I felt it was important in the title to not say first [01:04:00] contact with artificial general intelligence, or it's here now.

Cause my perspective was we were seeing glimmers, like true sparks, like a snapping arc in places. And at times that were like, bald us over, like, wait a minute, like what's going on here? This is really impressive. And it does have a number of the capabilities that people have talked about for years. Let's just come out of the, out of the shadows of experimentation with the system, show the examples as to why we're seeing these sparks and look at these 72 sparks and their cross links.

Now, for me, the sparks included, I have examples in what's called the AI anthology online of some of the examples I played with early on, but even the fact that these systems could see when they weren't trained, on imagery, like I

was drawing faces in, you know, with if ASCII and talking to the system and, you know, yeah, I see a [01:05:00] face there.

And I would change it into a little lunar lander, just like by editing. And the system would say, well, I see a face, I think still. I said, no, that's a lunar lander. Oh, I'm so sorry. Why did you say face? Well, you know, faces are so common. And so, these systems were like almost like having dialogues, even about things that they weren't necessarily trained to understand, learned through language associations over time, I mean, over the training corpora.

So, I can go on and on about this, each spark that came to our attention, but. The biggest word for me was polymathic. I'll use that word again, like the ability of the system to weave with fluidity across different disciplines and combine things together and do synthesis really was the main sparking arc for me in, in coming up with that title.

So, I agree. I think if you don't view AGI in a quasi-religious kind of way, and you just look at it in the [01:06:00] strict technical definition of the word, it seems very hard for me to argue that it is not a general kind of intelligence. I guess so. Let me just say, so we were basically, in some ways you might say, we were claiming that term back.

We were calling it back from the quasi-religious and saying, let's get serious about this. This is computer science. And I guess, but so then there's still the qualifier of the title, which are sparks. So, I guess the question that I'd like to ask you is when do the sparks hit kindling and become a full roaring fire?

How far are we away from that? It's hard to know. Do you think it will happen in your lifetime? I think, it's safe to say from my point of view and what I've been seeing that the next paper like this won't use the term sparks. It'll be a different, construct to capture a more comprehensive set of capabilities we'll be seeing.

So, I expect surprises that are exciting in my lifetime. And [01:07:00] this is what career's been all about. We're getting to this point now where, as I said, we're coming full circle back to the deep interest in what neural nets were doing, was at the foundations of our own cognitive substrate, and I think we'll be learning a lot from these systems that has implications for that.

As well as learning more generally about principles or the physics of intelligence. Alright, Eric. Well, I think that's a great place to end it. I think we've come full circle. So, thanks so much for joining us on *AI Grand Rounds*.

Well, thank you. It’s been a pleasure. And thanks for all the great questions and conversation.

Thanks so much, Eric. That was great.