[Microsoft AI CEO Mustafa Suleyman: Our AI Doctor Outperforms Human Diagnosticians Alex Kantrowitz]

AI in Medicine: A New Frontier

ai's application in medicine is one of the most fascinating uses of the technology both because of what can go right and what can go wrong And so today I'm thrilled to bring in Mustafa Sullean the CEO of Microsoft AI to talk about the application of the technology in medicine and a new development that Microsoft has on that front Mustafa great to see you Thanks for coming back on the show Yeah it's a pleasure Yeah great to see you too Thanks for having me So I was reading your blog post that you're going to talk we're going to talk a little bit about this new development that you have as far as putting AI to use in medicine but one stat struck uh struck me right before we get to the actual work itself and that is that 50 million queries come in across uh C-Pilot uh and your other AI properties that are trying to solve medical use cases Is that a good thing I mean I think it's incredible because we're already with just search engines making access to information super cheap and concise And now with Copilot the answers are much more conversational You can tone them down so they suit your specific level of knowledge and expertise And as a result more and more people are asking C-Pilot and Bing um health related questions So yeah you're right We have 50 million queries a day uh that are health rellated and they can range from anything from you know um uh a cancer issue that someone's dealing with to a death in a family to a mental health issue to just having a skin rash And so the variety is is huge But obviously we've got a really important objective here is to try and improve the quality of our consumer health products And so this is 50 million queries every day Do the queries that come into chat bots look any different than search I mean asking you know uh search product what's going on with you is not uh anything new but I imagine with the chatbot you might be able to go one level deeper So do you see anything different in the queries Co-pilot's

The Diagnostic Bot: Revolutionizing Patient Care

answers tend to be more succinct and tend to be um more sort of responsive to the style and tone of the individual person asking the question and that tends to encourage people to ask a second follow-up question So it turns it into more of a dialogue or a consultation that you might end up having with your doctor So they are quite different to a normal search query And so let's talk about this Speaking of dialogues let's talk about this new uh development that Microsoft is announcing today which is that you've created and let me know if I get this right a diagnostician bot that effectively will be able to dialogue with a patient's case file and then make a diagnosis So it's actually two bots One uh and it's a system So it's not just a Microsoft's uh bots uh but it can be on any bot where there one bot uh basically acts as a gatekeeper to all patients medical information and then the other one is basically acting as the diagnostician or the physician uh that goes in and asks questions about that history And you found some pretty uh incredible results when it comes to the effectiveness of this system to be able to diagnose uh correctly Yeah it's a great summary That's exactly right We essentially wanted to simulate what it would be like um for an AI to act as a diagnostician to ask the patient a series of questions to draw out their case history uh go through a whole bunch of tests that they may have had pathology and radiology and then iteratively examine the information that it's getting in order to improve the accuracy and reliability of its prediction about what your uh diagnosis this actually is Um and we actually use the New England Journal of Medicine case histories Uh hundreds of these past cases One of these cases

comes out every single week and it's like an ultimate crossword for doctors Um they obviously don't see the answer until the following week And it's a big guessing game to go back through five to seven pages of very detailed history and then try and figure out what the uh diagnosis actually turns out to be Okay And so what happens is these two bots work together in conjunction to figure out what the diagnosis is Why use a system like this I mean I thought one of the benefits of generative AI is it can sort of take in a lot of information and then come to these answers sometimes in one shot So what is the benefit of having these this dialogue between two bots So the big breakthrough of the last 6 months or so in AI is these thinking or reasoning models that can obviously query other agents or find other information sources at inference time to improve the quality of its response rather than just giving the first best answer It instead goes and you know consults a range of different um sources and that improves the quality of the information that it it finally gets to So we see that this orchestrator which under the hood uses four different models from the major

Accuracy and Efficiency: AI vs. Human Diagnoses

providers can actually improve the accuracy of each of the individual models and collectively all of them together by a very significant degree about 10% or so So it's a it's a big step forward and I think that as the AI models get commoditized um you know really all the value will be added in that final layer of orchestration product integration and that's what we're seeing with this um diagnostic orchestrator So a 10% increase in uh accurately diagnosing on top of the standard LLM Yeah And in fact we actually benchmarked that against human performance So we had a whole bunch of expert physicians play this simulated uh diagnosis environment game and they on average get about one in five right so about 20% Whereas our orchestrator gets about 85% accuracy so it's four times more accurate which you know in in like my career I've never seen such a big gulf between human level performance and the AI systems performance Many years ago I worked on like lots of uh diagnoses for radiology and uh head and neck cancer and me mammography and the goal was just to take a single radiology exam and predict you know yes or no does it have cancer and that was the most we could do whereas now it is not just producing a binary uh class output but it's actually producing a very detailed uh diagnosis and getting and doing that sequentious sequentially um through this interactive dialogue mechanism and so that massively improves the accuracy Okay so it can do 80% uh accurate diagnoses which sounds incredible and I have to pressure test this a little bit because what if you have the same thing happen to medicine as is happening with uh beginner level code where basically there are people who are learning to code uh using these co-pilots but then when something breaks it becomes harder for them to figure out what's going on So if you're a doctor relying on something amazing 80% uh accuracy um but if you don't have if you sort of outsource some of your thinking to these bots is that a problem down the line Yeah So this isn't just giving a blackbox answer That's why the sequential diagnosis part is so important because you can watch the AI in real time ask questions of the case history get an answer shape a new question get an answer present a new question then present then then ask for a different type of testing get those results interpret it then give an answer so the dialogic nature means that a human doctor can follow along and actually learn in a very transparent way it's almost like having an interpretive ility mechanism inside the black box of the LLM because you can see its thinking process in real time and in fact you don't just see the sort of chain of thoughts which is the you know inner monologue we've actually created five different uh types of agent which all have a debate and we call this chain of debate they negotiate with one another they try to prioritize you know certain different aspects like cost or efficiency um and it's the

coordination of those different skills skill sets among the doctors which actually met doctor agents is actually what makes this um so effective but I I want to ask again because even if a doctor can watch this go uh take place it it effectively turns uh their role in di let's say this becomes something that doctors use it turns their role in diagnosis to from something that's active and really thinking through to more of like a passive okay I'm watching the bots uh go through it and I do wonder if uh there is some benefit in having the doctor actually have to do that themselves because it helps the brain work in ways that it doesn't when it's just watching bots have a conversation Yeah I mean I I think that's totally true I just still think this is going to be an amazing education tool for doctors to actually learn about the breadth of cases they never would have encountered For example we actually ran the the DXO orchestrator last week on the most recent case study in the New England Journal of Medicine and it correctly diagnose uh diagnosed a case that had only ever been seen 1500 times in all of medical literature It was such an obscure longtail uh disease So very few doctors are ever going to get the chance to see that And so the ability to accurately and preventably

The Role of Doctors in an AI-Driven Future

detect um these kinds of conditions in the wild in production I think will massively outweigh um you know the risk of of of of doctors not being able to sort of exercise in in the way that you describe I think the tools just change how you work and you know obviously everyone will sort of have to adapt to that over time but the utility is just so unquestionably beneficial that I think it it it makes it worthwhile Now is it able to do that because the cases are potentially in the training data And even if they are does it really matter I mean it is if it is able to diagnose these rare conditions Um should we really mind if it's seen it before in the training data Well part of the reason why we partnered with the New England Journal of Medicine is because each week they put out a brand new case which has never even been digitized So there's no question that it's not in the training data this case for example from last week there's absolutely no way it's in the training data because it's literally just got published So and and you know we we think that's the case going back for all of the previous studies too cases too So I don't think there's any chance of that This really is doing a kind of um a sort of abstraction of judgment It's not just reproducing training data It is actually doing some kind of inference or or thinking based on the knowledge that it does already have Now I have to ask you you mentioned that basically what happens in this orchestration is that you're taking um a traditional model and turning it into some form of specialized reasoning model Um there was a interesting point in the paper that said that uh the results uh improvement over like the reasoning models like an 03 weren't as big as they were uh elsewhere So is it possible that the reasoning models that are state-of-the-art today will effectively learn how to do stuff like this and you won't need this type of specialized uh sequencing in order to be able to make these diagnosis in a accurate way like we're seeing here I I think that the the real value here long term is in in actually how you orchestrate a variety of different models with different types of expertise So you know each one of these five agents has been prompted and designed to have a different type of expertise and then have them jointly negotiate um and reason collectively It may be the case you know that that maybe they all get subsumed into a single model in the future I I don't know Right now it doesn't look like that Right now orchestrators are the thing that is able to drive much bigger gains The other thing that we see for example is that it's able to um optimize for cost as well and reduce the cost by avoiding unnecessary tests uh versus the humans So that's a

function of um cost being factored into the orchestrator at inference time which you know wouldn't be something that you could reconcile inside one of the um a single model um you know in in pre-training or post- training Yeah we we shouldn't skip over the cost thing I mean obviously in medicine cost is a factor to ignore it to get better You know you could order every single test and probably do better diagnosing people but it's just not a reality today And it is interesting to watch the bot work through which tests to order and then come in actually at a lower cost than typical doctors uh and come out with a better diagnosis So how does it work Well more tests also make people feel anxious they waste time especially if you have a condition then that could be treated another way So it's not just cost but it's actually the patient experience that gets optimized for as well And so how does it decide which test to order and how to optimize cost Is that another feature within the model or well I mean you can think of an unnecessary test as a kind of error a human error So really what the model is trying to do is to get to the best diagnosis with the minimum number of tests And you know obviously you know the the model has much broader you know sort of range in terms of awareness of which test results tend to correlate with which particular uh diagnostic outcomes And so given that it's seen so many more cases than any given human then it's obviously showing that it can do a better job of judging in this instance given this case history that it already knows about a patient What is the sort of minimum number of necessary tests to get the next piece of information to be able to continue the um the diagnosis and get it more accurate Could I tell you something else that surprised me It seemed like the bot struggled with more common type of uh uh diagnoses was really good at the complex diagnosis I'll just read it straight from your paper Although it excels at tackling the and I I called it a bot It's really the orchestrator It excels at tackling the most complex diagnostic challenges further testing is needed to assess the performance on more common everyday presentations Do you think it's just like waiting to diagnose that rare case so it skips over the fact that it could just be a stomach ache All all we're saying there is that we haven't applied it to your everyday sort of GP or primary care physician experience where you're sort of you know you have a a skin rash or you you know you you've got a pain with your knee and so um this does tend to be the longer tale of complex cases but it goes without saying that there is less of that information in the training data and we know that if there is more training data the models

Expanding AI's Reach: Beyond Medicine

do better So if so clearly by virtue of the fact that there are more cancers more diabetes you know more weight loss questions more you know knee pain than any of these longtail conditions the model is almost certainly going to do better in those primary care type environments than it's doing on the longtail Um but because we didn't develop the research with those environments today we will of course as a next step that's why we added that qualifier right Okay that explains a lot And let's talk a little bit about the role of the doctor moving forward I mean we touched on it previously Um but there's an interesting this the release that you put out in conjunction with the paper does touch on what the role of the doctor will be It says doctor clinical roles are much broader than simply making a diagnosis Uh they need to navigate ambiguity and build trust with patients and their families in a way that AI isn't set up to do Can I just take the other side of this I mean maybe AI is set up to do this I think that if you're talking with a bot every day you might trust it more than a doctor you see once a year or uh even a new specialist So does is AI poised to be able to take on some of that work as well It's possible that it will do some of that work Certainly I I hope one day

that it will be good enough to do that kind of work But nothing is going to replace the humanto human connection that you get in the physical real world at a moment of heightened anxiety and fear when you're just facing you know one of the biggest challenges in your life and you have a massive diagnosis ahead of you or when you just need everyday regular treatment and care So that's I think going to continue to be the role of uh a doctor and hopefully they get to spend more time face to face with patients and you know sort of less time reading the you know the the the long tale of the history and so on So doctors become something of auditors of the out in the future they become auditors of the output uh of these AI bots They use some of their training to either confirm or look at a different uh different possibility and then mostly what they're doing is effectively they're becoming shepherds that are shephering patients through their you know for lack of a better phrase care journey **I think there's still going to be a tremendous amount of judgment um that is required by expert human doctors both as part of the diagnosis and then secondly making the judgment about what works for the patient and factoring in you know and helping a patient decide what journey do** I want to take given that I now know I've got this diagnosis what treatments do I want to take and when and what are the trade-offs there um so that that is going to require a tremendous amount of judgment and so it's not just about the humanto human connection and being on your feet It's also thinking in a deeply empathetic way alongside a patient that's re received a diagnosis to plan their treatment course Okay And what other professions could you see that's being applied to this type of system Well I mean the the basic method of these orchestrators is that they tune different AIs to play very specific roles and then have the AIs negotiate with one another debate and discuss That setting obviously applies to a lot of different environments um you know be it in business or even in government in the

Future Aspirations: Integrating AI in Healthcare

future Um and so I think if this finding holds and applies to other domains I think it will be be very very promising because it's also how we collectively as a human species work right We generally consult very widely when we make decisions Often there's even consensus um before actually coming to a final conclusion So it has a lot of parallels to the the human world And lastly this isn't being rolled out broadly in a hospital setting yet So everyone who's panicking at this point can relax But is that the ultimate goal Like where do you see this Is it an education tool Um or does this actually become integrated in medical centers and hospitals in the coming years I mean obviously at the moment this is just early research and we're figuring out how best to deploy it But I think the fact that we're able to get a 4x improvement on human performance across the board on diagnosis with significantly reduced cost in super fast time I mean to me that feels like steps towards a true medical super intelligence and we would want to try and make that available as widely as possible as quickly as possible including for our 50 million uh daily health queries and so that's going to be our ambition is like get it in front of consumers as as as fast as possible in the safest way possible Okay I'm definitely one of those uh 50 million who are asking these medical questions Uh and so the better it gets the happier I'm going to be Although I think with the caveat um that when things get really hairy I'm going to go see a human Sounds like a good idea All right Mustafa great to see you as always Thanks for coming on Thanks man Appreciate you Take it easy