**CRAIG:** Hi, I'm Craig Smith and this is Eye on AI. This week, I speak to Rohit Prasad, Senior Vice President and Head Scientist for Alexa at Amazon. We spoke about the development of conversational AI and virtual assistants and the merging of IoT sensor data into ambient intelligence, AI that is always present and immediately accessible.

**CRAIG:** Before we begin, I want to mention the sponsor of this podcast, CLEARML, an open-source MLOps solution for anyone building machine-learning models. ClearML allows users to easily Track, Orchestrate, and Automate ML Workflows at Scale. If you're a Data Engineer, ML engineer, DevOps, or a Data Scientist, go checkout ClearML today at clear.ml and signup for free.

**CRAIG:** Meanwhile, I hope you enjoy the conversation with Rohit as much as I did.

**CRAIG:** Rohit Prasad I'm delighted to be able to talk to you. I've seen you on Lex Friedman's podcast. I'm I'm a big fan of Lex's.

**CRAIG:** I'm delighted to talk to you. I appreciate you making the time.

**CRAIG:** I'm particularly interested in ambient intelligence and I see that you maybe a year or two ago had a blog post about that. So I wanted to ask, first of all through Amazon's eyes how you see ambient intelligence developing .

**CRAIG:** You are, you're obviously connecting Alexa with things like the Astro robot and the ring doorbell system, but what are some of the other sensors that you see entering this mesh? .

**ROHIT:** Yeah, you're absolutely right in what you just stitched together. , Ambient intelligence is essentially about AI or the underlying AI service embedded in everything around you and then it's available when you need it. Our vision is to make it really useful for our customers everywhere. For instance, in Alexa, you can just beckon it by saying Alexa and it answers or takes an action for you or, it fades into the background when it's not needed. Of course in that process, it can often even anticipate your needs which is what a great assistant advisor and companion would do.

**ROHIT:** But that's essentially what ambient intelligence is. And as it's evolving, it's very easy for customers to use because it's just available every moment. The best way to get ambient intelligence done is what I think of is as more generalizable AI not to be confused with the dystopian, all knowing all capable, Uber, artificial, general intelligence.

**ROHIT:** Instead, it's very pragmatic in the sense that just like we as humans generalize our learning across many different activities. AI should do the same. It should also evolve with the environment. If their environment changes, it should adapt to it. And lastly, we as humans explore a lot on our own and learn our own.

**ROHIT:** And of course, getting input from others when needed AI should behave the same way. And this is where you talked about Astro. That's a. Incarnation of an AI that's embedded in a form factor that's not restricted in a static location. It moves, it explores the world, just like our pets and we do.

**CRAIG:** Yeah. I've been talking to a company called OpenSensors. It's a small startup and they have a software platform that receives data from all kinds of sensors. And they're focusing on the commercial real estate market. So a company can have these sensors all throughout its office. It measures everything from traffic, the use of different spaces to air quality, to sound levels. And it's not ambient intelligence. It's feeding it into a dashboard to monitor these things. But when you look forward at Amazon, what other kinds of sensors do you imagine? Not only in the home. I've spoken to Fei-Fei Li. She's very interested in ambient intelligence

**CRAIG:** for

**CRAIG:** healthcare, elder care use cases. Are there things that are not yet out there on the market that you're looking at?

**ROHIT:** Yeah I think the best glimpse to the future again, is what we see happening with Alexa working together with many different devices we already have.

**ROHIT:** And you're right in saying it's not just in homes. Alexa is in your cars, it's in in your phone. But to tie it back, this sensors are a huge part of it. If you think. Any AI, especially in the underlying services that build Alexa, if you think about everything it does on reactive basis, where you ask something and it does something or takes a proactive action on its own.

**ROHIT:** All these sensors become hugely important. Alexa is orchestrating 30 machine learning systems that process speech, video, ultrasound touch and it's growing . One of our products is Alexa Together for remote caregiving.

**ROHIT:** And if the person you're giving care to has not been active or fell, then the sensors, again, matter. Can Alexa notify you based on that. So these are the kind of things where even in home, there's a lot more to do, but of course, in enterprise settings, also, you can think of many applications just like the way you were describing.

**ROHIT:** And this is where I think our Services like Alexa voice service, where any enterprise can integrate Alexa in their environment become hugely important because you can imagine these sensory data can be used to make the best decisions for your application. So I think this is the perfect combination of how AI starts maturing. It has human senses, but also superhuman ones like ultrasound, which we can't sense as humans.

**ROHIT:** But those are the kind of things that make it very much about the human and machine working together to get the best outcomes in any setup.

**CRAIG:** How is ultrasound integrated into the Alexa ecosystem now?

**ROHIT:** Oh yeah, some devices where you don't need to even speak to it, to shut off the alarm action or so forth.

**ROHIT:** So that can be done as well.

**CRAIG:** And how does that work? You're not talking about vision.

**ROHIT:** Not vision in this case. So you can turn off the alarm for instance, by just a gesture, and that picks up ultrasound. It's not a motion detection, not a motion detection. That's ulrasound.

**CRAIG:** Oh, is that right? Yeah. And on fall detection, that's integrated already in this Alexa for care.

**ROHIT:** It's one of the things we've been thinking about and these are one other applications that you can imagine that can happen. \

**CRAIG:** The conversations I've had are about, for example, in an assisted living environment you have multiple sensors That learn the patterns of behavior of the person living in that environment, and then can recognize changes in that behavior and alert somebody.

**CRAIG:** The issue that Fei-Fei talks a lot about is privacy. How to protect privacy in those environments. I didn't quite understand why that is such an issue because if you are using a system like that it would be simple enough to sign a waiver to say that I'm happy to have this AI ingest all of this data in and around me for the purposes of monitoring my behavior.

**CRAIG:** can you talk a little bit about that privacy issue?

**ROHIT:** I think privacy is paramount. We think of it as trust and the AI service has to be trusted by the end user . And in this setup, which you're talking about in terms of assisted living, the rights of the person is super important.

**ROHIT:** This is where we need to come up with the right consent models there for who's. You suggested a waiver , but I still think it's the right should should be with the end user in terms of. They should understand what data is being collected and how it's being used. And that's been our approach from the very beginning.

**ROHIT:** We want to be very transparent about what data is being collected by Alexa or other services and how it's being used for you. And you should have control over that data. So I think in those setups, how well does the end user understand it and how well does the caregiver understand What has been collected is very important to be very transparent.

**CRAIG:** Yeah. The learning that takes place, for example, with Astro it's in the development of the AI. There's not much learning, taking place once Astro is in the home.

**ROHIT:** So one of the things we are working on it's paramount that the AI can adapt to you. So when it's in your home imagine teaching your pet some activity. The same way Astro, you should be able to teach it through natural language . And one of the capabilities we are working on , building on teachable aI, is where you can teach Astro that this is the backdoor and this is what it's open versus close means. And and imagine if the back door is left open, and you left for work, Astro can send you a notification that, Hey Craig, , the back door is open. That will be hugely beneficial. And you can imagine that everybody's home is slightly different.

**ROHIT:** And when you said all the AI is learned Ahead of time or an offline setup while true. You always need to adapt to the environment you are in. And in this kind of a teachable setup, the AI is learning on its own by asking you questions or you are teaching it very much as you do another human or a pet, and that's what we are working on.

**ROHIT:** And I'm super excited about this invention because it'll bring the ultimate vision of the AI adapting to you come to fruition versus most technology we have to adapt to it. and this with AI, I think it's really profound because it adapts to you.

**CRAIG:** What exactly is the Alexa teacher model. Something, as you say, that's being worked on or does that exist?

**ROHIT:** Oh, Alexa teacher model exists. As you know currently there's a lot of effort in making AI generalize, not as well as humans do . But because of large transformer models, the models are becoming more and more competent in generalizing across tasks, right? . So for instance, transformer models are being used for generating code from natural language or images from natural language or even natural responses in conversational settings. Alexa teacher model takes large transformer models and adapts to the Alexa use cases by pre-training again on Alexa data, not just the large corpora. We believe that's a very important step where it's not just about the size of the model. There's lot of competition right now in the parameter war. Do I have the biggest model, but it often matters what data it's trained on, how competent the model is. And those should be an equally important factor. And the Alexa teacher model is powering is the three Ms; the multitask learning. As in, can you learn multiple tasks within a modality? Take speech, for instance, there's the task of speech recognition, transcribing speech into text, or detecting whether you said the wake word, Alexa, or who is it? Is it Craig or his wife? All of that was usually done with independent models learned on their own. Whereas with shared large acoustic encoders that are transformer based, you can now share a lot of these learnings. So that's multitask intelligence.

**ROHIT:** The next one is multilingual intelligence. If you follow the transformer literature and how the big models are being trained in that literature, you find that they're mostly used on one language like English. But here again, we use our customer obsessed approach in Amazon. Our problem is that Alexa is available in 17 language variants. How can I learn the commonalities of the structure of languages across all languages so that any given language now is learned with far less data and Alexa performs with the highest accuracy possible.

**ROHIT:** There also, we have used Alexa teacher model again by training it on 12 languages simultaneously and then adapting it to the language and tasks of interest by using novel distillation techniques that cut down the parameters from billions to millions. So that it can still run in real time in a tractable way and in a sustainable way.

**ROHIT:** And lastly is the multimodal intelligence which is where we talked about the teachable AI and Astro being able to understand the scene. For instance, if you said is the window next to the garage door open , then you need to understand that scene. And now you're tying visual concepts to language And that also is being enabled by Alexa teacher model that is trained in a multimodal fashion. So think about Alexa teacher model is a foundational model that makes every task that Alexa is competent on much easier to model and much more efficient and much more accurate.

**CRAIG:** Yeah. Maybe for my own understanding as well as listeners, we can explode alexa's and talk about the different parts. So I have, for example an echo at home. is communicating with an Alexa server somewhere, right? It doesn't have any AI in that box. So when I speak to it, it's sending it out to an engine that, that understands the speech, translates it to text, goes hunts on the internet, for example, finds a response, brings it back, translates it to speech and then delivers it to the speaker. Where is that central server? Is there one massive central server? Are there many instantiations around the world for each region, each language.

**ROHIT:** Yeah, let me make it more concrete.

**ROHIT:** When you speak to your device you address it as Alexa or any other wake word, the device only listens for the wake word. It's just detecting whether you're someone is addressing me or not. And then when the wake word is detected, which is also a deep learning based model by the way.

**ROHIT:** So it has the AI on the device too. And when it's very confident, it heard a wake word versus something else. Like if it's not confident that you said Alexa, it wouldn't start streaming to the cloud. Only when it's confident the light ring comes on and the audio streams to the cloud and the audio goes to multiples AWS servers, right?

**ROHIT:** At any given instance, of course, to one of the servers where I mentioned there are 30 machine learning systems that do something with that audio or or the text that is produced. Like we, if you whisper to Alexa, Alexa whispers back, which means it needs to detect how you spoke to Alexa. It also personalizes, if you said, play my music.

**ROHIT:** If it, if you have enrolled yourself as Craig, then it knows. It recognizes who the speaker is and then adapts to it. And does the, so all these actions of generating a response that is being done by orchestrating around 30 plus machine learning systems happens in the server and then a response is generated and a text to speech audio comes out on your device. This is in most of your early generation devices. However, our Echo Show 10, which is a device that has motion and camera in it that also rotates with you, such that the screen and the camera can move with you. That is a capability of on device speech recognition as well.

**ROHIT:** So the audio is recognized only on device not sent to the cloud . It's sent to the cloud, but you can also choose to not send the voice recordings if you wanted that extra comfort of privacy. So Echo Show 10 has that functionality too, but it still operates in the default mode that you speak the wake word and only then it start streaming to the cloud .

**ROHIT:** However, we also introduced a new modality, which we call conversation mode early this year and what it does on Echo Show 10, again, our most powerful device, which has a new network accelerator, which makes lot of AI happen on the device itself. A lot of times you'd want to converse with Alexa and you don't want to say the wake word every time like you and I talking, I'm not saying Craig in every reference.

**ROHIT:** So if you wanted that, you can say, Alexa, join the conversation. And then a huge blue border comes around the screen to let you know it's in a conversation mode. So it's now participating in your conversation. So you don't have to say the wake would every time. And when that happens, it's fusing your audio and visuals to determine whether you are addressing the device. Sometimes I may be speaking to you, but not looking at you. But of course, when I look to you, it's an easier cue that Alexa is listening to me, or I want to talk to Alexa. So with conversation mode you don't have to keep repeating Alexa over and over again, when you wanna explore something, for instance,

**CRAIG:** So these are updates on the servers in the cloud that your device is connecting to is right?

**ROHIT:** So in the conversation mode, even your speech is recognized locally on the device. All the visuals is processed only locally on the device. None of the video goes to the cloud.

**ROHIT:** The audio goes, unless you have turned on the setting, 'do not send voice recordings,' you have that additional setting for your piece of mind, if you wanted to choose that. But the text is sent to the cloud and then it determines what's the best answer to give to you using a lot of deep learning based what should be the best action that Alexa should take, given what it has understood so far.

**CRAIG:** Okay. That's interesting. I didn't understand that. So the voice is not necessarily saying, but the text, there's a voice to text engine on the local device on the local device. And when you talk about the Alexa teacher model that's a one step back, right?

**CRAIG:** That's your team teaching the model to do new things, is that right?

**ROHIT:** Yeah. Alexa teacher model, think of it as offline intelligence, which learns from all the data in the world and also the Alexa interactions so that it can generalize and train many different models for whatever task you needed. Of course, some people think that there could be one model for everything , but Alexa teacher model is the closest to that, but it can train a lot of specialized task specific capabilities by adapting this generic model for more custom use cases. Think of it as more of a general language model.

**ROHIT:** And here, I don't mean just language because it's also multimodal. So it's a, it's an Uber model that helps with more generalization for the Alexa use cases.

**CRAIG:** And that exists on a hardware device somewhere in,

**ROHIT:** it's not on device. It's on cloud. Yes.

**CRAIG:** On, oh, it is on the cloud. Okay. But it's it when a user is interacting with Alexa, it's not interacting with the teacher model, the teacher model is receiving data from all the Alexa devices in the world and is integrating it into

**ROHIT:** yeah. Think of the teacher model as a teacher. It learns from everything and it is also teaching a lot of specialized task specific models. Like there's a natural understanding model . The natural understanding model's task is to the text that has been generated from speech recognition and interpret what the customer is looking for. That's one task. Similarly speech recognition is the capability of converting audio to text. That's also can be informed by the teacher model and we talked about multimodal intelligence where visual concepts like 'the green outfit in the garden' is interpreted with the textual rendition to describe that scene. So all that is powered by this multilingual multitask multimodal model, which we refer as the Alexa teacher model. So that's more offline. It's not running in real time.

**CRAIG:** But that model then is updating the Alexa models or algorithms in the cloud that then you are interacting with.

**ROHIT:** So the Alexa teacher model is very expensive to train, so that's not updating as frequently as the more specialized task specific models that are trained off it. So those are updated more frequently with available data. The teacher model is much less frequently updated.

**ROHIT:** There's a lot of research going on scaling the Alexa teacher model in the most cost efficient and most sustainable way so that you're not using a lot of power. Think about us as humans, we operate in so little power whereas these machine learning systems take ton of power and we need to be responsible about how much power is being consumed as well, even if cost was not a factor.

**CRAIG:** And then back to ambient intelligence. So right now it's a camera, a speaker that you can interact with and maybe a temperature control or something, that are tied together. Are they being tied together locally or is all of that going up to the cloud and being processed there.

**ROHIT:** Depends on the use case . of course, if you wanted to check the temperature in your room and one of our echo devices has a temperature sensor and the others don't you still want the other device to be able to answer it for you.

**ROHIT:** Asking what's the temperature inside will require that temperature, sensing data to go to the cloud. So it's still very early in terms of the ambient intelligence about what data to process locally versus what to process in the cloud. But one concrete example I gave you is essentially the conversation mode that not only just because of privacy reasons, that you don't want everything streaming to the cloud, like your visuals but also because the kind of round trip that would be needed to send data to the cloud and then process things and bring it back will be too long. Yeah. So you want to make sure that the latency is low as well. So processing things on the edge or on the device has that benefit of latency . And that's something to keep in mind when thinking about what should be processed on the edge versus in the cloud.

**CRAIG:** Yeah. In terms of, research going forward, you had a paper about zero and few shot knowledge seeking turn detection in task oriented dialogue systems. , I won't try and repeat what the paper said, I did read the paper, but can you talk about that?

**CRAIG:** And that that learning, is there a point at which the consumers will be able to teach the device and will that learning then converge in some central model ? We all know about the Microsoft example, Tai was the name of the model, I think, where people were allowed to teach it and it very quickly became toxic.

**CRAIG:** But can you just talk about that?

**ROHIT:** Yeah. There is a lot of things here and I will try to unpack some of that. The paper you referenced is more about how you can use knowledge, let's say common sense knowledge , which is where we have extended our work now.

**ROHIT:** We call it think before you speak or the AI should think before it speaks. Take these large language models, which you just referred to, they're trained on lot of open corpora data that is available on the web, lot of different kind of conversations happen , and of course the model the way it's trained can start just mimicking those.

**ROHIT:** And depending on the query or request, it can just give an answer which can be at the very least insensitive, if not toxic. And so you have to put a lot of guardrails, but more importantly, it should also be, not nonsensical answers ,right? Think before it speaks, we are looking at using common sense knowledge , which is available in knowledge graphs, like ConceptNet where you have a lot of word pragmatics encoded.

**ROHIT:** And that helps ground the conversation. it becomes more sensible. To give you an example, University teams are trying to meet our grand challenge of 20 minutes of conversation for our Alexa prize competition.

**ROHIT:** If you and I met and we didn't know each other, can we continue a conversation for 20 minutes? That means both of us should be knowledgeable enough. Interesting enough.

**ROHIT:** So that's super hard. But that's an ideal setting to to explore human-like reasoning with common sense.

**ROHIT:** And if I said it's it's Valentine's day and what should I get my wife? And and if the social bot came back and said I think you should get some roses, red roses . So that's where it's now using a ton of knowledge about it's Valentine's day, it's a romantic event roses, and especially red roses signify romantic relationship. That's super hard for these language models to learn right away. They can start mimicking it if they've seen the data, but this is something that is very efficiently encoded in these in these knowledge graphs. And that's what you want to do. Hence think which is use the common sense knowledge or the word pragmatics in your modeling and then use it in real time before you speak so that's what our new work is in this area.

**CRAIG:** The evolution of conversational AI or Alexa started with pre-programmed intents, is the word I've heard used, where you ask a question and there is a canned answer that's given. It's evolved now to where the response is more natural, but the answer is still sourced.

**ROHIT:** It wasn't that it was ever canned answers. What Alexa did at the very beginning was to interpret the natural language query and map it into certain structured interpretation that you can act on.

**ROHIT:** Iconic example is play music by Sting. One interpretation is play music by an artist and the artist type here is Sting. And that is essentially a query then to look up. Does this person have any songs from Sting or does he subscribe to a service that has songs from Sting?

**ROHIT:** That was very much for a command and control, like more limited interactions. But when we introduced Alexa conversations in 2019, people don't want a single turn interaction. they want to engage and talk just like you and I do.

**ROHIT:** And in that setup, this whole model of structured interpretation to only certain actions is not sufficient. So we did the next part of it, which was the you just bring your API and examples of how in a natural language way will you interact with that API? And if you bring a few examples, Alexa will automatically figure out all the possible ways it can interact with your system or your service to complete the user request, which can take one or multiple turns. But let's say you were trying to buy movie tickets. You wouldn't hand code in your skill that first get the number of people the time, what movie? With Alexa conversations, you can do them in any order.

**ROHIT:** So that was a big extension. The third extension was in the context of people don't just stick to task oriented dialogues, where they're trying to accomplish a task, like buy movie tickets or concert tickets or so forth. People may often have a question in context or they would like to explore more things.

**ROHIT:** And more things started happening. Like I'll give you a concrete example of that with what I call conversational explorations, which is a new capability that we are working on. And where often you'll ask a question. And an example we just used recently is What's the oldest living tree.

**ROHIT:** And then Alexa gives you an answer that it's the bristle cone and it grows in this parts of the world. You can ask few more questions or it'll give you more recommendations and low and behold you're out learning about what hike is most popular in white mountains. And when you are learning about a certain topic, you don't get constrained by exactly what you started looking.

**ROHIT:** The conversation starts evolving very quickly into many different areas. And that is just not possible with a structured interpretation or just with APIs. You have in context question answering, you have more human- like language generation. Ultimately what this does is , if you asked me what's the best hike in this area. I would give you one answer. I'll answer it like an expert saying you should do this particular hike. Or if you said what's is the best Indian restaurant to eat at, I'll give you my recommendation . So this is where it becomes more daunting.

**ROHIT:** Alexa can't just spit a few links at you. Instead it should give you a recommendation based on the expertise in the topic, and then let you explore. So it changes the complete mental model. And where we started nine years back is progressed in a big way.

**ROHIT:** And now that you can have explorations with Alexa and you can ask today what's happening with inflation and it'll give. a bunch of stories around inflation and you can explore at your own pace.

**CRAIG:** It's amazing. When you say nine years, it certainly seems like much longer ago than nine years.

**CRAIG:** Where do you think we are? I'm sure you get asked this question all the time on the timeline or the curve are. Are we still climbing steeply and I'm talking about conversational AI generally. Are we at a plateau and we're looking for a breakthrough and if we are climbing steeply, how far in the future do you think we'll reach the plateau and where will it go from there?

**ROHIT:** I'm very optimistic about where we are actually. And how far things have come in the last decade I would say Alexa revolutionized how you could interact with a device from a distance. That was science fiction. we grew up watching in star Trek and then even in the hardest DARPA efforts, we used to think that's too hard.

**ROHIT:** But that happened, that was a huge advancement. Then the second thing that happened was the early set of self-serve tools for developers started happening with Alexa skills kit and Alexa voice service where suddenly the barrier to entry into AI where you didn't have to be a scientist to participate.

**ROHIT:** You had the tools for AI and machine learning. Though they were still crude. I wouldn't say we are done with those. We are still at the very beginning of self serve. But I think that was a huge advancement. And the third thing that happened with these general language models and teachable AI in combination, the serve service is not restricted to developers. It's now end user that can customize the AI for its need. And as this becomes more and more powerful, I truly believe ambient intelligence powered by generalizable intelligence will make everyday AI happen for everyone everywhere. So I'm very optimistic about this.

**CRAIG:** That's it for this episode. I want to thank Rohit for his time. I also want to thank Amazon for making it possible for me to interview Rohit and see much of what the company is doing. For those who are interested, you can find a transcript of this episode on our website, eye-on.ai.

**CRAIG:** I also want to thank Clear ML, our sponsor. Do me a favor and check out clear.ml.

**CRAIG:** And remember, the singularity may not be near, but AI is about to change your world. So, pay attention.