The Deloitte On Cloud Podcast

Gary Arora, Chief Architect for Cloud and AI Solutions at Deloitte

Title: **Intel's Brent Collins on how to deploy, govern, and scale AI agents to the enterprise**

Description: In this episode, Gary Arora and Intel's Brent Collins explore the rise of Agentic AI and what that means for enterprise leaders. They dive into agent design, necessary operational changes, and governance issues. For Brent, architecture and orchestration are critical to making Agentic AI secure, reliable, and scalable. Gary and Brent also look at the future of AI and how it has the potential to democratize expertise and reshape how work is done,--and how decisions are made.

Duration: **00:18:37**

Gary Arora:
Welcome back to the On Cloud podcast. I'm your host, Gary Arora, chief architect for cloud and AI solutions at Deloitte. Everyone is talking about AI agents right now: autonomous workflows, multiagent systems, orchestration layers. It is the next big thing in enterprise AI, but do we really understand what it takes to build and scale these systems responsibly? How do architecture, hardware, security, and governance involve in this new world?

Joining me is someone who lives at the intersection of AI, strategy, system architecture, and enterprise scale. He is the Vice President of enterprise AI Strategy at Intel. Please welcome Brent Collins to the show.

Brent Collins:
Hey, Gary. Thanks for having me. I'm really excited about the conversation today.

Gary Arora:
Likewise, Brent. Let's start with the big picture. The AI world today seems split between two visions, one is chasing a single, all powerful AGI, and the other is embracing a constellation of federated, specialized agents that can collaborate and solve problems. So, my first question to you is are these two visions actually in conflict and then how should tech leaders decide where to place their bets over the next 18 months?

Brent Collins:
It's a great question. I don't think they're in conflict at all. I think they're just serving two different purposes. So, if I look at the way humans interact today, and I'll take health care as one example, you might go see a general practice doctor to get to the bottom of where you need to put more time in and then you go see a specialist to go into a little bit more detail. I think the way that you see AI evolving is very similar.

AI for me is very similar to a human function, and that's why you see generalists and you see specialists, but I think we're going to see a lot of this evolve and you're seeing this with even the large language models today have mixture of experts behind them. So, you're going to have an interface and then you're going to go into somebody that might know physics better or somebody that knows research or something of that nature.

So, even those models are being broken out, but I do think as we look at the market, you don't need a general-purpose mega model for everything that you do. Sometimes you just need something that's very simple and very good at a specific function. So, for example, if you look at how you book a vacation, and I just got back from vacation, you need somebody who intakes all of the different things that you want to do and then you might have somebody that's really good at booking hotels or somebody that knows airfares and how air travel works and all that.

So, you might have just a very specialized niche, but smaller, agent to go do that. You don't need to spend the money for a brilliant mind to go book air tickets. You just need something that's very, very efficient. So, I think when you look at the market and where enterprises are going, they're going to be making decisions based on not just the technical aspect of what's best technically, but also on the financial side, which is what's going to be most efficient. I think that we're going to start to see decisions being made across both technical engineering as well as financial engineering moving forward, and that's how we're going to the best AI.

Gary Arora:
You mentioned health care. I spend a lot of my time as my day job is there. We've got so many workflows here that really require some efficient automation and not so much superintelligence, as you mentioned. Let's talk about architecture.
Now with Agentic AI, you do need an orchestrator, something that can translate human intent into the right sequence of micro-model calls. For our architects and developers who are listening, what core design principles or standards are absolutely non-negotiable for you?

Brent Collins:
Well, I think you hit an operative word, which is standards, and I think we're starting to see standards evolving and you see the same thing that you see everywhere in technology. I've been at it for about 25 years, some of you've been at it longer or not quite as long, but we see the same thing everywhere, which is you need solid standards and then you see those standards being broken out a little bit into standard plus standard-plus.

So, we see companies with off-the-shelf products and then you're going to have more in the general market, open-source tools that are being used to orchestrate. So, I think you're going to start to see this bifurcation a little bit into off-the-shelf tools that have their own way of putting together agents within their construct, and then you might see something a little more open source or free out there, eventually though those things will converge. They need to talk to each other.

For example, you see Model Context Protocol, which is how agents interact with one another in the wild, and we're going to start to see those become much more sophisticated with how they do things. Now, you mentioned what's non-negotiable. I think it's a great assertion and what I mean by that is you see things like privacy and security coming up a lot. So, you're going to need a really robust governance model to ensure that data chain of custody is really important.

So, you're going to see agents exchanging data. Well, when they exchange data, you're going to have to make sure they're exchanging the right data. If they're exchanging the wrong data and it gets out in the wild, that's a problem. So, data labeling behind the scenes is going to be really important as well as tracking how that data moves between agents and then ultimately gets to where it's going. On top of that, you want to make sure you're confident that the data is the right data. It's the most up-to-date data.

So, things like real-time streaming, real-time checkpoints, things of that nature are going to be really, really important as you see more and more agents interacting. Again, no different than humans. Humans exchange data all the time. You want to make sure the right human has the right data. It's the same thing in AI.

Gary Arora:
So, privacy, security, and governance to ensure that right data is being shared by the right models as your non-negotiables, cannot argue with that, but let's double click on it. Agentic systems do need fast bidirectional data sharing, but if you look at any enterprise, the data lives behind compliance walls, across geographies, departments, and legal zones. In what you have come across, how can teams federate data safely without triggering these compliance nightmares?

Brent Collins:
Well, we're working internally on this exact problem, and it is a problem because there is no clear answer. If somebody tells you we've got it figured out, one, I would ask them to check their math, and two, I would ask them to prove it. It really becomes difficult as you get more real time with the data in particular. So, if I move back, there's some fundamental core principles that need to come into place, and we've talked about it, which is your data has be properly labeled; those labels have to carry through.

So, over time, that data may change, it may drift, and you need to make sure that you've got compliance checkpoints to make sure that you've got the right people have access, the right people are restricted, the data is up to date. So, one of the biggest problems that we see is that data, it atrophies over time. So, you got to make sure that you're replacing that data with the most up-to-date data in a chain. With some workflows, it might not matter. If I'm looking at something that's my tax returns, it's not going to change over time, but if you look at stock prices, they might change in a millisecond.

So, we really need to take a look at what the agents are working on, who they're interfacing with, and the quality of data that's needed for that and quality could mean the data is correct or quality could mean it's up to date. So, those are two things that are a little bit different and how we look at data as it flows through agents and various applications.

Gary Arora:
Data atrophy and data quality are really top concerns here, especially when agents start calling each other's output and sharing memory, you have new threat models showing up, things like prompt injection, decision loops, data leakage across agents. From your vantage point, where should we draw the line between helpful autonomy versus runaway risk?

Brent Collins:
Well, I tend to be a person who takes a little bit more risk in the name of innovation. So, I think early on, I would say it's OK to err a little bit on the side of innovation velocity, but I think over time, it's going to be like everything else. You innovate as quickly as you can, and you might take a few gambles as you go do that. Then, as you get a little bit longer and things get more mature, you've got to lock things down a lot more. I think there's never a perfect line there.

You do the best you can with the information you have and then over time, I think security gets their hooks into more of these things. Also, as legal gets involved and you start to see more exposure, companies are going to have to take that into account, but I would say take advantage of the liberty, right now, that you've got to innovate, innovate quickly, fail fast if you need to, and move on.

Gary Arora:
100%. Based on what I'm seeing in the marketplace, this space is still growing and evolving. There are new tools and standards coming into the picture, but let's talk hardware. I've got a few small models running on my home servers. I'm a big fan of lightweight and local AI. With Intel's Xeon 6 that is now proving that CPU-only inference is viable for specialized models, how should tech leaders decide when to stay CPU-only versus when to mix CPU and accelerators and perhaps when to double down on dedicated AI silicon?

Brent Collins:
Great question. I think I'm going to go somewhere you probably wouldn't expect from the leader in CPUs, which is I would rather not be talking about silicon at all. I would rather be talking about the workload first and the needs of the business and then working your way into whatever makes the most sense. So, you mentioned you have it on your local PC and you could have it running on a Core Ultra, which is an Intel product for clients, and it would run a Llama model just fine on your local machine.

But what we're seeing with a lot of companies is there is a different operating model for GPUs, for example, versus a CPU and they're much more comfortable and/or they already have the CPUs on hand. So, there's nothing that you don't need to have a religious debate about GPU versus CPU versus NPU. It really is about the workload that you're trying to run, can you get the performance out of what you already have or what you're going to buy?

Then, again, I think we get caught in this technical dilemma and a lot of times it's a financial dilemma, which is, OK what's the most efficient way to go do this? Sometimes, "efficient" doesn't mean "lowest cost." It just means, "OK, I've already got this on hand, and I can use it or if I go purchase this." We see that 25% of pilots get out of pilot and go into production.

So, do you want to buy dedicated infrastructure that can only do one thing, or do you want to buy something that can run a lot of different things in case you decide to go in a different direction? So, again, for me, it's always going to start with the application, the workload, the user, and then work your way back into, OK, does this perform well for the user, is it financially viable, is it secure?

All those things come into account. Then, the last thing is operationally introducing something into your data center or into your cloud environment that you may not be comfortable with or familiar with, that could be a problem. So, "Can I run these well, do I have people with the technical skills that are required to go do this, or am I going to have to make a new investment?"

So, no different than most applications, but again as a silicon company, you would expect me to start there. For me, it's really let's start with the workload of the application, what you needed to do, and then work your way back into all the other things that dictate that decision.

Gary Arora:
Now, I like how you summarize that. A lot of times it is a financial dilemma or a technical constraint. So, starting with the user and then the workloads and then working backwards makes perfect sense. You also mentioned OP models. So, I want to talk more about that. Beyond the technology, what has to change in an organization and an OP model in terms of MLOps and perhaps now Agent Ops for managing and versioning and monitoring hundreds of different independent agents without creating chaos?

Brent Collins:
Well, I would say first of all, everything needs to change if you're really looking at it holistically over the next 10 years because we're not looking at traditional applications or even traditional. If I look at the way hardware runs, the way you operate hardware is very different than the way that you manage people and the way you manage agents is going to be the same way you manage people, are similar I should say.

So, the operational models are going to have to change dramatically, but if I go back to infrastructure a little bit and you may have been surprised to hear me mention that, but you have people that are very familiar with existing legacy hardware, for example, or even in the cloud how to operate these things and one of the biggest problems I had dealing with enterprise clients. And it sounds weird coming from a technologist was simply power and energy in a data center. We had data centers that introduced AI, and all of a sudden, I was at 40% space capacity and 100% power capacity or cooling capacity.

So, the model with AI changes everything all the way down to the data center and we like to say you don't have a technology problem. You have a backhoe problem, which is I need a backhoe to basically dig up and deliver more power into my data center, more effectively cool that data center, but I like where you're going with the higher-level operational side, which is even the way that you do applications today is very different than the way you're going to have to manage agents because agents have agency, which means they can make decisions on their own.

There's going to be a certain level of autonomy, and if you go, I think what's really instructive might be automation, which is where do you have a human checkpoint in that workflow. Similar to what you might do with automation, you're going to have to do the same thing with agents. The other thing that we're going to have to really be cognizant of is, and we used to call it paving the cow path in automation, which is you don't want to just go where you've already been. So, it's a really good time to do value stream mapping to really understand the way that workflows should work versus the way they do work.

And I think that actually Agentic and some of the Generative will do that anyway because the way that it thinks is a little bit different than the way that we've traditionally thought, but it really provides an opportunity to rethink how these things interact with one another, how they support your business, and then the way that you operationalize that is going to be different as well.

Gary Arora:
Great response, Brent. The way you manage people is the way you manage agents. Now, that's the mindset shift we talk about because you're managing really agency here. Speaking of the shift, what's one common belief do you think most CIOs, CTOs, or even AI strategists need to completely rethink as they plan their AI road map over the next, say, three years?

Brent Collins:
Well, I don't think it's necessarily the way you rethink or think as much as you just have to be open to new ways of doing things and new ideas, and what I found is if you have an assumption on AI, it's going to be probably ungrounded in weeks, if not days. So, be open to new things, be open to new ways of doing things.

I know that it might sound like a cop out on the answer, but what I found is one of the biggest inhibitors is, I'm going to use AI to automate the way that humans do things or maybe do things in a similar way just faster. I think that it does things completely differently. It can do things that you may not be thinking of right now, and I think being open to some of those new ideas and some of the new things and the new ways that it will do that is going to really be beneficial to most organizations. So, if I'm a CTO or a CIO, I would say be open, and be creative, because there's all sorts of things that you can do with this, that you may not be aware of today, that you're going to be able to take advantage of moving forward.

Gary Arora:
Be open, be creative, golden words. Let's wrap up on a hopeful, optimistic note. So, fast forward five years, Brent, what kind of problems now seem solvable for the first time and how might the, say, day-to-day, experience of work actually become simpler and more human friendly?

Brent Collins:
It's a good question. I mean five years in AI is like 25 years in any other market. So, it's really hard to say. I would say as we move more towards general intelligence, as some people think we're already at AGI, but I would say as we look at superintelligence, one of the things we used to say before I came over here and I'll steal it from a good friend of mine, Tim Brooks, is what if anyone knew what everyone knew.

It's a really weird way to say we're going to have expertise, deep expertise, at our hands and everybody will have it. So, I don't need to go contact the world's best physicist. I already have that at my disposal. I don't need to contact the person who's the best at you name it. If I want to go learn tennis or my son is in Calc 3 right now and he uses it as a tutor for Calc 3, there's just you have at your disposal all of this expertise, and I think the way that that comes together through humans is going to be really powerful.

So, I think when you democratize intelligence is what I would say, you're going to be able to do so much more if you just open up your mind and really take advantage of that. The ability to ask really good questions of these experts is going to dictate the results you get. I think when we look at moving forward again, the democratization of intelligence is just, man, it's so powerful.

It means you can do so many different things and it's not just the most wealthy or affluent they're going to be able to have access to this, it's everybody. My son, he's a dead broke college student, but he's got access to the very best tools and the very best tutors for what he's working on, and there's no substitute for that. It's just going to change everything.

Gary Arora:
100%. As you said, democratization of intelligence, I think it's the single biggest inflection point that we are in right now. What if everyone knew what everyone knew? I want that tongue twister on our t-shirt.

Alright, well, that's it for this episode. Thank you for tuning in. A big thank you to our guest, Brent. If you liked this episode, please be sure to leave us a review so we can continue to keep it real and bring you more insights from the trenches. Thanks for listening to the On Cloud podcast. Until next time. I'm Gary Arora.

Operator: