On Cloud Podcast
July 2024

The Deloitte On Cloud Podcast

Mike Kavis, Chief Cloud Architect, Deloitte Consulting LLP

Title: **Avoiding AI anarchy: DevOps leader John Willis on why and how organizations should embrace Generative AI**

Description: In part three of our series with DevOps leader John Willis, Mike Kavis and John extend their discussion on AI. John strongly believes that, in light of Generative AI’s growth, it’s unrealistic for organizations to continue to block its use because it will simply lead to shadow-AI anarchy. Instead, John says it’s essential to embrace AI, but in the right way—which includes setting realistic objectives for its use, developing AI governance policies, and training workers on how to follow them.

Duration: **00:17:39**

Mike Kavis:

Hey everyone and welcome back to the On Cloud podcast. Today we will be finishing up our three-part series with John Willis, thanks so much for joining us during this series, and I hope you are enjoying it! We are going to jump right in where we left off in part two and continue our discussion on SRE in the cloud. In part two, we ended the episode by highlighting SRE and how it can actually improve workflow and efficiency in technology management. We discussed the challenge of balancing security and usability, emphasizing that well-structured constraints can lead to better management and operations at scale. Let’s continue our conversation with John.

So, here's the challenge. This is what I saw in cloud and I'm sure I'm going to see in AI. The people making the decisions on the controls do it in the context of what's good for their silo and not what's good for the people who consume those. So, too often you have amazing clouds. where you can do so much. They put controls on it with only consideration of security and governance. And then people can't get work done. And it winds up having the opposite effect. It's like, "I can't do my job because you put so much controls on it." So, there has to be a balance. The people deciding on these controls must be customer-focused and the customers, the builders.

John Willis:
This is why I'm such a big fan of SRE. One of the big things when you go back to Deming, Deming hated MBOs. He hated metrics that were basically results-based because you didn't learn anything from them. And the problem with almost all organizations like the ones you described is we create those type of environments where you're blocking and locking.

And it may be like, "Oh, no, OKRs, John, no, they're not—they don't –" Baloney. You tell people that their motivation is based on results, you're locking stuff down. And what Deming would say, it was about the method. You'd always say, "By what method? What was the mean?" Some people call it measurement by means. "What is the method that you used?" And this is why I love SRE, because if you look at a KPI, MBO, and even OKR management structure for controlling stuff, you could say, "Well, it's your bonus," but you're putting results-oriented structures on human behavior.

When you do it by the method or the means, you're basically allowing people to experiment and learn and try to discover, and if I say the result is it has to be X by this by this, then I'm either going to hit that number, I'm going to lie, or I'm going to miss that number. But if I say that the objective is to constantly iterate on the improvement cycle. And then, going back to SRE, the whole idea of an SLI and SLO is more consistent with Deming's ideas.

I would say that—OKRs are Tayloristic. If you know, this is Frederick Williams-Taylor and that structure of command and control structure. I know they weren't designed that way. I've read *Measure of Matters*; I've read about Intel's Andy Grove but the point is they're implemented in a way that's Tayloristic.

And service level indicators and service level objectives are more about the method. So, again, when you take the example of the thing that you've seen and we've both seen in cloud, if we take the MBO result or OKR approach, we normally get controls and people then have to do workarounds.

But if we take sort of the SLI, SLO approach, if it's operated in the right way, you're having a discussion with the different components that will make up that service. So, you'll say, "I need to bring the business owners in, I need to bring the operations, I need to bring the developer," depending on how scaled or horizontal it needs to get. And what do we do? We design together the indicators that are going to make up the objectives.

And it's a far better learning, because first off, we start off with that it isn't hard or fast. "Your OKR is to have this by the end of this quarter, the end of this year." Whereas a properly defined methodology for implementing SRE with SLIs and SLOs is we're constantly learning. For those who listened to the podcast last week with Deming, you're using a theory of knowledge or a scientific method to say, "Hey, I think that the SLI for this service should be 10 millisecond latency. We don't know, but let's start there. And then let's keep documenting the methods that we're using to calibrate and look at it." And the SLI and SLO isn't a static definition that will get reviewed every quarter. It is something that's sort of ongoing and living and the monitoring is connected to it. So, again, go back to Generative AI. Let's apply sort of an SLI, SLO objectives to the constraints that we believe our hypothesis need to be done.

You can either block, ignore, and embrace. Embrace doesn't mean just turn it off and turn it on and let everybody just go hog wild. It means you are at least admitting to yourself that you don't know the answers. We're going to start at some starting point and we're going to learn and adapt and we're going to create a feedback loop. And so, yeah, right now—paying attention, some of the things that you relied on in the past to be canonical sources for your security, like the whole OVAL, STIG, CVEs, they're so far behind right now. And everybody else, they're just struggling to keep up.

Mike Kavis:
And I'm going to go back to someone who's been building for almost 40 years and blocked for almost 40 years. Controls are great as long as it's in the context of the company delivering what they promised to deliver. And too often, things are locked down with only the consideration of the people who own security governance and those types of things.

John Willis:
Yeah, yeah. Absolutely. I mean, that's DevOps. I'm considered one of the founders of the DevOps movement. I'm also behind the scenes one of the founders of the DevSecOps movement. Mark Miller and I created the first DevSecOpsDays . And that was really a big part of how do we do what we did with Dev and Ops with security? And then, I wrote a paper that turned into a book called *Investments Unlimited*. It's like, "How do you deal with automated governance?"

And so, there's been a lot of movement in general IT of how to sort of eliminate that sort of these ivory tower people sitting who knows where are demanding that everybody do it this way. And I've done qualitative analysis for large banks where these young kids will say," John, just find this person that owns this control so I can just have a conversation with them about how insane this control is." I mean, there's large organizations, that are like "Who owns that risk?" It can't even be navigated. It's the companies are so complex and big that no one person could actually tell you, "Oh no, that risk is owned by this guy. Here's his location. Here's the number. Just call him and why don't you have a conversation about how insane that risk that was developed 25 years ago."

One of the worst risks I ever saw, was they had to do a contingency plan for data. You really don't need a contingency plan for data storage. And the whole planet would have to be destroyed before you'd lose your data.

Mike Kavis:
So, back to shadow AI in case people didn't get this out of the earlier conversation, is—with shadow IT and cloud you had to be highly technical. You still had to build code. You still had to be able to manage infrastructure. With AI, you just need to be able to prompt an engine. So, not only does it expand the universe of the shadows, but the technical knowhow of the shadow.

So, if I'm not technical, I ask ChatGPT something, I just take for granted it's right. And I know from trying to use it to build code, you have to iterate. You have to be more precise than just, "Build me X." So, that's what caught my attention is that I wasn't thinking about that until you mentioned it. And I'm like, "Oh, my, now anyone in a company can build something with zero validation and it would be all over the place. So, that's what caught my eye. So, any closing two-minute thoughts on that?

John Willis:
There's a great presentation I just saw, the keynote speaker gave a compelling argument about you don't really program anymore. The Von Neumann sort of structure of computing is completely different with models, the way you program through natural language. It's not even programming. Now, under the covers it's writing code, but this is the scary—part if only five percent of an organization was coding and that was how all your business services were being created, at least you could put a microscope on it to a certain extent.

But imagine now that's three orders of magnitude more. The shadow is three orders of magnitude bigger. And a large bulk of that is still code. It's just that the human is having a natural language conversation with the model and the model is programming under the covers. And you're right, that is a complexity of technical debt that is sort of almost incomprehensible.

And so, you sort of start thinking about that and you think about the new way to program is through sort of NLP and models. And at some point, whether it's writing code or not, especially when you talk about mixture of agent and agents, the way people are now, if you're sort of following what's going on, what you can do the sort of agent structure, this is a whole other level. I've been working with autonomous dev tools. I just got access to Microsoft's Copilot workspace.

This is just Copilot. This is a tool that you open up an issue in a GitHub repo and it goes and analyzes it, it gives you a work plan for fixing it and asks you, "Do you want to fix it?" It allows you to sort of just code, so you can go in and say, "Well, no, no, no. I don't want that. Do this, do this." But you hit the button; it fixes the problem. We are moving to—and this is all—if you think about when ChatGPT came out or Copilot came out, or a couple of years, we've already moved up two levels of abstraction. We've gone from code helpers to autonomous coding to autonomous software engineering in less than three years.

Mike Kavis:
Yeah, it'll be interesting because currently, it's only as good as the prompter.

John Willis:
Well, again, this is where the sort of mixture of experts or these sort of autonomous agents working together, because all humans have bias? I know we're been running out of time, but the one thing I hate when people say, "Oh, you can't use AI because it has bias and it hallucinates," or "It's not as good coder as a human coder."

So, what's happening now is the models are now being the judges and the models can make decisions in very specific, narrow AI, they call this, very narrow AI focus, can make decisions as good or in some cases better than a human. And if you start combining them as agents to work together, you're finding that your output is actually getting better.

There are people that are combining hundreds of autonomous agents working together to solve problems where every one of it is at least 50 percent as good in a narrow focus as a human.

Mike Kavis:
Yeah, I agree with that. I still think you have to ask the right questions to get the right answers, though.

John Willis:
But we're seeing more and more is it's going to be the Generative agent, the LLM, that's going to ask the question. So, it's going to by default ask as good a question as a user and learn how to ask better questions at a scale that humans can't learn it. And to me, again, go back to the premise of this conversation, is all that is fantastic for business and human knowledge, in my opinion. The question I have is: Is it going to destroy your organization because you didn't pay attention to the potential technical debt tsunami that's going to happen because of shadow AI?

Mike Kavis:
Yeah, there's two parts to that. If you block it, you're going to destroy your organization, so you will fall behind. If you ignore it, then what you just said, it's going to destroy you. So, I guess the art of it is how do you embrace in a continuous learning process?

John Willis:
That's right. How do you learn, create feedback loops? How do you sort of get everybody involved? One of the large healthcare insurance companies who has been a vanguard in the insurance industry for cloud, what they've done is they've created a global training for—what they want every employee to understand, what are the Social Security implications and all that. It's like when you go to work for any big company.You go through that sort of mandatory training where you literally take the class, you can't skip it, you can't go any faster than it'll allow you to go, you've got to answer the questions, and when you're done, you sign off on it, and it goes into your record as an employee. And that's how a corporation protects itself. You're in the banking industry; you know there are rules of things you just cannot do with system of record data or personal account of user. And you can't raise your hand and say, "Oh, well I didn't know that. I thought you could copy that file to this thing." No, no, no, you literally took a mandatory training class and you signed off on it and it's in your record of being.

And so, this idea of embracing is instead of ignoring it, let's give everybody this training, admit that we don't know all the answers right now, but we have some hard and fast theories about what we should and what we shouldn't do, and you've got to sign off on this to say that you learned this, so you can't (a), ignore it, but more importantly, you can't do the, "Do now and ask forgiveness later." I mean, you can but you could get fired, because, "Well, no, I didn't know we couldn't do that…" “No, no, no, no, in section 4.3 of that training that you signed off on, that you agreed as part of your employment to agree with, it said specifically you can't do this.” "Oh, I forgot about that." “No, no, sorry.”

So, again, I think there are a lot of cool techniques that we can learn. In this one insurance company, they're not putting a blind eye, they're not just going out and just hiring a chief AI officer; they're treating it as a CIO or infrastructure and operations. How do we protect our corporation, our brand, our risks, and all that? And we're trying to come up with things that we know work pretty well in not only cloud, but how we did big data--data classification is going to be huge. People are going to create LLMs that are going to wind up going out to the customers.

And if you don't know the data provenance—let's say data that you've licensed for a very specific use case, maybe GIS data or some very specific data that—in your contract for that data—was contractually only to be used here, somebody, when they were building an LLM were trying to get out there really fast, were grabbing all data sources, they were throwing it into the LLM, the vector database or whatever they're using. And then all of a sudden, two years later, that vendor figures out part of that data came from a very specific license use case and it wasn't licensed to be used for general customer consumption. Oops, we got ourselves a big lawsuit here, not counting the brand reputation.

And when we built data lakes we learned a lot. We don't just stuff everybody in there; we do sort of staging of data.

Mike Kavis:
On that note, we'll close it out, but I appreciate the time. It's always a pleasure. So, that's it for today. I hope you enjoyed our podcast. Make sure to like this, leave a review, and subscribe. You can also check out previous episodes where you listen to your favorite podcasts. I'm always on X, madgreek65. Or you could hit me up on e-mail, mkavis@Deloitte.com. Feel free to ping me and let us know what topics you want to listen to. And until next time, to the clouds.

Operator:
This podcast is produced by Deloitte. The views and opinions expressed by podcast speakers and guests are solely their own and do not reflect the opinions of Deloitte. This podcast provides general information only and is not intended to constitute advice or services of any kind. For additional information about Deloitte, go to Deloitte.com/about.

This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor.

Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

Visit the On Cloud library
www.deloitte.com/us/cloud-podcast

About Deloitte

As used in this podcast, "Deloitte" means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.
Please see www.deloitte.com/about to learn more about our global network of member firms.