

3 critical layers for containing AI risk

Transcript

SPEAKERS:

Vikrant Rai, Associate Partner, Internal Audit & SOX

Mark Owens, Partner, Transformation Advisory Services

Katie MacQuivey, Partner, AI, Transformation & Customer Strategy

VIKRANT RAI: Hallucinations are part of what AI delivers. And I think if you look at how you can contain that risk, it's typically applying three critical layers of understanding — people, process and technology. You've got to train your people to do it right, to leverage the prompts and the frameworks that are available. From a process standpoint, we want to be sure that as AI solutions are being adopted mainstream, that there is a thought-out process for how those are being taken out. And then thirdly, from a technology perspective, it's important to understand what kind of data sets the AI models are being trained on. Are they being trained in a way that is more effective and giving the solutions and outcomes that you want to see? And then lastly, looking out for any deviations of behaviors. We can expect to see more evolution from that standpoint where you have agents monitoring behaviors of other agents so you can separate the outcomes and avoid one model from getting contaminated by the outputs of the other, but that process is still evolving and we're likely going to see it materialize in the time to come.

MARK OWENS: The risk profiles with hallucinations are being mitigated both programmatically within the models themselves, and then particularly in terms of some of the technology, the capabilities, whether it be LangGraph or CrewAI, the ability from an orchestration standpoint to be really less probabilistic and more deterministic. So it does come back to a little bit, what's the use case for AI? If this is a machine-learned model, you really don't have to worry about hallucinations at the end of the day. But if it does venture into, hey, how does the ChatGPT, the Claude ultimately start to make these decisions, it really involves some context engineering and strong prompt structures to ultimately enable a better outcome.

KATIE MACQUIVEY: You always need a human in the loop. That will be true no matter how mature and how capable AI becomes. The element that you can't create is that creativity, is the connecting of the dots, and it changes what you need in that manager that's actually looking at the output. Now all of a sudden, you need systems thinking,

critical thinking, management skills. A lot of those later-in-career skills you need upfront because you're going to have your most junior people trying to do that QA on the output from an agent. Put the guardrails in place and the stage gates, and make sure you have a consistent process that everyone can adopt to ensure that your output is correct.