Meet the man helping Alexa rule the world

John Davidson

Amazon has let its artificial intelligence assistant off the leash, letting it learn by itself without human supervision, but the man in charge says it won't hatch global domination plans ... yet.

When the AI formerly known as Alexa attains hyper intelligence and takes over the world, historians picking through the ruins of humanity may one day point to a day in September, 2018, as the beginning of the end.

It was September 20 when Amazon quietly announced it had allowed its hugely popular voice-controlled digital assistant, Alexa, to begin teaching herself. That "game changer", as the scientist in charge of Alexa, Amazon vice-president Rohit Prasad puts it, started off simple enough.

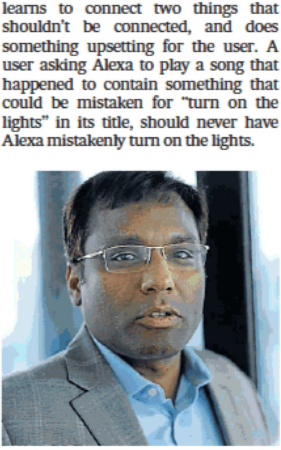
Rather than have customers continually interrupting Alexa (or "barging in" as it's called in AI parlance) to make the same corrections over and over when they were trying to get Alexa to play a certain piece of music or a certain radio station, Alexa has now started to learn from its mistakes.

If enough Alexa users ask Alexa to "play Sirius XM Chill", and then moments later barge in to ask Alexa to "play Sirius Channel 53", Alexa will learn that Sirius XM Chill is another name for Channel 53 on the US-based satellite radio station Sirius, and from then on it will get it right.

Such "unsupervised" learning based on corrections made by the public, says Prasad, obviates the need for humans working at Amazon to make such a correction in Alexa's neural network, but it's not without its risks, even in the early days when Amazon's experiment in unsupervised machine learning is limited only to music, and limited only to the US.

(There's no telling when the feature will be rolled out here in Australia, he says, nor when it will be expanded beyond music.)

Right now, the chief risk is that Alexa learns to connect two things that shouldn't be connected, and does something upsetting for the user. A user asking Alexa to play a song that happened to contain something that could be mistaken for "turn on the lights" in its title, should never have Alexa mistakenly turn on the lights.



Rohit Prasad: You have to be careful when pushing boundaries.

"I'm more worried about that right now, because that would be very damaging," says Prasad. "What if there's a baby sleeping?"

To counter that, Amazon has put strict limits on Alexa, not on what it can learn from the public (the inputs), but on what it can actually do with what it's learned (the outputs).

In the early days of the experiment, the output of any unsupervised learning has to be limited to music.

Even if Alexa mistakenly decided that someone asking it to play a certain song meant it should open the garage door, it wouldn't be able to do so, because "cross domain" output, that in this example takes something learned in the domain of music and applies it to the domain of home automation, is blocked.

"That doesn't mean that we won't go cross domain, but it's too early to say how well that would work," Prasad told *The Australian Financial Review* in an interview.

The other risk, that the artificial intelligence learns something from the public that it shouldn't learn, has to be managed too, of course.

Many such errors would be corrected by the very same process that brought them about in the first place. If Alexa incorrectly learnt that Sirius XM Chill was on channel 63 rather than channel 53, and it started playing Christian music instead of deep house, then users would immediately start to barge in again, and Alexa would quickly correct its own error.

But what if Alexa learns something that doesn't immediately result in an error? After all, neural networks are essentially a "black box" to humans: humans can see what goes into them, and humans can see what comes out of them, but there's no way of telling what connections the computer has made inside them.

The experiment in unsupervised machine learning is being monitored by the team at Amazon in charge of maintaining the AI's personality, so it won't suddenly start making racist utterances, Prasad says.

"Clearly this is an area where we have to keep on applying a lot of judgment as we keep rolling it out."

And all of this is a far cry from what it would take for Alexa to learn something completely new, such as the desire to take over the world, and what to do with that desire.

"It's very hard for [Alexa] to learn a new concept and what it means. It can detect anomalies, it can detect patterns it doesn't understand, but to attach a concept to that, and for it to know what action to take on that concept, is a non-trivial problem."

Still, getting Alexa to learn new concepts for itself is something that would be desirable, says Prasad.

"Definitely we want it to become more teachable. But it's early days.

"When you're trying to push the boundaries, you have to be really careful. You have to exercise high judgment."