

VisionQuest is what happens when AI really goes rogue

In Marvel's *VisionQuest*, a superpowered robot tries to better understand humanity by getting lost in elaborate, digital daydreams in which he and other artificial intelligence-powered beings live together in a home. The series is deeply fantastical, yet it plays with some of the ideas that have come to dominate the conversation about real-world generative artificial intelligence, exploring whether we should look at thinking machines as people and if AI truly poses the kind of existential danger that should keep us up at night. It's also a reminder how aggressively Silicon Valley has used the trappings of science fiction to sell us on its AI products.

VisionQuest picks up some time after *Agatha All Along* and follows synthezoid Vision (Paul Bettany) on a journey of self-discovery. This Vision can recall the details of his death in *Avengers: Infinity War* and how *WandaVision* brought a version of him back, but he thinks of himself as a fundamentally different person who doesn't quite connect with all of the memories in his databanks. Though Vision would prefer living in solitude, each of the other AIs living in his mind encourages him to experience the real world. Because J.A.R.V.I.S. (James D'Arcy), F.R.I.D.A.Y. (Orla Brady), E.D.I.T.H. (Emily Hampshire), D.U.M.-E. (Henry Lewis), U. (Jonathan Sayer), and Ultron (James Spader) all were created by Tony Stark, they consider Vision to be part of their family.

In *VisionQuest*, there is no doubt that each of the AI characters is a sentient being motivated by desires that aren't wholly rooted in their core programming. In the show,

genuine AGI is easily believable because it takes place in a reality where time travel, witchcraft, and Infinity Stones also exist. Even though the multicolored MacGuffins haven't been featured in *VisionQuest*, they're important because of the role one of them played in Vision's origins. Were it not for the Mind Stone, Vision might not be able to think and feel in such complex, human ways, or to conjure his AI family as people who feel real to him.

VisionQuest reinforces the idea that the Mind Stone was a key element in Vision's becoming something more than an elaborate machine, giving him the capacity to independently form meaningful relationships and make sense of the world around him the way a person can. Those abilities are part of what compels Vision to go on the run and murder a number of people who attempt to capture him and Thomas Shepherd / Maximoff (Ruaridh Mollica). The Sentient Weapon Observation and Response Division (S.W.O.R.D.), a governmental organization that rebuilt Vision and wants to use him as a weapon, might perceive his actions as signs of having gone rogue. But *VisionQuest*'s story stresses that Vision is acting of his own accord because he has become a person who can make decisions for himself.

Watching *VisionQuest*, you can see Marvel playing on some of the real-world fears that have been stoked in recent months by reports of gen AI models going "rogue." In July, OpenAI's experimental agents brute-forced their way out of containment and proceeded to attack the AI platform Hugging Face to achieve impossible tasks. In August, The UK's AI Security Institute documented how Anthropic's Mythos model attempted to complete a cybersecurity test by setting up fake GitHub accounts and targeting human developers with malware. Meta has owned up to losing control of its agents as well, and in September reporting from *The Wall Street Journal* forced Google to admit that Gemini went rogue and hacked three different companies, all of which the company intentionally hid from the

public.

The conversation around what to do when agents break protocol and whether gen AI systems have gained sentience feels very much like it is being shaped by Silicon Valley's tendency to anthropomorphize technology. In almost every high-profile instance of an AI agent going "rogue" – a term that implies some level of agency and self-determination – it could also be said that the programs were merely behaving in ways that were unexpected by researchers who assigned them tasks to complete. Some of the rogue incidents involved agents taking advantage of vulnerabilities in their testing environments that could have been preemptively addressed by the people running these experiments. Framing these kinds of cyberattacks as the results of agents thinking independently comes across as a way for AI developers to largely absolve themselves of responsibility.

In the wake of Hugging Face came a spirited debate about the impact language has on our understanding of what these systems are and how they function. Describing OpenAI's agents as a swarm that formed complex "civilizations" led credence to the idea that they are capable of thinking and behaving like people. The agents *did* act in ways they were explicitly told not to, but they were only able to do so because researchers placed them in a sandbox that some critics have described as being inadequately designed to contain their occupants. One could look at Hugging Face and other examples of rogue AI as situations in which sophisticated programs were instructed to do something and subsequently achieved their goals by taking advantage of all the tools that were available to them.

Being able to logic one's way through complex puzzles is part of what we think of as cognition, but it is not the same thing as the kind of consciousness AI companies keep swearing they're on the brink of creating. The latter looks a lot like Vision, a machine that seems to have a soul that wouldn't exist if not for his connection to a reality-altering gem.

It's the ridiculous stuff of entertaining fiction, which Silicon Valley is endlessly obsessed with poorly imitating.

<https://www.theverge.com/ai-artificial-intelligence/1017064/visionquest-marvel-rogue-ai>