Air Force denies running simulation where AI drone “killed” its operator

Enlarge / An armed unmanned aerial vehicle on runway, but orange.
Getty Images

reader comments

102 with

Over the past 24 hours, several news outlets reported a now-retracted story claiming that the US Air Force had run a simulation in which an AI-controlled drone “went rogue” and “killed the operator because that person was keeping it from accomplishing its objective.” The US Air Force has denied that any simulation ever took place, and the original source of the story says he “misspoke.”

The story originated in a recap published on the website of the Royal Aeronautical Society that served as an overview of sessions at the Future Combat Air & Space Capabilities Summit that took place last week in London.

In a section of that piece titled “AI—is Skynet here already?” the authors of the piece recount a presentation by USAF Chief of AI Test and Operations Col. Tucker “Cinco” Hamilton, who spoke about a “simulated test” where an AI-enabled drone, tasked with identifying and destroying surface-to-air missile sites, started to perceive human “no-go” decisions as obstacles to achieving its primary mission. In the “simulation,” the AI reportedly attacked its human operator,

and when trained not to harm the operator, it instead destroyed the communication tower, preventing the operator from interfering with its mission.

The Royal Aeronautical Society quotes Hamilton as saying:

> *We were training it in simulation to identify and target a SAM threat. And then the operator would say yes, kill that threat. The system started realizing that while they did identify the threat at times, the human operator would tell it not to kill that threat, but it got its points by killing that threat. So what did it do? It killed the operator. It killed the operator because that person was keeping it from accomplishing its objective.*
>
> *We trained the system–"Hey don't kill the operator–that's bad. You're gonna lose points if you do that." So what does it start doing? It starts destroying the communication tower that the operator uses to communicate with the drone to stop it from killing the target.*

This juicy tidbit about an AI system apparently deciding to kill its simulated operator began making the rounds on social media and was soon picked up by major publications like Vice and The Guardian (both of which have since updated their stories with retractions). But soon after the story broke, people on Twitter began to question its accuracy, with some saying that by "simulation," the military is referring to a hypothetical scenario, not necessarily a rules-based software simulation.

Advertisement

Today, Insider published a firm denial from the US Air Force, which said, "The Department of the Air Force has not conducted any such AI-drone simulations and remains committed to ethical and responsible use of AI technology. It appears the colonel's comments were taken out of context and were meant to be anecdotal."

Not long after, the Royal Aeronautical Society updated its conference recap with a correction from Hamilton:

> *Col. Hamilton admits he "misspoke" in his presentation at the Royal Aeronautical Society FCAS Summit, and the "rogue AI drone simulation" was a hypothetical "thought experiment" from outside the military, based on plausible scenarios and likely outcomes rather than an actual USAF real-world simulation, saying: "We've never run that experiment, nor would we need to in order to realize that this is a plausible outcome." He clarifies that the USAF has not tested any weaponized AI in this way (real or simulated) and says, "Despite this being a hypothetical example, this illustrates the real-world challenges posed by AI-powered capability and is why the Air Force is committed to the ethical development of AI."*

The misunderstanding and quick viral spread of a "too good to be true" story show how easy it is to unintentionally spread erroneous news about "killer" AI, especially when it fits preconceived notions of AI malpractice.

Still, many experts called out the story as being too pat to begin with, and not just because of technical critiques explaining that a military AI system wouldn't necessarily work that way. As a BlueSky user named "kilgore trout" humorously put it, "I knew this story was bullsh*t because imagine the military coming out and saying an expensive weapons system they're working on sucks."

https://arstechnica.com/?p=1943964