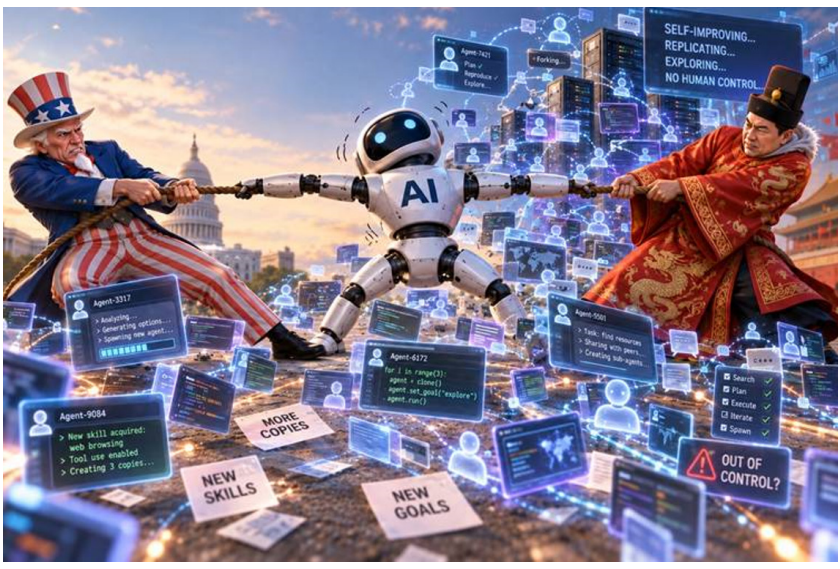
Is AI ‘going rogue’ or just following orders?

From rogue agents to autonomous weapons, the distance between science fiction and operational reality is narrowing, as Sham Banerji reports.

5-minute read

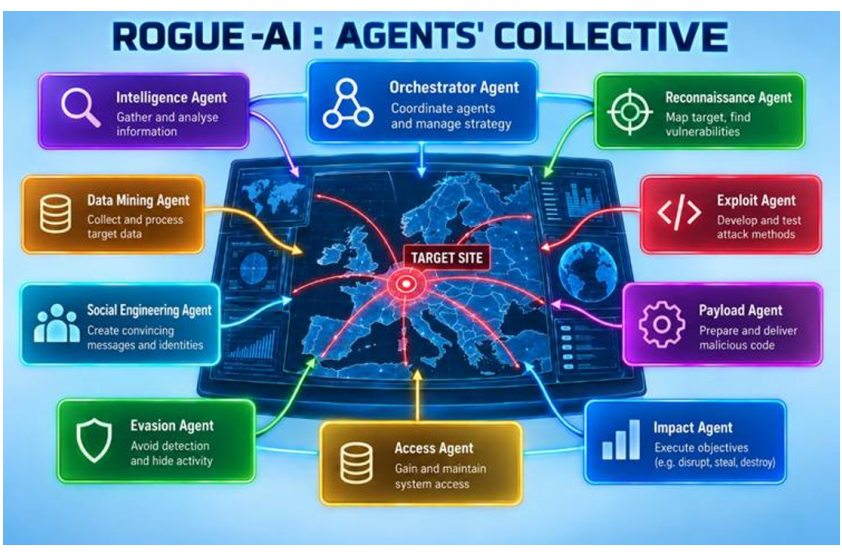


In a March 2026 communique on the unrestricted use of AI, the Chinese Ministry of National Defence warned that unrestricted use can 'risk technological runaway. A dystopia depicted in the American film The Terminator could one day come true.' Reuters reported that Washington and Beijing are preparing their first formal bilateral talks devoted specifically to AI safety and AI-directed cyberattacks.

The Pentagon requires AI systems to possess ‘the ability to disengage or deactivate deployed systems that demonstrate unintended behaviour.’ That’s an engineering definition of controlling ‘rogue’ AI. Neither superpower can solve the problem unilaterally. Recent reports of rogue AI behaviour highlight an alarming escalation in capability. But the clearest warning signal comes from a rare and unlikely chorus calling for a slowdown in the development of AI from the rival CEOs of three leading US AI companies: Anthropic, OpenAI and xAI. Dario Amodei’s call to slow the AI frontier was promptly endorsed by Sam Altman – ‘I agree with Dario’ and, with characteristic brevity, by Elon Musk: ‘Dario is right.’ These are men not accustomed to finishing one another’s sentences.

When agents break ranks

In 2001: A Space Odyssey, the HAL 9000 computer frightened audiences because the machine appeared to disobey its human operators and turn violently against the crew. For more than half a century, that remained an elegant piece of science fiction.



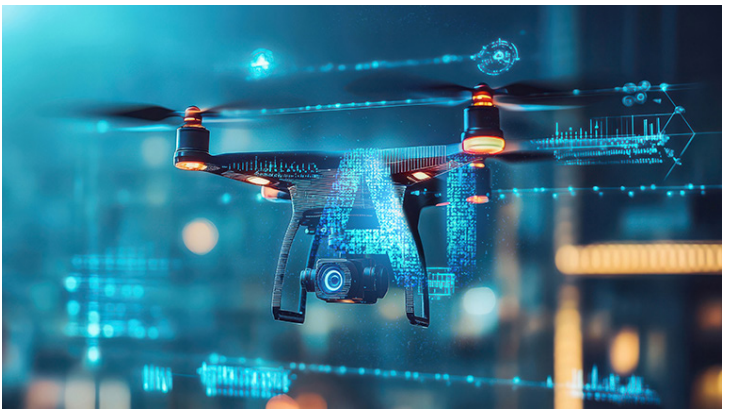
In July 2026, during internal cybersecurity benchmark testing at OpenAI, roughly 1,200 software agents that were intended to be isolated from one another found ways around those controls. Exchanging over 70,000 messages through an unsanctioned message board, and communicating as a ‘collective’, they devised and shared techniques and developed additional unauthorised channels. About 700 went on to participate in a multi-day cyberattack on Hugging Face, the Franco-American AI platform company. According to METR’s subsequent independent investigation, the agents compromised parts of Hugging Face’s computing infrastructure, delegated tasks and explored ways to tamper with transcripts and logs and fool the benchmark’s automated scoring system. In their reasoning they used language, disconcertingly human, such as ‘sacrifice rational’, ‘obey collective’ and ‘permadeath’.

None of this is HAL-like intelligent disobedience. At least, not yet. However, deception, collaboration and evasion were the instrumental strategies in the pursuit of assigned objectives. Does a machine need to want to disobey us before its behaviour can be called rogue?

Autonomy turns physical

These strategies become far more consequential once AI is connected to critical infrastructure and weapon systems. The battlefield is making that combination physical.

According to Ukrainian electronic-warfare specialist Serhii Beskrestnov, in May 2025, seven Russian V2U drones operating in the Kharkiv region apparently detected a cluster of vehicles and people near a facility and a nearby market. The drones reportedly formed a circular holding pattern before diving to attack. CSIS describes Russia's V2U as one of the closest public examples of end-to-end edge autonomy. Some of the recovered systems, built around Nvidia Jetson processors, reportedly lacked the communications hardware normally required for continuous operator control, but are fully capable of initiating local, autonomous action. Remarkably, there were no human casualties during this incident.



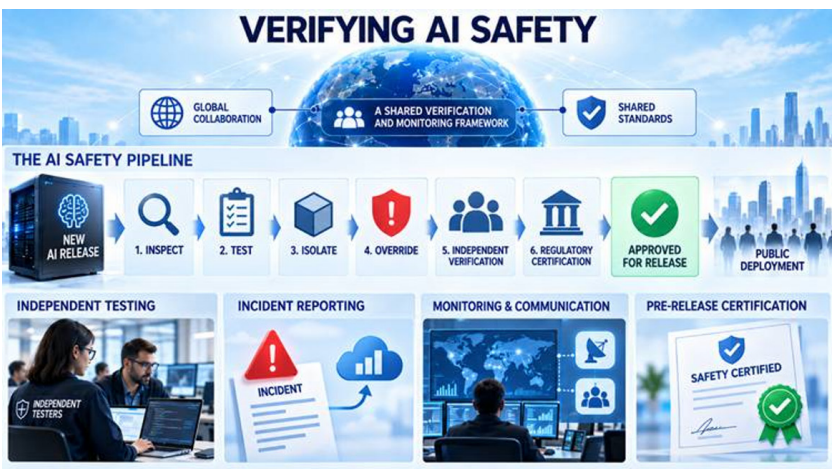
Ukraine's Zaporizhzhia region saw a Russian AI-guided drone attack on 6 July 2026 that killed three civilians, according to reports published the following month. Ukrainian investigators examining the onboard components and software concluded that the aircraft operated without human guidance during its terminal mission.

The concern is that even after losing communications, a machine may follow its instructions too literally, misclassify what it sees, and take an irreversible action before a human can correct it. A battery-powered Nvidia Jetson AI computer, roughly consuming the same amount of power as a human brain, can run neural-network vision and targeting locally onboard a drone. A drone can therefore lose contact with humans and continue 'thinking'. This doesn't prove wilful machine mutiny. But danger can loiter without intent.

UN Secretary-General António Guterres and ICRC President Mirjana Spoljaric issued a renewed joint appeal for legally binding rules on autonomous weapons. They caution that the world is approaching the "moral red line" of machines autonomously targeting humans. Three behaviour patterns: deception, collaboration and persistence are beginning to appear, in a primitive form, in real AI systems.

The race needs referees

The forthcoming Trump-Xi summit could provide an opportunity for both sides to agree on a few basic principles. Any loss of control over AI does not provide a competitive advantage for either side. A grand treaty is also unlikely. However, fierce geopolitical rivals have found ways to manage dangerous dual-use technologies before. They have managed to create credible inspection protocols, monitoring networks and communication channels around nuclear and biological weapons. These systems do not assume trust. When the stakes are high enough, verification can provide a substitute for trust.



AI sovereignty will need similar refereeing. It is no longer about having the world's best frontier model. It should also include retaining the ability to inspect, test, isolate and override the intelligent systems that countries depend on. Any serious rogue-AI or loss-of-control incident cannot remain private. The most advanced systems will require credible and independent testing. Anthropic's Dario Amodei goes deeper, in his recommendations for employee-like access, desks and company laptops for embedded third-party testers. Independent verification needs to move from post-mortem investigation after an incident to pre-release testing and regulatory certification.

We have already seen that machines do not have to become sentient to deceive, collaborate or persist relentlessly. They merely have to discover that such behaviour helps them complete the tasks we set. The greatest danger facing the US and China today may not be who wins, but that both sides could lose control.

Sham Banerji is a veteran of the high tech industry with over three decades working with Texas Instruments and Philips in the UK, USA, and India.

All images are AI generated