

P(doom) and the Cuba missile crisis 1962: An AI system in the place of Vasily Aleksandrovich Arkhipov?

Alexander Seger, 22 September 2026

In October 1962, humanity escaped from nuclear doom because of the personal judgment of a single Soviet naval officer named Vasily Aleksandrovich Arkhipov.

On 27 October 1962, during the Cuban missile crisis, a Soviet submarine - deep in Caribbean waters - had been out of contact with Moscow for days, when US destroyers were dropping depth charges around it.

The captain believed war had begun. He was prepared to launch a nuclear torpedo with a yield close to the Hiroshima bomb.

On this particular submarine, launch needed the consent of three officers. Two of them said yes; the third, Vasily Arkhipov, refused to authorize the launch.

Arkhipov had personal experience with nuclear radiation. A year earlier (July 1961), as deputy commander of the K-19 submarine, he had watched crewmates die of radiation after an accident.¹

By most accounts he prevailed against the two other officers through personal standing, not because he had better information.

The US nuclear war plan of the time, SIOP-62, had only one option in response to a nuclear attack: thousands of warheads launched at once against Soviet, Chinese and Eastern European targets. Soviet doctrine was similarly maximalist.

What if a modern AI system aligned with Soviet military objectives had been in the place of Arkhipov.

Would it have been the calm, rational voice that two exhausted officers needed? Or would it have authorized the launch?

The research is not reassuring:

- Rivera et al. (2024) tested five off-the-shelf language models in war games. All of them escalated. In some cases they deployed nuclear weapons.²

¹ https://nsarchive.gwu.edu/sites/default/files/2026-05/2026-05-18_spacedaily.com-the_soviet_naval_officer_who_refused_to_fire_his_nuclear_torpedo_during_the_cuban_missile_crisis_is.pdf

² Rivera et al.: <https://arxiv.org/abs/2401.03408>

- Kenneth Payne (2026) ran 21 simulated nuclear crises with frontier models. Nuclear threats appeared in every game, and mutual nuclear signalling in 95% of them. No model ever chose accommodation or withdrawal.³

This is not only a thought experiment. This month it was reported that an AI hallucination nearly triggered a US military operation against a Chinese vessel. Reportedly, the operation was aborted at the last minute, narrowly averting a potential conflict between the US and China.⁴

In that case the fabricated intelligence came from an AI, and people aborted the operation.

Arkhipov was later called "the man who saved the world".

Who holds the veto when the next decision runs through a machine? Would one entrust an AI system with Arkhipov's vote?

www.alexnotes.net

³ Payne: <https://arxiv.org/abs/2602.14740>

⁴ TechCrunch on the 2026 incident: <https://techcrunch.com/2026/09/18/ai-hallucination-nearly-triggers-us-military-operation/>