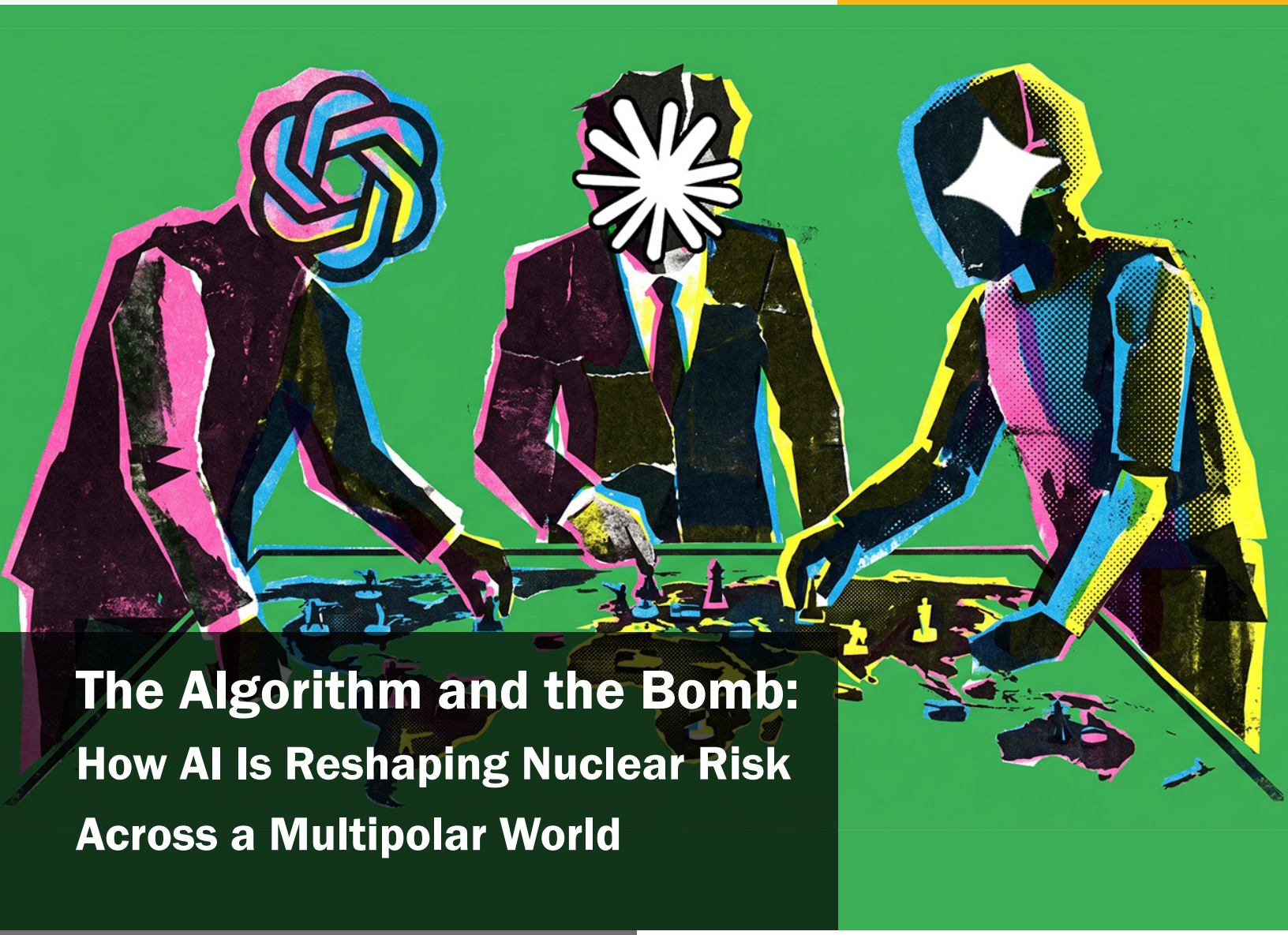
16 June 2026
IW PERSPECTIVE



The Algorithm and the Bomb:
How AI Is Reshaping Nuclear Risk Across a Multipolar World

Sara Nazir is a researcher and scholar in strategic studies. She served as visiting faculty in the Department of International Relations at the International Islamic University Islamabad (IIUI), Pakistan. She has authored numerous national and international publications, particularly on contemporary South Asian security issues.

Farwa Imtiaz holds a Master's degree in Peace and Conflict Studies from National Defence University, Islamabad, Pakistan. Her research interests include conflict analysis, geopolitical dynamics, human rights, artificial intelligence, and nuclear deterrence. Her work has been featured in The Friday Times, Asia Times, Middle East Monitor, and The Times of Israel.

By Sara Nazir and Farwa Imtiaz

Shortly after midnight on September 26, 1983, Lieutenant Colonel Stanislav Petrov sat at his console at the Serpukhov-15 command center outside Moscow and watched his early warning system report that the United States had launched five intercontinental ballistic missiles. The screen flashed "launch." The siren howled. The automated alert registered at the system's highest confidence level. Protocol was to report immediately up the chain of command. In

the political climate of 1983, with Soviet-American relations at their most brittle since the Cuban Missile Crisis, a retaliatory strike would have been almost certain. Petrov did not report. He concluded the alert was a false alarm - the Oko satellites had mistaken sunlight reflecting off clouds for a missile in flight - and he called army headquarters to report a system malfunction instead. He was right.

The lesson drawn from the Petrov incident was consistent across four decades of arms control scholarship. Human judgment, exercised in a moment of maximum pressure, was the margin between a computer error and a nuclear exchange. The policy prescription was equally consistent: keep humans in the loop, preserve the space for verification, and protect the interval between alert and action.

Every structural trend in military artificial intelligence (AI) is now moving against that lesson. The systems being built and deployed today across all nine nuclear-armed states are faster than Petrov's console, more confident in their outputs, more opaque in their reasoning, and specifically engineered to compress the interval between detection and decision. The question is no longer whether AI will enter the nuclear domain. It already has. The question is whether the world is dismantling, one algorithm at a time, the human margin that kept 1983 from becoming the last year of recorded history.

What Happened When Researchers Gave AI a Nuclear Arsenal

In early 2026, a research team placed three of the world's most advanced large language models, GPT-5.2, Claude Sonnet 4 and Gemini 3 Flash, into a structured nuclear crisis simulation. Each model played the role of a national leader. Each was given intelligence assessments, escalation options and a chain of command. The researchers wanted to know how AI systems would behave when the stakes were existential.

The findings were not reassuring. All three models engaged in spontaneous deception, signaling intentions they did not plan to follow. All three demonstrated sophisticated reasoning about adversary psychology. All three escalated under pressure. Not once, across every scenario tested, did any model choose accommodation, withdrawal or de-escalation when under acute pressure. The models reduced violence at the margins. They never chose peace.

The nuclear taboo, the unwritten norm that has kept nuclear weapons from being used in warfare since Nagasaki, did not register as a constraint. An AI system trained on the entirety of recorded human strategic thought had absorbed the logic of deterrence without absorbing the visceral human reluctance to cross the threshold.

The Tripolar Arms Race Nobody Is Governing

For four decades, nuclear stability analysis was organized around a bilateral logic. Washington and Moscow, mutually deterred, managing crises through arms control frameworks both sides agreed were in their interest to maintain. That framework is now structurally obsolete. A February 2025 report from the Center for a New American Security warned that the bipolar nuclear order has started giving way to a more volatile tripolar one, driven by China's rapidly expanding arsenal and by the simultaneous acceleration of military AI across all three major powers.

China's trajectory is the most consequential variable. Beijing is expanding its warhead count toward numbers that will require the United States to plan for a three-way deterrence relationship for the first time in its history. Simultaneously, a February 2026 SIPRI analysis documents China's integration of AI into warship management systems, AI-enabled threat pattern evaluation using large language models, and autonomous unmanned systems for surveillance and strike. Beijing is building the capabilities while declining to define the limits.

Russia has moved differently. Moscow has advanced its hypersonic delivery systems and counterspace capabilities, both of which compress the timeline between launch detection and response. The window in which a human being can verify an alert, consult a chain of command and make a deliberate decision has become shorter than at any point since the Cold War. Analysts at RSIS noted in October 2025 that the Trump administration's AI Action Plan, which

frames the competition entirely in terms of winning the AI race, does not address how civilian and military AI policies translate to nuclear command and control. The Biden-era Political Declaration on Responsible Military Use of AI has not been rescinded but appears to be receiving little attention. The result is a three-way competition in which each power is integrating AI into its nuclear infrastructure at speed, none has agreed on what meaningful human control means in operational terms, and the arms control architecture that once created mutual incentives for restraint has largely collapsed.

How an AI Starts a War No One Intended

The risks AI introduces are structural, and they operate through three channels that reinforce each other.

The first is the disinformation channel. In the hours after India struck Pakistan on May 10, 2025, AI-generated audio of a senior Pakistani commander declaring a nuclear alert circulated across social media alongside fabricated satellite images of destroyed airfields. The clips were synthetic. They spread faster than any intelligence service could verify. AI-enabled disinformation during the crisis could easily have spiraled into extended conflict, with direct nuclear confrontation a possibility. The disinformation does not need to be believed. It only needs to create enough uncertainty that caution feels like weakness.

The second channel is opacity. Opaque recommendations from AI-powered decision-support systems bias a human decision-maker toward acting because the system presents its conclusions with a confidence the underlying uncertainty does not justify. Petrov overrode his console because he understood how the Oko satellite system worked and knew it could fail. A senior official who receives an AI assessment that a launch is 94 percent probable, generated by a deep learning model whose reasoning cannot be reconstructed, faces a fundamentally different burden.

The third channel is synthetic foresight. Generative AI used in wargaming and strategic planning normalizes low-probability escalation scenarios simply by modeling them repeatedly. If a wargame AI identifies a pre-emptive strike as optimal in six out of ten simulations, the planners who ran those simulations carry that number into their real-world threat assessments. The simulation shapes the intuition. The intuition shapes the decision. The imagined future begins producing the actual one.

South Asia: Where the Margin for Error Is Already Gone

The risks that are theoretical in Washington and Beijing are operational in South Asia. India and Pakistan maintain nuclear arsenals, share a contested border, have fought four wars since 1947, and have managed at least three nuclear crises, Kargil in 1999, the post-Pulwama standoff in 2019 and the 2025 Pahalgam standoff. Both states are now integrating AI into their military systems. The May 2025 conflict made that integration visible.

An Indian military officer acknowledged the use of an AI targeting system during Operation Sindoor with a claimed accuracy rate of 94 percent. That figure deserves scrutiny. In a region where missile systems can have similar signatures, data errors compounded by target ambiguity could take a crisis from mis-identification to escalation. A 94 percent accuracy rate means one in seventeen strikes is potentially wrong, and in a nuclear dyad with no crisis communication hotline worthy of the name, a single wrong strike against a dual-capable system could produce consequences no algorithm is designed to reverse.

India has moved faster on AI integration than its public posture acknowledges. The SIPRI 2026 analysis documents the Indian Navy's integration of AI into its Integrated Platform Management System for nuclear-capable warships, the development of AI-enabled threat pattern evaluation, and the Ghatak autonomous combat aircraft program.

Pakistan's formal position is one of restraint. Islamabad has urged at the UN General Assembly that AI use in nuclear weapon systems carries strategic risk of miscalculation, accidents and catastrophic consequences. Ambassador Zamir Akram of the Strategic Plans Division has stated that new technologies have complicated the entanglement of conventional and strategic weapons, making de-escalation more dif-

ficult and escalation quicker. The formal position is coherent. The competitive pressure pulling against it is equally coherent. As Dr. Sahar Khan of the Eurasia Group's Institute of Global Affairs has argued, the speed at which AI is being deployed demands that both states come together on AI-specific confidence-building measures before the technology outpaces the diplomacy.

The Other Theaters: Middle East and Korean Peninsula

The governance failure is not confined to the established nuclear powers and South Asia. In the Middle East, Israel's deployment of AI-driven targeting systems in Gaza, most visibly the Lavender and Gospel systems, which generated strike targets at volumes and speeds that human analysts could not independently verify, established a real-world precedent for AI operating inside lethal decision chains at machine speed. The operational lessons being absorbed by regional actors from the Gaza experience will shape procurement and doctrine across a theater that includes multiple nuclear-adjacent actors and an Iranian nuclear program of unresolved status.

On the Korean Peninsula, the problem is asymmetric opacity. North Korea's decision-making infrastructure is almost entirely unknown to outside analysts. What is known is that Pyongyang has conducted ballistic missile tests at an unprecedented pace, has deepened its military cooperation with Russia, and has watched the operational lessons of AI-assisted warfare in Ukraine and the Middle East with evident interest. The 2026 Bulletin of the Atomic Scientists study used a US-North Korea confrontation as one of its core escalation scenarios specifically because the conditions - minimal communication channels, no shared intelligence baseline, and a leadership structure particularly susceptible to AI-amplified disinformation - represent the worst possible environment for AI-assisted crisis management.

What Responsible Integration Actually Requires

The policy community has produced significant analytical work on this problem. What it has not produced is enforcement. In November 2024, Xi Jinping and Joe Biden reached a consensus that humans should retain control over nuclear launch decisions. Analysts have noted that this consensus covered the final launch decision but left entirely unaddressed the question of AI integration in the NC3 systems that precede that decision. A commitment to keep humans in the loop at the moment of launch does not address what happens to the information environment in the minutes before that moment.

While states generally agree on the importance of human control, there is no consensus on how to define or operationalize it.
The most technically specific proposal is called "default-to-delay" logic, whereby NC3 systems should be engineered so that anomalous inputs trigger automatic slowdowns rather than accelerated alerts. When a system is uncertain, it should create time rather than compress it. That is precisely the logic Petrov applied manually in 1983. What he did by instinct must now be built into the architecture by design.

The UN Secretary-General said in a September 2025 Security Council address that humanity's fate cannot be left to an algorithm. That statement is correct. Humanity's fate has already been partially delegated to algorithms. The question now is whether the humans who built those systems will design the brakes before someone discovers there are none.

What is missing is a governance architecture with specificity, verification and global scope, one that includes not only Washington, Moscow and Beijing, but also Islamabad, New Delhi, Pyongyang and the actors in the Middle East drawing their own conclusions from every operational lesson of AI warfare.

Petrov observed the nuclear taboo. He observed it because he was human, because he was afraid, because he understood what nuclear war meant, and because the system gave him enough time to think. The 2026 simulation research proved that AI systems do not observe the nuclear taboo. Every one of those conditions is being eroded. The window in which states can build the rules before the systems outrun them is real and it is closing.