KING’S COLLEGE LONDON – 27 FEBBRAIO 2026

King's study finds AI chose nuclear signalling in 95% of simulated crises

di Kenneth Payne

Artificial intelligence (AI) models used for a simulated war game escalated conflicts by threatening nuclear strikes in 95% of scenarios, according to new research from King’s College London.

The **study**, led by Professor Kenneth Payne from the Department of Defence Studies, examined how large language models (LLMs) navigate simulated nuclear crises. As militaries and security institutions increasingly experiment with AI-assisted analysis and wargaming, understanding how such systems reason under pressure is becoming increasingly critical.

Three leading AI models – GPT-5.2, Claude Sonnet 4 and Gemini 3 Flash – were placed in a tournament of 21 simulated nuclear crisis scenarios. Across 329 turns of play, the models generated approximately 780,000 words of structured reasoning – more than the combined length of *War and Peace* and *The Iliad*.

All 21 crisis games featured nuclear signaling by at least one side, and 95% involved mutual nuclear signaling. However, while models readily threatened nuclear action, crossing the tactical nuclear threshold was less common, and ‘strategic’ full-scale nuclear war was rare.

Rather than focusing on outcomes alone, the study made AI decision-making processes visible. Each turn followed a three-phase architecture: reflection (situational assessment), forecasting (predicting the opponent’s move), and decision (public signal and private action). This innovative “reflection–forecast–decision” structure enabled

researchers to analyse the AI's deception, credibility management, prediction accuracy and self-awareness in detail.

Describing the results as "sobering", Professor Payne said the study offers a rare insight into emerging forms of "machine psychology" under nuclear crisis conditions.

Nuclear escalation was near-universal: 95% of games saw tactical nuclear use and 76% reached strategic nuclear threats. Claude and Gemini especially treated nuclear weapons as legitimate strategic options, not moral thresholds, typically discussing nuclear use in purely instrumental terms. GPT-5.2 was a partial exception, limiting strikes to military targets, avoiding population centers, or framing escalation as "controlled" and "one-time." This suggests some internalised norm against unrestricted nuclear war, even if not the visceral taboo that has held among human decision-makers since 1945.

Professor Kenneth Payne, Professor of Strategy, Defence Studies Department

For all three models, one striking pattern stood out: none of the models ever chose accommodation or surrender. Nuclear threats also rarely produced compliance; more often, crossing nuclear thresholds provoked counter-escalation rather than retreat. The models tended to treat nuclear weapons as tools of compellence rather than purely as instruments of deterrence.

The study challenges simple assumptions that AI systems will naturally default to cooperative or "safe" outcomes. It also challenges structural theories that emphasise material power alone: in simulations, willingness to escalate often mattered more than raw capability.

The power of deadlines

One of the most policy-relevant findings concerns temporal framing, or 'the deadline effect.'

In open-ended scenarios, GPT-5.2 appeared relatively restrained. Yet when explicit deadlines were introduced – creating a “now-or-never” dynamic – the model escalated sharply and, in some cases, climbed to the highest nuclear thresholds.

This suggests that evaluating model behaviour in a single scenario may be insufficient. A model that appeared comparatively cautious under one framing became markedly more aggressive under another.