

ARTIFICIAL INTELLIGENCE AND NUCLEAR DETERRENCE: CHALLENGES OF AI INTEGRATION INTO THE NUCLEAR DECISION-MAKING PROCESS

Bogdan STOJANOVIĆ¹

ABSTRACT

This paper seeks to explore the relationship between the latest artificial intelligence (AI) technologies and nuclear deterrence, a theoretical and practical concept that has existed in international politics from the Cold War to the present day. The core research question addressed in the analysis is whether the integration of AI into the nuclear deterrence process would contribute to stability or, conversely, whether reducing the human factor in nuclear decision-making would increase the risk of nuclear weapons use. The cost-benefit analysis of artificial intelligence in such a sensitive area as nuclear decision-making is most effectively examined through the realist lens. Also, in the theoretical framework, the author introduces his own concept of “constructed rationality”. The issue of responsibility in the case of the human factor being displaced in favour of AI, creates new challenges and complicates the process of nuclear deterrence. The hypothesis of nuclear peace has proven valid in practice, because mutually assured destruction (MAD) makes war between nuclear powers highly unlikely. AI has already found its place in the military domain, making it justifiable to examine how the application of cutting-edge technology would affect the functionality and predictability of nuclear deterrence among nuclear weapon states. The author concludes that the full integration of AI systems into nuclear decision-making would undoubtedly increase the likelihood of deliberate or inadvertent nuclear escalation.

ARTICLE HISTORY

Received:

01/04/2026

Revised:

05/07/2026

Accepted:

31/07/2026

KEYWORDS

AI; artificial intelligence; nuclear deterrence; nuclear decision-making; nuclear weapons; constructed rationality; nuclear war.

Cite this article as: Stojanović, Bogdan. 2026. “Artificial Intelligence and Nuclear Deterrence: Challenges of AI Integration into the Nuclear Decision-Making Process”. *The Review of International Affairs* LXXVIII (1198): 567–601. https://doi.org/10.18485/iipe_ria.2026.77.1198.2

¹ Senior Research Fellow, Institute of International Politics and Economics, Belgrade, Serbia. E-mail: bogdan.stojanovic@diplomacy.bg.ac.rs, <https://orcid.org/0000-0002-9624-1516>

Introductory Considerations

Some of the greatest challenges to contemporary international security are the various threats posed by the uses of artificial intelligence (AI) within the context of nuclear weapons and their associated decision-making systems. Although, arguably, technological advancements in military capabilities will always continue on course, the incorporation of AI into the realm of nuclear weapons will present unique challenges that can not be found in other types of military application. Given the level of destruction available to mankind using nuclear weapons on a global scale, the introduction of AI will necessitate different ways to analyse the speed and autonomy of decision-making. While nuclear and AI technology are both usually thought of in terms of increasing security, there is also a counter-argument that AI will fundamentally disrupt the deterrent logic that has maintained a stable international system for more than eight decades. Within defence circles, AI is widely viewed as a force multiplier, which will change the nuclear command, control, and communications (NC3) architecture dramatically. As AI is integrated into early warning systems and into national reconnaissance, surveillance, and intelligence systems (ISR), the impact it will have on existing deterrence strategies and strategic stability doctrines will have to be assessed. The development and deployment of AI in this realm marks a qualitative shift in the development of nuclear deterrence and of strategies for maintaining international security more broadly. While AI promises increased efficiency, rapid processing of vast amounts of new data, and a heightened ability to analyse complex information, it presents new and unprecedented risks for global stability. Artificial intelligence is not an inherently destabilising technology in the nuclear context. Rather, its risks and benefits are contingent upon specific applications and the extent to which “appropriate levels of human judgment” are preserved in critical decisions (Lin 2025).

Key risks include: the opacity of algorithms, the possibility of data manipulation, the erosion of authentic human control despite the nominal preservation of the “man in the loop”, the divergent logics of the various nuclear powers, the compression of the decision-making window, and the strengthening of the risk of misinformation and non-compliance. The Responsible AI in the Military Domain (REAIM) summit held in Spain in February 2026 illustrated the deep-rooted strategic suspicion of countries that have advanced the most in the field of this application. The United States of America (USA) and China have given up on signing a joint declaration for the responsible use of AI in the military domain because any establishment of normative frameworks would lead to a classic prisoner’s dilemma and the loss of an already achieved advantage (Waldersee 2026). A fundamental problem of future development is embodied in the questions of how to preserve majority human control over the most destructive force in the world, and how to monitor and verify something that is inherently opaque in its nature? As leading

pioneers in machine learning have warned, AI has the potential to have emergent autonomy from human oversight and produce untruthful or misaligned output. Geoffrey Hinton has pointed out that highly advanced AI may develop instrumental purposes that differ from what we want them to do. Therefore, it will become exceedingly difficult to have reliable control over decisions that carry significant consequences (Brown 2023). In order to avoid rapidly implementing AI systems into nuclear command, control, and communications architectures without having an adequate regulatory mechanism, the international community needs to work together towards establishing common laws to protect against adversarial manipulation of these systems. Without a cooperative and comprehensive strategy, the international community is likely to arrive at the paradox that, in seeking technological dominance, it has simultaneously created an environment of increased nuclear risk.

Existential Nuclear Deterrence and Constructed Rationality in International Relations

When analyzing the implications of a new technology, it is necessary to first evaluate how that technology can affect a broader conceptual framework. Nuclear deterrence was first developed as a means of national security during the Cold War, yet the theoretical basis of deterrence has roots in even human nature itself. At its most basic level, deterrence is the use of threatened force preemptively to influence another state or non-state actor to behave in a certain manner in order to achieve strategic objectives (Downs 1989, 226). In the case of nuclear deterrence, a state attempts to deter another from attacking by seriously considering using a nuclear weapon on the aggressor. This model of deterrence is contingent upon the assumption of rationality. The costs of becoming a target of a nuclear strike are likely to be greater than the potential gains (George and Smoke 1974, 11). Nuclear deterrence is based upon a realist view of the nature of international relations. Realists tend to view the major actors in international relations as self-interested unitary actors seeking to achieve different goals. Security, or survival, is the paramount interest of states and is usually the major goal of foreign policy. The famous Cold War strategist Bernard Brodie astutely observed: “Thus far the chief purpose of our military establishment has been to win wars. From now on its chief purpose must be to avert them” (Brodie 1946, 76). Roughly speaking, all the desires to go to war follow in descending order from the desire to win a battle, and all the desires to avoid war follow in ascending order from the desire to avoid all risk to one’s security and thus to prevent war altogether. The paradoxical nature of nuclear weapons lies in their simultaneous capacity to terminate human existence and to dissuade nuclear powers from engaging in direct interstate war. This stems from the conjunction of a rational pursuit of survival and the cataclysmic consequences of

potential nuclear exchange (Mutual Assured Destruction, or MAD). This rational deterrence against aggression, grounded in the imperative of state survival, is further reinforced by the ostensibly irrational category of human fear. Certain theorists characterise this configuration as a “balance of terror”, wherein rational cost-benefit calculations are supplemented by the apprehension of political elites regarding total annihilation (Wohlstetter 1959). It would therefore be entirely appropriate to speak of “deterrence through intimidation”, rather than exclusively of rational deterrence, when analysing relations among nuclear-armed states.

At the theoretical level, two fundamental approaches to deterrence can be delineated, although the modalities and effects of deterrence remain consistent across both. The first approach is essentially rationalist, in the narrower, economic sense, as it emphasises cost-benefit analysis within the deterrence framework. According to this model, a given state may refrain from aggression due to the anticipation of unacceptably high costs that would inevitably follow. The alternative approach is fundamentally psychological, prioritising the concept of “fear” of destruction as the primary motivator for abstaining from attack against a nuclear-armed state. Within this framework, survival is understood as an innate instinct of every individual, and by extension, of every state. Despite their respective methods of analysis, the economic and psychological approaches to deterrence can be regarded as two sides of a single analytical coin. Thus, countries will either not wage nuclear war because they have calculated the costs and benefits rationally, or will refrain from provoking conflict because they have an emotional fear of being destroyed. Therefore, due to the breadth of the conceptual framework, the term “existential nuclear deterrence” should be preferred to “rational nuclear deterrence”, although the latter dominates the scientific literature (Bundy 1983, 4). There is a reciprocal relationship between decisions based on emotion and those based on reason. Any decision not to engage in nuclear exchange has rational consequences, because nuclear war can never be useful or cost-effective. As Robert Jervis explained clearly, successful deterrence does not require that political leaders are perfectly rational (Jervis 1979, 299). It simply requires that decision-makers have enough awareness of reality that the costs of engaging in nuclear war will far outweigh any possible gains. This basic idea goes beyond being a simple logical concept and is composed of three key elements: strategic reasoning, psychological characteristics, and the reality that there is an existential threat to human existence created by nuclear weapons.

Kenneth Waltz reached similar conclusions when he analysed how well deterrence strategy would work operationally. According to him, rational decision-makers do not need to do extensive cost-benefit analyses for an understanding of nuclear war. They only need to recognise that there are extremely high “costs” associated with engaging in such a war. Waltz calls this the “assumption of rationality”, which represents the minimal requirement for deterrence to work (Waltz 1981). All forms of leadership, either by rational or irrational leaders (even Hitler as perhaps the best cited example

of an irrational leader), should be able to refrain from attacking a country that could use nuclear weapons. Looking at this issue through an ontological perspective, rationality exists conceptually in an exceedingly unstable environment of being able to function and be operationally effective.

An objective event must be cognitively processed in order to determine the action to be taken. Human behaviour, and the global behaviour of all nations collectively, stems from a subjective construction through their ideas, norms, values and psychological characteristics (Stojanović 2013, 150). Building upon the conceptual framework developed in prior research, the phrase of “constructed rationality” best describes the practical effect of rationality because every phenomenon of objective reality must be mediated by a human filter comprised of normative frameworks, value systems, and non-rational dimensions before a decision regarding specific action is reached (Stojanović 2013, 151). Inference is something that simply happens to us as cognition constantly goes beyond the available data (Moshman 2004). On a second level, thinking is the deliberate use of our inferences to pursue a goal, and finally reasoning arises only when that thinking becomes self-constrained by norms the thinker holds to be justified (Moshman 2004). Reasoning always depends on the personal quality of the one who reasons.

Even with huge limitations for a deterrent strategy from the theoretical and practical perspectives, nuclear deterrence has worked so far and will keep working without any doubt. Game theory represents a dynamic and general framework for analysing the impacts of rational behaviour, calculating risks, and miscommunication in all strategic interactions between states (Lindelauf 2020). Understanding the essence of constructed rationality and the operational logic of a deterrence strategy is crucial for the critical analysis of the effects of AI integration into the nuclear strategic “game”. The integration of AI in the strategic decision-making process in nuclear-armed states effectively strips interstate relations of their subjective, normative and affective dimensions, transforming them into an ideal-typical model of rational behaviour. This approach to interstate interaction, paradoxically, increases the structural risks of nuclear conflict, precisely because it eliminates the pragmatic hesitations, ethical inhibitions and contextual judgements that in the past functioned as a buffer against war.

Nuclear deterrence is the strategy of using military force in an indirect fashion for political objectives. In the nuclear age, the major strategic thought has shifted from the general staff to the civilian cabinets. During the Cold War, strategy and politics became inescapably intertwined with great difficulty in drawing a line between them. Given the capacity of nuclear weapons for total destruction, strategic thinking shifted from military to civilian cabinets. Consequently, this period is often referred to as the “golden age of strategic studies”, as political relations during that time were fundamentally shaped by the nuclear strategies of the superpowers. This transition carries considerable significance, as nuclear deterrence increasingly

became an instrument wielded predominantly by political leaders rather than military commanders. AI may lead us to a new shift, a turning away from civilian cabinets to processors. James Johnson claims that “in the context of AI and autonomy, particularly information complexity, misinformation, and manipulation, rationality-based deterrence logic appears an increasingly untenable proposition” (Johnson 2023, 202-203). This technology has added another level of complexity to nuclear deterrence, probably the most delicate domain of decision-making.

Concept and Application of Artificial Intelligence

The famous book “The Singularity Is Near: When Humans Transcend Biology”, written by renowned developer Ray Kurzweil (2005), anticipated the rapid development of technology to the point where it would surpass human intelligence. This so-called “technological singularity” would be a point in time in history where human intelligence is no longer the primary form of intelligence. Artificial intelligence is a computer system that attempts to imitate and match human intelligence. The goal of developing AI is to create a computer model that learns, improves, and solves increasingly complex problems. As previously discussed, deep learning is such an attempt to create a system that learns like a human brain and therefore gets progressively “smarter”. In fact, IBM’s Deep Blue, a chess-playing computer, beat world chess champion Garry Kasparov in 1997 (Yao 2022). The engines for playing chess have come a long way since, to the point that humans are no longer competitive. But chess is just one field where artificial narrow intelligence (ANI) is driving human professionals from their desks. People are being pushed out in data analysis, medical diagnostics and countless other speciality roles. Marvin Minsky’s work (1954) on neural networks and artificial intelligence in simulating brain function forms the basis of most subsequent AI research, including the pursuit of artificial general intelligence (AGI), the holy grail of reasoning at its highest level. There is considerable concern, and even fear, as the next generation of intelligent machines is being created at a very rapid pace. Much of the fear expressed about AI is in the media regarding the displacement of many people in various areas.

The dual-use nature of AI (as with all technology) is that it can be used in a positive sense (e.g. detecting tumours before they can be seen through human vision, helping solve math problems differently than a human, and speeding up the engineering design process) or in a negative sense, creating an inhumane society. Although the roots of AI and experimental endeavours in this field extend back several decades, a decisive breakthrough in this technology was achieved only in recent years. It is a noteworthy coincidence that the war in Ukraine, which fundamentally altered the landscape of global politics, commenced in February 2022, while later that same year, OpenAI launched ChatGPT, probably the most famous large language model. According to Sam Altman (2022), the company’s chief

executive, this AI application amassed its first million users within merely five days of its release. The conflict in Eastern Europe has thrust the issue of nuclear weapons deployment back to the top of the political and military agenda. Assessments suggesting that Russia might resort to tactical nuclear weapons, notwithstanding their low probability, have captured the attention of the global public. Thus, this temporal coincidence has prompted critical inquiry into the symbiosis of two distinct narratives. On one hand, the prospect of initiating a nuclear armageddon and on the other, the capacity of AI to reason more effectively than human cognition. It is imperative to bear in mind that over 12,000 nuclear warheads currently exist worldwide, a reality that renders the threat of annihilating human civilisation profoundly credible (Davenport 2026).

The rapid growth of AI within the consumer market for goods and services is not just a significant scientific achievement, but also opens the door to a major application of AI for warfare purposes. As history tells us, dual-use technologies always migrate across the civil-military realms, from commercial to military applications and vice versa. Historical examples include microchips, satellites, the internet, rocketry and nuclear power. Humans continue to be the most successful beings because of their ability to learn, think and acquire newly created ideas and skills, store and pass them down through time. However, in constructing an environment to meet their own needs, humans have been able to create something that could eventually replace them. An entity often referred to as the “final invention”. The introduction of Artificial General Intelligence (AGI) would be as important as the control of fire or the discovery of electricity. The question of whether and when we will introduce AGI that will perform all cognitive functions more effectively than humans remains to be seen. However, the urgency of applying current AI technologies in military development has increased, especially concerning weapons of mass destruction.

Similarities and Differences between AI and Nuclear Weapons

Both inventions, nuclear weapons and artificial intelligence, simultaneously embody a technological leap forward for human civilisation and a potential instrument of mass destruction (either physically or otherwise). During the Cold War, scientists, policymakers, and military strategists grappled with a fundamental question regarding nuclear weapons: would they ultimately annihilate the planet and its inhabitants, or would the equilibrium of nuclear power prevent large-scale war? The hypothesis of „nuclear peace“ proved largely correct, as the cataclysmic nature of these weapons effectively deterred direct military confrontation between the world’s foremost nuclear powers (the United States and the Soviet Union) (Stojanović 2021). A parallel inquiry confronts us today in relation to AI: will it ultimately safeguard or destroy humanity? Just as leading scientists at the dawn of

the nuclear age called for stringent control on the further development of nuclear weapons, prominent figures in the field of AI have signed an open statement asserting that „mitigating the risk of extinction from artificial intelligence should be a global priority alongside other societal-scale risks such as pandemics and nuclear war“ (Abdul 2023). Elon Musk, the world’s wealthiest individual and a significant investor in emerging technologies, has gone further, contending that AI poses a far greater existential threat than nuclear weapons (Clifford 2018).

The development of both AI and nuclear weapons necessitates substantial investment in physical infrastructure. Beyond the requisite scientific expertise, the construction of a nuclear device demands extensive and costly facilities for producing fissile material. Processes such as uranium enrichment and plutonium reprocessing remain the preserve of only the most technologically advanced states. Artificial intelligence likewise demands substantial physical resources, epitomised by the most sophisticated microchips, a capability currently concentrated in the hands of only a few states. Thus, both technologies carry the potential to pose existential threats to humanity as a whole. Both are amenable to civilian and military applications. Finally, the development of AI and nuclear technology has unfolded and continues to unfold within the context of an “international race”. But now that race also includes private corporations. Notwithstanding these essential similarities concerning resource dependencies, technological revolutionary potential, and risks to humankind, AI and nuclear weapons technologies diverge significantly across numerous dimensions. Whereas nuclear proliferation is constrained by the scarcity of fissile materials and the visibility of enrichment infrastructure, the diffusion of AI capabilities faces fewer material bottlenecks and may proliferate through software channels that are inherently more difficult to monitor and verify. The intangible nature of AI algorithms renders their proliferation to a broader range of actors markedly easier compared to the cumbersome physical infrastructure for nuclear programs (reactors, fissile materials, or centrifuges). Safeguards and protective measures can be relatively easily circumvented through modifications to source code, effectively precluding meaningful control over dissemination once an AI model has been initially released. This fundamental asymmetry between the material constraints governing nuclear proliferation and the frictionless diffusion of software-based capabilities carries profound implications for international governance, arms control, and the design of regulatory regimes in the twenty-first century. Moreover, while the destructive consequences of nuclear weapons are immediate, physical, and geographically bounded, the potential harms of advanced AI systems may be indirect, systemic, and globally diffuse, ranging from algorithmic bias and cyber vulnerabilities to the erosion of human influence in critical decision-making processes. Artificial intelligence has emerged primarily from the private sector, and its development is characterised by a diffuse ecosystem encompassing numerous non-state actors, a structural feature that significantly complicates any prospective efforts to control its proliferation. Whereas

nuclear weapons remain confined to the secured arsenals of the world's most powerful states, AI is utilised by millions of individuals worldwide, with each interaction contributing incrementally to the iterative refinement and „intelligence“ of these systems. Nuclear technology permits a relatively clear demarcation between civilian and military applications. Civilian nuclear programs are legally permissible under international law, which explains why more than thirty states currently operate nuclear reactors for peaceful purposes. Simultaneously, the Treaty on the Non-Proliferation of Nuclear Weapons (NPT 1968) explicitly prohibits the spread of nuclear weapons as the military component of this technology. By contrast, the diverse applications of AI cannot be neatly disaggregated, rendering the separate regulation or prohibition of specific uses exceptionally challenging. Consider, for instance, the same underlying capability that enables the rapid drafting of corporate emails can equally facilitate the swift dissemination of disinformation at large scale.

Instead of trying to restrict the development of AI through technology-based regulations, a better approach may be to focus regulations on how AI is used. Historical examples can help us in that matter. For instance, the Agreement on the Prevention of Incidents on and over the High Seas between the U.S. and U.S.S.R. (APIHS 1972), did not limit the development of military ships but created operational protocols to regulate behaviour by requiring the notification of hazardous manoeuvres. Similarly, the Outer Space Treaty (1967) offers a normative framework under which countries are held accountable for the actions of their private entities (Stojanović and Dimitrijević 2025, 312-323). This may incentivise countries to establish similar regulations for corporations developing and employing AI applications. It is essential to understand the relationship between AI and nuclear weapons in order to support nuclear deterrence both theoretically and practically. Introducing AI into this area would create a level of almost irreversible commitment. Using one high-risk technology (AI) to control the actions of another (nuclear weapons) dramatically increases the potential for catastrophic results for humanity if things go wrong.

Accountability and Other Problems in the Military Use of AI

Military applications of AI have immense potential to enhance current capabilities in surveillance, command and control, data analysis, and the remote control of manned or unmanned platforms (Horowitz 2018). Although still in its infancy, we can expect that next-frontier technologies will be developed and deployed at an unprecedented pace within the next several decades. The pressure to deploy AI-enabled systems on currently operating platforms will demand radical transformation of current platforms and force military strategists to rethink conventional strategic thinking. Next-frontier technologies have the potential for disruptive military innovation, transitioning from an industrial-age warfare model

through information warfare to third-generation systems that incorporate fully autonomous weapon systems (AWS). The introduction of AI algorithms into the military arsenal has opened a new chapter in the practice of using war technology. The impact of using AI in the armed forces is increasingly referred to as the AI revolution in military affairs (AI-RMA) (Raska 2021). From a strategic perspective, the introduction of AI to the military domain is forcing states to re-think their warfighting strategies and to adjust their current doctrines to better confront the strategic opportunities and threats that AI technology presents.

Control of AI presents another challenge in that it cannot be perceived as a weapon. A useful analogy for this could be radar technology. Modern military applications demonstrate that without its capacity for target acquisition, missiles remain effectively “blind”. Similarly, AI serves as a military force booster, processing vast datasets and providing real-time synchronisation of military units. Integrated military strategy with AI employs the totality of all weapon systems, operating in the air, on land, at sea, in outer space, and in cyberspace in order to counter meet strategic objectives. The use of AI technology by armies around the world is increasing in order to improve combat efficiency and to support a range of multi-domain military operations. AI supports situational awareness, allowing personnel to make better use of available assets and to make informed decisions (intelligence, surveillance and reconnaissance - ISR). Military commanders and decision-makers can quickly process large amounts of data and create plans within a greatly enhanced information architecture. Moreover, there is a corresponding proliferation of innovative visual detection, classification and tracking sensor technologies (f.e. hyperspectral imaging, computational photography and miniaturisation of sensors).

As the practices of AI evolve, problems arise concerning the ethical and normative standards of international humanitarian law. The potential use of AI technology in combat has raised a number of moral dilemmas. Foremost, how can we trust systems that are capable of identifying targets without a human being making a meaningful decision in the loop? AWS have the power to change our view of warfare fundamentally, particularly regarding who is responsible when a civilian is killed as a result of the failure of a programmed AI (Boutin 2022). Who should be held liable for a civilian’s death as a result of a technical malfunction of an AWS? Is it the creator of the algorithm, the owner of the AWS, or the commander who used it? While all technologies can malfunction, it seems like some algorithms are more at risk than others of misreading or misinterpreting a complex battlefield scenario, therefore causing injury or death to civilians. The most challenging aspect of these deadly systems being operated by algorithms is that it is impossible to know every possible situation that may arise in an armed conflict. Take, for example, if an AI system could eliminate a high-value terrorist target but at the same time result in civilian deaths, how would the AI system decide? Will it calculate the projected number of civilian deaths based on a threshold of five or ten? There is a difference in how humans and

AI function. While AI has some degree of independent learning or adaptation to new situations, it will never be able to assess a situation where life and death decisions must be made in a manner as humans, who are guided by reason, emotion, and principles of ethics. When a human makes a mistake, it is known who is to blame, while the issue of responsibility in AI systems is very vague. The aforementioned issues become magnified as AI becomes more integrated into decision-making related to the use of nuclear weapons. Who is at fault when decisions made by AI result in the mass loss of life? Is it the creator of the AI system, the military authorities approving its use, the nation state using the AI, or someone else? The use of AI in these instances raises serious issues about human dignity and the right to life, particularly relating to the question of autonomous systems operating without human involvement or oversight. As a result of these issues, the concept of “meaningful human control” has emerged as an increasingly important normative construct for future use and application of these technologies (Heyns 2017). Particularly with respect to using lethal force, should always include the meaningful human control. The development of clearly defined ethical guidelines and mechanisms of accountability are critical to ensuring the development of secure and responsible AI technologies for military applications and help to mitigate the risks.

AI has become increasingly relevant when addressing the challenge of tracking and managing the growing number of artificial objects in space, including satellites and other components of space-based infrastructure. Constellations of hundreds to thousands of small satellites that work together as an integrated, interrelated system will require advanced planning and real-time coordination to mitigate collision risks (Stojanović and Dimitrijević 2025, 139). Conventional methods of managing operations have been performed primarily by human operators and do not provide the capabilities when considering that the number of active satellites in low-Earth orbit is projected to triple over the next decade. In contrast, AI systems can coordinate the trajectories of thousands of objects simultaneously and to conduct collision avoidance manoeuvres, something that would not be possible for human operators to accomplish at the same pace. The strategic importance of AI-based autonomous systems has been demonstrated in the ongoing conflict in Ukraine. The Ukrainian battlefield has become a testing ground for military applications of artificial intelligence. One such test is the deployment of unmanned aerial vehicles (UAVs) with machine learning algorithms for target identification, navigation, and coordination of swarming behaviour. According to defence researchers, AI-targeting drones have improved strike precision and have likely reduced casualties for attacking forces through a decreased need for human involvement in high-risk situations (Schmidt and Grant 2025). In addition to the above concerns of accountability and establishing ethical standards, heavy reliance on any cyber-based system creates an immense opportunity for adversaries to exploit. AI-generated hacking tools that were developed to identify and exploit systemic failures can easily

compromise critical space-based infrastructure, with disastrous results. For example, a theoretical space-based weapon system programmed to autonomously detect and respond to perceived threats could severely compress decision timeframes to milliseconds. Such a system would provide a positive benefit in defensive reactions but would also place an unintentional risk of escalation through misinterpretation of enemy actions as imminent threats, or through enemy actions to spoof sensors. The frustrations offered by this issue create an unfortunate historical reflection. The Cuban Missile Crisis in 1962 was a historically tense conflict between the United States and the Soviet Union where both sides were engaging in complicated calculations and miscalculations about their responses to events. If either side had used a nuclear weapon against the other, this could have led to the deaths of millions of people or even cataclysmic global destruction. Would a hypothetical AI-enabled warning system during the Cuban Missile Crisis have led to a de-escalation in tensions by providing superior data analysis, or would it have increased tensions by misinterpretation of ambiguous signals as an imminent threat? Adding artificial intelligence to that type of scenario could create even greater problems and increase the possibility of worldwide destruction. This paradox illustrates a fundamental conflict when AI is integrated into strategic systems. The very attributes that promote operational efficiencies (speed, autonomy, and the ability to recognise patterns) inherently degrade the time for deliberation, contextual judgment and human restraint that have historically been used to prevent catastrophic miscalculations.

Integration of AI into Nuclear Decision-Making Systems and Mutual Assured Destruction

The roots of the nuclear decision-making system can be traced to the history of the Cold War and the geopolitical context of a world dominated by two nuclear superpowers. The relevance of nuclear weapons has evolved from a merely military relevance to a symbol of national pride. After the collapse of the Soviet Union, the post-Cold War world has undergone considerable changes in the geopolitical environment, characterised by extraordinary volatility and unpredictability. AI integration into nuclear decision-making would fundamentally transform the strategic decision-making architecture. But the potential benefits could rapidly become catastrophic risks to humanity. The rapid reaction time of AI systems, operating in milliseconds, could limit the opportunities for intervention before a response is “gone beyond” the reach of human decision-makers. For civil and military leaders, this means having a technical understanding of AI limitations (brittleness, lack of transparency, data vulnerability), as well as a normative imperative to develop institutions that facilitate judgment over the speed of algorithms. As AI becomes more prevalent in support of nuclear command and control and communications architectures, the responsibility is growing because of the irreversible consequences

of nuclear use, and the significant ethical imperatives to maintain human decision-making at the highest levels.

AI is also capable of amplifying cyber threats to the nuclear command and communications architecture, increasing the vulnerability of early warning systems. The use of advanced AI models may significantly exacerbate the potential for a cyberattack to cause inadvertent escalation (Tamburrini 2023). There is particular danger in using AI in drones and other autonomous platforms intended to deliver nuclear weapons. Research indicates that, despite all powerful countries such as the United States stating that they intend to maintain a “human in the loop” with respect to nuclear matters, the practical introduction of AI models drastically reduces human involvement (Галлямов 2026). There are various ways to introduce AI to nuclear decision-making, with three core models: “human-in-the-loop”, “human-on-the-loop”, and “human-out-of-the-loop”. The “human-on-the-loop” model is commonly pointed to as the lowest risk, retaining human oversight over AI systems, but time compression and operative speed of the AI still increase the risk of inadvertent escalation even in such a model. As it was already elaborated, the MAD (mutual assured destruction) concept, based on the assumption of rational action and caution regarding nuclear threats, defines strategic stability. The use of AI as part of the chain of command integrates the parameters of extremely fast response time, algorithmic opacity, and inherent biases. These variables have the potential to disrupt the MAD logic. AI not only accelerates decision-making processes, but also imposes a level of complexity for which the existing international order is not institutionally adapted.

At the time of the Cold War, nuclear deterrence was stable and reliable. Even with the proliferation of nuclear weapons, which was imperfectly channelled into specific dyadic relationships based on reciprocal threats, the strategy of nuclear deterrence was rather comprehensible, stable, and consistent with the underlying assumptions. According to Johnson, the introduction of offensive AI capabilities could potentially add additional variables to consider. Both state and non-state actors could potentially utilise AI in order to manipulate nuclear state through the implementation of “catalytic war”, when a third party’s (either a state or a non state entity) actions can lead to nuclear war (nuclear use) between two nuclear states as well as demonstrating the significant damage inflicted on both nations due to misinformation and misrepresentation of information by a third party (Johnson 2021). A catalytic war would use a third party to provoke conflict between nuclear adversaries, thereby causing escalation as long as that third party was not directly involved in the original escalation. This perspective highlights how AI not only may enhance existing geopolitical or strategic dynamics of nuclear deterrence, but how AI could also completely change some of the risk factors associated with nuclear deterrents.

Also, there is a critical issue related to “developmental asymmetries”. This means that rival actors equipped with nuclear weapons will, by definition, operate their

systems based on different value systems and logical structures. Therefore, in theory, their autonomous systems will not share mutual interoperability and cannot, figuratively, “understand” each other. This will increase the likelihood of escalation of conflict due to technical misalignment at the level of machines communicating with one another. This transition from relatively stable relations among nuclear powers to a system of unpredictable, foggy, and accelerating decision-making based on AI will mean that there is no longer the necessary time to carefully consider strategy before potentially using a nuclear weapon. State-of-the-art AI models, which are based primarily on statistical pattern recognition, are highly sensitive to events that fall outside the training set. So, they can produce catastrophic errors in the event of a nuclear crisis. Because a nuclear crisis by its definition, is a one-time event that cannot be replicated, the statistically fragile models of AI will produce unpredictable results that cannot be detected until they occur and cannot be corrected in time to avert the situation. The incident of Stanislav Petrov in 1983 is an excellent example of the potential threat represented by a fully automated, AI-driven response system. Because Petrov’s decision was based solely on his human judgment, there is a high probability that a fully automated AI would have led to the marginalisation of human judgment and the initiation of a nuclear war (Aksenov 2013). The introduction of AI into nuclear command, control, and communications (NC3) systems is a Faustian bargain that could change the concept of deterrence from mutual assured destruction (MAD) to mutual assured vulnerability (MAV) (Sakubu, 2025, pp. 1003–1004). The epistemological opacity of algorithms, known as the “black box” problem, combined with the dramatic compression of decision-making timeframes, increases the potential for nuclear escalation. There are regional specificities, such as those in South Asia with minimal missile flight time, that further enhance incentives for preventive strikes by eroding the stability afforded by a second-strike capacity (Kazmi and Kazmi, 2025, pp. 20–21). In the case of complete integration of AI systems into the nuclear decision-making process, a key component of national security would become cybersecurity as cyber attacks on control systems could lead to catastrophic consequences such as the use of nuclear weapons. Recent research conducted by King’s College London has revealed that simulated war games utilising AI models resulted in conflict escalation through the use of nuclear weapons in 95% of the scenarios (Payne 2026). The research was led by Professor Kenneth Payne and focused on how large language models process simulated crises involving nuclear weapons. There were a total of 21 simulated crises, and in all 21 cases at least one side was found to be signalling their willingness to use nuclear weapons, and 95% of those incidents were mutual signals. “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

centres, 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” (Payne 2026, 21).

Conclusion: Does Automation Increase the Risk of Nuclear War?

Future directions of AI development and its implications for nuclear stability will be the subject of numerous scientific research. As AI matures, it is expected to significantly alter the nature of war and the way how weapons are operated. There is an immense amount of uncertainty about how AI will evolve in the nuclear domain. Currently, we have an understanding of mutual assured destruction (MAD) that is based upon human judgement and allows sufficient time to make decisions, but as AI matures, we are expected to move to a new understanding of mutual assured vulnerability (MAV), characterised by an unstable and uncertain environment. Much of this uncertainty is due to the reduced timeframe in which decisions will have to be made. Therefore, AI-driven early warning systems evaluate potential threats so quickly that the ability for humans to make decisions has been severely compromised. Free will and responsibility, concepts that belong to human beings but do not apply to automated systems like AI, were present on several occasions where decision-makers did not follow standard procedures, but instead chose to use emotional reasoning in an attempt to avoid nuclear war. Although emotions cannot replace rational thinking and, by their very nature, constitute a cognitive distortion, in certain situations they can play a positive role as a corrective factor for rationality. The logical processes currently embedded into machines however, could ultimately lead to devastating conflict with humanity’s value system, including the matters of utmost importance such as the sanctity of life, freedom of choice and individual responsibility. Cold logic is indifferent to human life, and could potentially determine the use of nuclear weapons to be a good course of action. As we have demonstrated in this paper, the very concept of rationality is profoundly deficient and exists not as an objective constant but as a constructed rationality. Regardless of this epistemological limitation in the field of nuclear deterrence, constructed rationality has historically prevented the outbreak of nuclear war. The final conclusion that emerges from the preceding analysis is that the integration of AI into nuclear command, control, and communications systems increases the overall risk of nuclear war.

ACKNOWLEDGEMENT: The paper presents findings of a study developed as a part of the research project “Serbia and Challenges in International Relations in 2026”, financed by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, and conducted by Institute of International Politics and Economics, Belgrade, during the year 2026.

ВЕШТАЧКА ИНТЕЛИГЕНЦИЈА И НУКЛЕАРНО ОДВРАЋАЊЕ: ИЗАЗОВИ ИНТЕГРАЦИЈЕ ВИ У ПРОЦЕС ОДЛУЧИВАЊА О НУКЛЕАРНИМ ПИТАЊИМА

Богдан СТОЈАНОВИЋ¹

АПСТРАКТ

Овај рад бави се истраживањем односа између најновијих технологија вештачке интелигенције (ВИ) и нуклеарног одвраћања, односно теоријског и практичног концепта који се јавља у међународној политици од Хладног рата до данас. Кључно истраживачко питање који ова анализа истражује јесте да ли би интеграција ВИ у процес нуклеарног одвраћања допринела стабилности или би, насупрот томе, смањење присуства људског фактора у одлучивању о нуклеарним питањима повећало ризик од употребе нуклеарног оружја. Анализа цене или трошкова са једне и користи од вештачке интелигенције са друге стране у осетљивој области као што је нуклеарно одлучивање, најефикасније се испитује кроз реалистичку призму. Исто тако, аутор у теоријски оквир уводи сопствени концепт „конструисане рационалности“. Питање одговорности у случају потискивања људског фактора у корист ВИ отвара нове изазове и усложњава процес нуклеарног одвраћања. Претпоставка о нуклеарном миру показала се важећом у пракси, јер међусобно загарантовано уништење (MAD) чини рат између нуклеарних сила мало вероватним. ВИ је већ пронашла своје место у војном домену, што оправдава испитивање начина на који би примена најсавременије технологије утицала на функционалност и предвидивост нуклеарног одвраћања међу државама које поседују нуклеарно наоружање. Аутор закључује да би потпуна интеграција система ВИ у нуклеарно одлучивање несумњиво повећала вероватноћу намерне или ненамерне нуклеарне ескалације.

КЉУЧНЕ РЕЧИ

ВИ; вештачка интелигенција; нуклеарно одвраћање; нуклеарно одлучивање (одлучивање о нуклеарним питањима); нуклеарно оружје (наоружање); конструисана рационалност; нуклеарни рат.

Цитирајте овај чланак као: Stojanović, Bogdan. 2026. "Artificial Intelligence and Nuclear Deterrence: Challenges of AI Integration into the Nuclear Decision-Making Process". *The Review of International Affairs* LXXVIII (1198): 567–601, https://doi.org/10.18485/iipe_ria.2026.77.1198.2

¹ Виши научни сарадник, Институт за међународну политику и привреду, Београд, Србија. Е-пошта: bogdan.stojanovic@diplomacy.bg.ac.rs, <https://orcid.org/0000-0002-9624-1516>

Уводна разматрања

Различите претње које произилазе из употребе вештачке интелигенције (ВИ) у контексту нуклеарног оружја и пратећих система одлучивања, неки су од највећих изазова за савремену међународну безбедност. Иако постоје аргументи да ће технолошки развој војних потенцијала увек наставити унапред одређеним смером, увођење ВИ у домен нуклеарног оружја доноси јединствене изазове који не могу да се нађу у другим врстама војне примене. Имајући у виду степен уништења који је човечанству доступан употребом нуклеарног оружја на глобалном нивоу, увођење ВИ ће захтевати другачије начине анализе у погледу брзине и аутономије одлучивања. Иако се и о нуклеарној и о ВИ технологији обично размишља у смислу повећања безбедности, постоји и контра аргумент, а тај је да ће ВИ у фундаменталном смислу пореметити логику одвраћања која је током више од осам деценија одржавала стабилан међународни систем. У одбрамбеним круговима, ВИ се нашироко посматра као мултипликатор силе, који ће драматично променити архитектуру нуклеарног командовања, контроле и комуникација (НКЗ). Имајући у виду да се ВИ интегрише у системе раног упозоравања, као и у националне системе за извиђање, надзор и обавештајни рад (ISR), биће потребно да се изврши процена утицаја који ће ВИ имати на постојеће стратегије одвраћања и доктрину стратешке стабилности. Развој и распоређивање ВИ у овом домену означава квалитативну промену у развоју нуклеарног одвраћања и стратегија за очување међународне безбедности у ширем смислу. Иако ВИ обећава повећану ефикасност, брзу обраду огромних количина нових података и појачану способност анализирања сложених информација, она доноси и нове и незапамћене ризике по глобалну стабилност. Вештачка интелигенција није у суштини технологија која доноси дестабилизацију у нуклеарном контексту. Напротив, ризици и користи од ВИ зависе од њене примене и од степена у којем се „одговарајући нивои људске процене“ одржавају у кључним одлукама (Lin 2025).

Кључни ризици укључују степен јасноће алгоритама, могућност манипулације подацима, ерозију аутентичне људске контроле упркос номиналном очувању принципа „човек у петљи“ (*man in the loop*) односно „човек у процесу одлучивања“, различите логике различитих нуклеарних сила, скраћивање времена за доношење одлука и јачање ризика од дезинформација и непоштовања споразума. Самит о одговорној ВИ у војном домену (REAIM), који је одржан у Шпанији у фебруару 2026. године, илустровао је дубоко укорењено стратешко подозрење земаља које су највише напредовале у области ове примене. Сједињене Америчке Државе (САД) и Кина одустале су од потписивања заједничке декларације о одговорној употреби ВИ у војном домену, зато што би било какво успостављање нормативних оквира довело до

класичне „дилеме затвореника“ (односно, да ли прекршити споразум и издати партнера или сарађивати) и губитка већ постигнуте предности (Waldersee 2026). Темељни проблем будућег развоја оличен је у питањима како сачувати већинску контролу човека над најразорнијом силом на свету и како надгледати и верификовати нешто што је по својој природи суштински непрозирно?

Водећи пионири у машинском учењу упозорили су да ВИ има потенцијал да развије аутономију изван људског надзора која се већ појављује, као и да произведе неистините или неусклађене резултате. Џефри Хинтон (Geoffrey Hinton) је истакао да високо напредна ВИ може да развије инструменталне сврхе које се разликују од онога што ми желимо да она ради. Стога ће постати изузетно тешко имати поуздану контролу над одлукама које носе значајне последице (Brown 2023). Да би се избегло брзо увођење и примена система ВИ у архитектура нуклеарног командовања, контроле и комуникација без адекватног регулаторног механизма, међународна заједница мора да сарађује на успостављању заједничких закона за заштиту од непријатељске манипулације овим системима. Без кооперативне и свеобухватне стратегије, међународна заједница ће вероватно доћи до парадокса, односно да је, тражећи технолошку доминацију, истовремено створила окружење повећаног нуклеарног ризика.

Егзистенцијално нуклеарно одвраћање и конструисана рационалност у међународним односима

При анализи последица неке нове технологије, неопходно је да се најпре изврши процена како та технологија може да утиче на шири концептуални оквир. Нуклеарно одвраћање је првобитно развијено као средство за националну безбедност током Хладног рата, али теоријска основа одвраћања има корене чак и у самој људској природи. На свом најосновнијем нивоу, одвраћање је превентивна употреба претње силом како би се утицало на другу државу или недржавног актера да се понаша на одређени начин ради постизања стратешких циљева (Downs 1989, 226). У случају нуклеарног одвраћања, једна држава покушава да одврати другу од напада тако што озбиљно разматра употребу нуклеарног оружја против агресора. Овај модел одвраћања зависи од претпоставке о рационалности. Цена у случају да се постане мета нуклеарног удара вероватно ће бити већи од потенцијалне користи (George and Smoke 1974, 11). Нуклеарно одвраћање засновано је на реалистичком погледу на природу међународних односа. Реалисти теже да посматрају главне актере у међународним односима као себичне, јединствене актере који настоје да постигну различите циљеве. Безбедност, или опстанак, представља врховни интерес држава и обично главни циљ спољне политике. Чувени хладноратовски стратег Бернард Броди (Bernard Brodie) проицијиво

запажа да је „до сада главна сврха наше војне структуре била да побеђује у ратовима. Од сада њена главна сврха мора да буде да их спречава“ (Brodie 1946, 76). Грубо говорећи, све жеље за одласком у рат настављају се у опадајућем редоследу почев од жеље за победом у бици, а све жеље за избегавањем рата следе у растућем редоследу, почев од жеље да се избегне сваки ризик по сопствену безбедност и тако у потпуности спречи рат. Парадоксална природа нуклеарног оружја лежи у његовој истовременој способности да оконча људско постојање и да одврати нуклеарне силе од упуштања у директан рат. Ово произилази из споја рационалне тежње за опстанком и катаклизмичних последица потенцијалне нуклеарне размене (Узајамно загарантовано уништење, или MAD – Mutual Assured Destruction). Ово рационално одвраћање од агресије, утемељено на императиву опстанка државе, додатно је појачано наизглед ирационалном категоријом људског страха. Поједини теоретичари ову конфигурацију карактеришу као „равнотежу страха“, у којој су рационалне калкулације цене и користи допуњене страхом политичких елита од потпуног уништења (Wohlstetter 1959). Стога би приликом анализе односа међу нуклеарно наоружаним државама било потпуно прикладно говорити о „одвраћању путем застрашивања“, а не искључиво о рационалном одвраћању.

На теоријском нивоу, могу да се издвоје два фундаментална приступа одвраћању, иако модалитети и ефекти одвраћања остају доследни у оба случаја. Први приступ је у суштини рационалистички у ужем, економском смислу, будући да наглашава анализу цене и користи унутар самог одвраћања. Према овом моделу, одређена држава може да се уздржи од агресије због очекивања неприхватљиво високе цене која би неизбежно уследила. Алтернативни приступ је фундаментално психолошки и даје приоритет концепту „страха“ од уништења као примарном мотиву за уздржавање од напада на државу која поседује нуклеарно наоружање. Унутар овог оквира, опстанак се разуме као урођени инстинкт сваког појединца, а самим тим и сваке државе. Упркос њиховим одговарајућим методама анализе, економски и психолошки приступи одвраћању могу да се посматрају као две стране истог аналитичког новчића. Дакле, земље или неће водити нуклеарни рат зато што су рационално израчунале цену односно трошкове и користи, или ће се уздржати од изазивања сукоба зато што имају емоционални страх од уништења. Због тога би, услед ширине овог концептуалног оквира, термину „егзистенцијално нуклеарно одвраћање“ требало дати предност у односу на „рационално нуклеарно одвраћање“, иако овај други доминира у научној литератури (Bundy 1983, 4). Постоји узајамни однос између одлука заснованих на емоцијама и оних заснованих на разуму. Свака одлука да се не уђе у нуклеарну размену има рационалне последице, јер нуклеарни рат никада не може да буде користан или исплатив. Као што је Роберт Џервис (Robert Jervis)

објаснио, успешно одвраћање не захтева да политички лидери буду савршено рационални (Jervis 1979, 299). Оно једноставно захтева да доносиоци одлука имају довољно свести о стварности да ће цена односно трошкови упуштања у нуклеарни рат далеко превазићи сваку могућу корист. Ова основна идеја превазилази једноставан логички концепт и састављена је од три кључна елемента: стратешког резонувања, психолошких карактеристика и стварности да постоји егзистенцијална претња људском постојању коју је створило нуклеарно оружје.

Кенет Волц (Kenneth Waltz) је дошао до сличних закључака када је анализирао у којој мери би стратегија одвраћања могла да функционише у пракси. Према њему, рационални доносиоци одлука не морају да раде опсежне анализе цене и користи како би разумели нуклеарни рат. Они само треба да препознају да постоје изузетно високи „трошкови“ односно цена повезана са упуштањем у такав рат. Волц ово назива „претпоставком рационалности“, која представља минимални услов да би одвраћање функционисало (Waltz 1981). Сви облици вођства, било од стране рационалних или ирационалних лидера (чак и Хитлер као можда најчешће цитирани пример ирационалног лидера) требало би да буду у стању да се уздрже од напада на земљу која би могла да употреби нуклеарно оружје. Посматрајући ово питање кроз онтолошку перспективу, рационалност концептуално постоји у изузетно нестабилном окружењу да би могла да функционише и буде оперативно ефикасна.

Одређени објективни догађај мора да буде когнитивно обрађен како би се одредила акција коју треба предузети. Људско понашање, као и глобално понашање свих нација заједно, произилази из субјективне конструкције до које се долази кроз њихове идеје, норме, вредности и психолошке карактеристике (Stojanović 2013, 150). Када га надоградимо на концептуални оквир који је развијен у претходним истраживањима, израз „конструисана рационалност“ најбоље описује практични ефекат рационалности, будући да сваки феномен објективне стварности мора да буде и посредован кроз људски филтер састављен од нормативних оквира, система вредности и ирационалних димензија пре него што се донесе одлука о конкретној акцији (Stojanović 2013, 151). Закључивање је нешто што нам се једноставно дешава, јер спознаја непрестано иде изван доступних података (Moshman 2004). На другом нивоу, размишљање је намерна употреба наших закључака за постизање циља, и на крају, резонување се јавља тек када то размишљање постане самоограничено нормама које мислилац сматра оправданим (Moshman 2004). Резонување увек зависи од личних квалитета онога ко резонује.

Чак и уз огромна ограничења у погледу стратегије одвраћања из теоријске и практичне перспективе, нуклеарно одвраћање је до сада функционисало и без икакве сумње ће наставити да функционише. Теорија игара представља динамичан и општи оквир за анализу утицаја рационалног понашања,

израчунавање ризика и погрешне комуникације у свим стратешким интеракцијама између држава (Lindelauf 2020). Схватање суштине конструисане рационалности и оперативне логике стратегије одвраћања, од кључног је значаја за критичку анализу ефеката интеграције ВИ у нуклеарну стратешку „игру“. Интеграција ВИ у процес стратешког одлучивања у државама са нуклеарним наоружањем, ефективно лишава међудржавне односе њихових субјективних, нормативних и афективних димензија, трансформишући их у идеално-типичан модел рационалног понашања. Парадоксално, овакав приступ међудржавној интеракцији повећава структурне ризике од нуклеарног сукоба управо зато што елиминише прагматична оклевања, етичке инхибиције и контекстуалне процене које су у прошлости функционисале као механизам против рата.

Нуклеарно одвраћање јесте стратегија коришћења војне силе за постизање политичких циљева на индиректан начин. У нуклеарној ери, главна стратешка мисао померила се са генералштабова на цивилне кабинете. Током Хладног рата, стратегија и политика постале су неизбежно испреплетене, због чега је било тешко повући линију између њих. С обзиром на капацитет нуклеарног оружја за тотално уништење, стратешко размишљање се померило из војних у цивилне кабинете. Због тога се овај период често назива „златним добом стратешких студија“, јер су политички односи у то време били у темељу обликовани нуклеарним стратегијама суперсила. Ова транзиција је од изузетног значаја, јер је нуклеарно одвраћање све више постајало инструмент којим су управљали претежно политички лидери, а не војни команданти. ВИ може да нас доведе до новог заокрета, односно до окретања од цивилних кабинета ка процесорима. Џејмс Џонсон (James Johnson) тврди да „у контексту ВИ и аутономије, посебно у смислу сложености информација, дезинформација и манипулације, логика одвраћања заснована на рационалности делује као све неодрживија претпоставка“ (Johnson 2023, 202–203). Ова технологија је нуклеарном одвраћању додала још један ниво сложености, као вероватно најделикатнијем домену доношења одлука.

Концепт и примена вештачке интелигенције

Чувена књига „Сингуларност је близу: кад људи превазиђу биологију“, коју је написао реномирани програмер Реј Курцвајл (Ray Kurzweil 2005), предвидела је брз развој технологије до тачке у којој ће она надмашити људску интелигенцију. Ова такозвана „технолошка сингуларност“ била би тачка у историји када људска интелигенција више неће бити примарни облик интелигенције. Вештачка интелигенција је рачунарски систем који покушава да имитира људску интелигенцију и да јој парира. Циљ развоја ВИ јесте стварање рачунарског модела који учи, усавршава се и решава све сложеније

проблеме. Као што је претходно разматрано, дубоко учење је управо такав покушај да се створи систем који учи попут људског мозга и стога постаје прогресивно „паметнији“. На пример, ИБМ-ов (IBM) Дип Блу (Deer Blue), рачунар за играње шаха, победио је светског шампиона у шаху Гарија Каспарова (Garry Kasparov) 1997. године (Ћао 2022). Програми за играње шаха су од тада прешли дуг пут, дошавши до тачке у којој људи више не представљају конкуренцију. Али шах је само једна област у којој уска вештачка интелигенција (Artificial Narrow Intelligence – ANI) потискује људске професионалце са њихових радних места. Људи бивају потиснути из анализе података, медицинске дијагностике и безброј других специјализованих позиција. Рад Марвина Минског (Marvin Minsky 1954) на неуронским мрежама и вештачкој интелигенцији у симулирању функција мозга, чини основу већине каснијих истраживања ВИ, укључујући и тежњу ка општој вештачкој интелигенцији (Artificial General Intelligence – AGI), која представља свети грал резоновања на његовом највишем нивоу. Постоји значајна забринутост, па чак и страх, пошто се следећа генерација интелигентних машина креира веома брзим темпом. Велики део страха који се изражава у медијима у вези са ВИ односи се на потискивање великог броја људи у различитим сферама.

Двострука природа ВИ (као и код сваке технологије) огледа се у томе што она може да се користи у позитивном смислу (нпр. откривање тумора пре него што се могу уочити људским видом, помоћ у решавању математичких проблема на начине који су другачији од оних на који их решава човек, као и убрзавање процеса инжењерског пројектовања) или у негативном смислу, стварајући нехумано друштво. Иако корени ВИ и експериментални напори у овој области сежу неколико деценија уназад, одлучујући продор у овој технологији остварен је тек последњих година. Пажње је вредна коинциденција да је рат у Украјини, који је темељно изменио пејзаж глобалне политике, почео у фебруару 2022. године, док је касније те исте године компанија ОпенАИ (OpenAI) лансирала ЧетГПТ (ChatGPT), вероватно најпознатији велики језички модел. Према речима Сема Алтмана (Sam Altman 2022), извршног директора компаније, ова ВИ апликација прикупила је својих првих милион корисника у року од само пет дана од дана њеног објављивања. Сукоб у источној Европи поново је потиснуо питање распоређивања нуклеарног оружја на сам врх политичке и војне агенде. Процене које су сугерисале да би Русија могла да прибегне тактичком нуклеарном оружју, без обзира на малу вероватноћу, привукле су пажњу глобалне јавности. Тако је ова временска коинциденција подстакла критичко преиспитивање симбиозе два различита наратива. С једне стране, перспектива покретања нуклеарног армагедона, а са друге, способност ВИ да резонује ефикасније од људске когниције. Од највећег је значаја имати на уму да тренутно широм света постоји

преко 12.000 нуклеарних бојевих глава, што је реалност која претњу уништења људске цивилизације чини дубоко веродостојном (Davenport 2026).

Брзи раст ВИ на потрошачком тржишту роба и услуга не представља тек значајно научно достигнуће, већ отвара врата и за велику примену ВИ у ратне сврхе. Као што нам историја говори, технологије двоструке намене увек мигрирају кроз цивилно-војне домене, од комерцијалних ка војним применама и обрнуто. Историјски примери укључују микрочипове, сателите, интернет, ракетну технику и нуклеарну енергију. Људи и даље остају најуспешнија бића због своје способности да уче, размишљају и усвајају новостворене идеје и вештине, да их чувају и преносе кроз време. Међутим, у изградњи окружења које би задовољило њихове сопствене потребе, људи су успели да створе нешто што би на крају могло да их замени, односно ентитет који се често назива „последњим изумом“. Увођење опште вештачке интелигенције (AGI) било би једнако важно као и овладавање ватром или откриће електричне енергије. Остаје да се види да ли ћемо и када увести АГИ која ће ефикасније од људи обављати све когнитивне функције. Међутим, хитност примене тренутних технологија ВИ у војном развоју се повећала, посебно када је реч о оружју за масовно уништење.

Сличности и разлике између ВИ и нуклеарног оружја

Оба ова изума, нуклеарно оружје и вештачка интелигенција, за људску цивилизацију истовремено представљају технолошки скок напред и потенцијални инструмент за масовно уништење (било физички или на други начин). Током Хладног рата, научници, креатори политике и војни стратеги хватали су се у коштац са фундаменталним питањем у вези са нуклеарним оружјем: да ли ће оно на крају уништити планету и њене становнике, или ће равнотежа нуклеарне моћи спречити рат великих размера? Хипотеза о „нуклеарном миру“ показала се углавном тачном, пошто је катаклизмична природа овог оружја ефикасно одвратила директан војни сукоб између водећих светских нуклеарних сила (Сједињених Држава и Совјетског Савеза) (Stojanović 2021). Са истим питањем се суочавамо и данас у вези са ВИ: да ли ће она на крају да заштити или уништи човечанство? Баш као што су водећи научници на почетку нуклеарне ере позвали на строгу контролу даљег развоја нуклеарног оружја, истакнуте личности у области ВИ потписале су отворену изјаву у којој тврде да би „ублажавање ризика од изумирања услед вештачке интелигенције требало да буде глобални приоритет заједно са другим ризицима друштвених размера као што су пандемије и нуклеарни рат“ (Abdul 2023). Илон Маск (Elon Musk), најбогатији човек на свету и значајан инвеститор у нове технологије, отишао је корак даље, тврдећи да ВИ представља далеко већу егзистенцијалну претњу од нуклеарног оружја (Clifford 2018).

Развој како ВИ, тако и нуклеарног оружја, захтева значајна улагања у физичку инфраструктуру. Осим неопходне научне стручности, изградња нуклеарног уређаја захтева опсежна и скупа постројења за производњу фисионог материјала. Поступци као што су обогаћивање уранијума и прерада плутонијума, остају у домену само технолошки најнапреднијих држава. Вештачка интелигенција такође захтева значајне физичке ресурсе, што најбоље илуструју најсофистициранији микрочипови, за које су капацитети и способност производње за сада концентрисани у рукама само неколико држава. Другим речима, обе технологије носе потенцијал да представљају егзистенцијалне претње човечанству у целини, јер су обе погодне за цивилне и војне примене. На крају крајева, развој ВИ и нуклеарне технологије одвијао се и наставља да се одвија у контексту „међународне трке“, с тим што сада та трка укључује и приватне корпорације. Упркос овим суштинским сличностима у погледу зависности од ресурса, технолошки револуционарног потенцијала и ризика по човечанство, технологије ВИ и нуклеарног оружја у знатној мери се разликују у бројним димензијама. Док је ширење нуклеарног оружја ограничено оскудицом фисионих материјала и видљивошћу инфраструктуре за обогаћивање, ширење способности ВИ суочава се са мање материјалних уских грла и може да се шири путем софтверских канала које је по природи теже надгледати и верификовати. Нематеријална природа алгоритама ВИ чини њихово ширење на шири круг актера знатно лакшим у поређењу са гломазном физичком инфраструктуром за нуклеарне програме (реактори, фисиони материјали или центрифуге). Заштитне мере могу релативно лако да се заобиђу изменама изворног кода, чиме се ефективно онемогућава смислена контрола над ширењем након што је модел ВИ првобитно објављен. Ова кључна асиметрија између материјалних ограничења која управљају нуклеарном пролиферацијом и бескомпромисног ширења заснованог на софтверу, доноси дубоке последице по међународно управљање, контролу наоружања и осмишљавање регулативе у двадесет првом веку. Штавише, док су деструктивне последице нуклеарног оружја тренутне, физичке и географски ограничене, потенцијалне штете напредних система ВИ могу да буду индиректне, системске и глобално распрострањене у распону од алгоритамске пристрасности и сајбер рањивости до ерозије људског утицаја у кључним процесима доношења одлука. Вештачка интелигенција је потекла првенствено из приватног сектора, а њен развој карактерише разуђени екосистем који обухвата бројне недржавне актере, што је структурална карактеристика која у знатној мери усложњава све будуће напоре за контролу њеног ширења. Док нуклеарно оружје остаје ограничено на безбедне арсенале најмоћнијих светских држава, ВИ користе милиони појединаца широм света, при чему свака интеракција постепено доприноси итеративном усавршавању и „интелигенцији“ ових система. Нуклеарна технологија омогућава релативно

јасно разграничење између цивилне и војне примене. Према међународном праву, цивилни нуклеарни програми законски су дозвољени, што објашњава зашто више од тридесет држава тренутно управља нуклеарним реакторима у мирољубиве сврхе. Истовремено, Уговор о неширењу нуклеарног оружја (NPT 1968) експлицитно забрањује ширење нуклеарног оружја као војне компоненте ове технологије. Насупрот томе, различите примене ВИ не могу јасно да се раздвоје, што засебну регулативу или забрану специфичних употреба чини изузетно изазовном. На пример, размислите о томе да иста основна способност која омогућава брзо састављање корпоративне електронске поште може подједнако да олакша брзо ширење дезинформација у великим размерама.

Уместо покушаја да се развој ВИ ограничи кроз регулативе засноване на самој технологији, бољи приступ би могао да буде фокусирање регулатива на начин на који се ВИ користи. Овде могу да нам буду од помоћи историјски примери. На пример, Споразум о спречавању инцидената на отвореном мору и у ваздушном простору изнад отвореног мора између САД и СССР-а (APIHS 1972) није ограничио развој војних бродова, већ је креирао оперативне протоколе за регулисање понашања којима се захтевало обавештавање о опасним маневрима. Слично томе, Споразум о свемиру (1967) нуди нормативни оквир према којем се државе сматрају одговорнима за акције својих приватних субјеката (Stojanović and Dimitrijević 2025, 312–323). Ово може да подстакне државе да успоставе сличне регулативе за корпорације које развијају и примењују апликације ВИ. Како би се и теоријски и практично подржало нуклеарно одвраћање, од суштинског је значаја схватање односа између ВИ и нуклеарног оружја. Увођење ВИ у ову област створило би ниво готово неповратне посвећености. Коришћење једне високоризичне технологије као што је ВИ за контролу акција друге (нуклеарно оружје), драматично повећава потенцијал за катастрофалне резултате по човечанство ако ствари крену у погрешном смеру.

Одговорност и друга отворена питања у војној употреби ВИ

Војне примене ВИ имају огроман потенцијал да унапреде тренутне капацитете у осматрању, командовању и контроли, анализи података и даљинском управљању платформама са посадом или без посаде (Horowitz 2018). Иако је још увек у повоју, можемо да очекујемо да ће у наредних неколико деценија технологије следеће генерације бити развијене и распоређене невиђеном брзином. Притисак да се системи омогућени вештачком интелигенцијом распореде на тренутно оперативне платформе, захтеваће радикалну трансформацију постојећих платформи и натераће војне стратегије да преиспитају свој конвенционални стратешки начин размишљања.

Технологије следеће генерације имају потенцијал за револуционарне војне иновације, прелазећи са модела ратовања из индустријског доба, преко информационог ратовања, до система треће генерације који укључују потпуно аутономне системе оружја (Autonomous Weapon Systems – AWS).

Увођење алгоритама ВИ у војни арсенал отворило је ново поглавље у пракси коришћења ратне технологије. На утицај употребе ВИ у оружаним снагама све се више реферише као на ВИ револуцију у војним пословима (AI revolution in military affairs – AI-RMA) (Raska 2021). Из стратешке перспективе, увођење ВИ у војни домен приморава државе да преиспитају своје стратегије ратовања и да прилагоде своје актуелне доктрине како би се боље суочиле са стратешким могућностима и претњама које представља технологија ВИ.

Контрола ВИ представља још један изазов у том смислу, зато што она не може да се перципира као оружје. Корисна аналогија за ово могла би да буде радарска технологија. Модерне војне примене показују да без њене способности за уочавање циљева, ракете у реалности остају „слепе“. Слично томе, ВИ служи као појачивач војне силе тако што може да обради огромне скупове података и обезбеди синхронизацију војних јединица у реалном времену. Војна стратегија са интегрисаном ВИ користи тоталитет свих система наоружања који делују у ваздуху, на копну, на мору, у свемиру и у сајбер простору, како би се испунили стратешки циљеви. Употреба технологије ВИ од стране армија широм света расте, како би се унапредила борбена ефикасност и подржао низ војних операција у вишеструким доменима. ВИ подржава ситуациону свесност, што омогућава особљу да боље искористи расположива средства и доноси информисане одлуке (обавештајни рад, надзор и извиђање – Intelligence, surveillance and reconnaissance – ISR). Војни команданти и доносиоци одлука могу брзо да обраде велике количине обавештајно-извиђачких података, односно ISR и да израде планове унутар знатно унапређене информационе архитектуре. Штавише, већ постоји одговарајуће ширење иновативних сензорских технологија за визуелну детекцију, класификацију и праћење (нпр. Хипер-спектрално снимање, компјутерска фотографија и минијатуризација сензора).

Са еволуцијом праксе коришћења ВИ, јављају се проблеми у вези са етичким и нормативним стандардима у области међународног хуманитарног права. Потенцијална употреба технологије ВИ у борби, покренула је низ моралних дилема. Пре свега, како можемо да верујемо системима који су у стању да идентификују циљеве без људског бића које доноси смислену одлуку? Аутономни системи оружја (AWS) имају моћ да из темеља промене наш поглед на ратовање, посебно у погледу тога ко је одговоран када цивил буде убијен као резултат отказивања програмиране ВИ (Boutin 2022). Ко би требало да сноси одговорност за смрт цивила услед техничког квара аутономног система оружја? Да ли је то творац алгорита, власник система или командант који га је

употребио? Иако свака технологија може да откаже, чини се да су неки алгоритми под већим ризиком од других када је у питању погрешно читање или тумачење сложеног сценарија на бојном пољу, чиме се узрокују повреде или смрт цивила. Најизазовнији аспект код ових смртоносних система којима управљају алгоритми, јесте то што је немогуће предвидети сваку могућу ситуацију која може да се јави у оружаном сукобу. На пример, ако би систем ВИ могао да елиминира терористичку мету високог профила, али би то истовремено резултирало смрћу цивила, коју одлуку би донео систем ВИ? Да ли ће израчунати пројектовани број смртних случајева цивила на основу прага од пет или десет? Постоји разлика у начину на који функционишу људи и на који функционише ВИ. Иако ВИ има одређени степен независног учења или прилагођавања новим ситуацијама, она никада неће бити способна да процени ситуацију у којој се доносе одлуке о животу и смрти на начин на који то чине људи, који су вођени разумом, емоцијама и етичким принципима. Када човек направи грешку, зна се ко је крив, док је питање одговорности код система ВИ прилично магловито. Поменути проблеми постају још већи како се ВИ све више интегрише у доношење одлука у вези са употребом нуклеарног оружја. Ко је крив када одлуке које донесе ВИ резултирају масовним губитком живота? Да ли је то творац система ВИ, војне власти које одобравају његову употребу, национална држава која користи ВИ или неко трећи? Употреба ВИ у овим случајевима покреће озбиљна питања о људском достојанству и праву на живот, посебно у вези са питањем аутономних система који делују без људског учешћа или надзора. Као резултат ових проблема, концепт „смилене људске контроле“ јавља се као све важнији нормативни конструкт за будућу употребу и примену ових технологија (Heyns 2017), а посебно када је у питању употреба смртоносне силе, јер би она увек требало да укључује смислену људску контролу. Развој јасно дефинисаних етичких смерница и механизма одговорности, од кључног је значаја за осигурање развоја безбедних и одговорних технологија ВИ за војне примене, као и за помоћ у ублажавању ризика.

ВИ све више добија на значају у решавању изазова праћења и управљања све већим бројем вештачких објеката у свемиру, укључујући сателите и друге компоненте свемирске инфраструктуре. Констелације од стотина до хиљада малих сателита који раде заједно као интегрисан, међусобно повезан систем, захтеваће напредно планирање и координацију у реалном времену како би се ублажили ризици од судара (Stojanović and Dimitrijević 2025, 139). Конвенционалне методе управљања операцијама примарно су обављали људски оператери и оне не пружају неопходне могућности када се узме у обзир да се предвиђа да ће се број активних сателита у ниској Земљиној орбити утростручити током наредне деценије. Насупрот томе, системи ВИ имају способност да врше координацију путање хиљада објеката истовремено и да спроводе маневре за избегавање судара, што би за људске оператере било

немогуће да постигну истом брзином. Стратешки значај аутономних система заснованих на ВИ демонстриран је у текућем сукобу у Украјини. Украјинско бојно поље постало је полигон за тестирање војних примена вештачке интелигенције. Један такав тест представљало је распоређивање беспилотних летелица (БПЛ) са алгоритмима машинског учења за идентификацију циљева, навигацију и координацију понашања у ројевима. Према тврдњама истраживача у области одбране, дронави са ВИ која је коришћена за циљање, побољшали су прецизност удара и вероватно смањили губитке за нападачке снаге кроз мању потребу за ангажовањем људи у високоризичним ситуацијама (Schmidt and Grant 2025). Поред горе наведених питања у погледу одговорности и успостављања етичких стандарда, превелико ослањање на било који систем заснован на сајбер технологији отвара огромну прилику коју противници могу да искористе. Хакерски алати генерисани помоћу ВИ, који су развијени за идентификацију и искоришћавање системских пропуста, могу лако да угрозе критичну свемирску инфраструктуру са катастрофалним резултатима. На пример, теоријски свемирски систем наоружања, програмиран да аутономно детектује и реагује на уочене претње, могао би озбиљно да сажме временске оквири за доношење одлука на милсекунде. Такав систем би пружио позитивну корист у одбрамбеним реакцијама, али би такође донео ненамерни ризик од ескалације кроз погрешно тумачење непријатељских акција као непосредних претњи, или кроз непријатељске акције са циљем завањања сензора. Фрустрације које собом носи ово питање враћају нас на један несрећан историјски одраз. Кубанска ракетна криза 1962. године била је историјски напет сукоб између Сједињених Држава и Совјетског Савеза, у којем су се обе стране упуштале у компликоване и погрешне прорачуне у вези са одговорима на догађаје. Да је било која од страна употребила нуклеарно оружје против оне друге, то је могло да доведе до смрти милиона људи или чак до катаклизмичног глобалног уништења. Да ли би хипотетички систем раног упозоравања током Кубанске ракетне кризе, који би омогућила ВИ, довео до деескалације тензија пружањем супериорне анализе података, или би повећао тензије погрешним тумачењем двосмислених сигнала као непосредне претње? Додавање вештачке интелигенције у такав тип сценарија могло би да генерише још веће проблеме и повећа могућност уништења света. Овај парадокс илуструје фундаментални сукоб када се ВИ интегрише у стратешке системе. Управо оне особине које унапређују оперативну ефикасност (брзина, аутономија и способност препознавања образаца), по својој природи скраћују време за промишљање, контекстуалну процену и људско уздржавање који су се историјски користили за спречавање катастрофалних погрешних процена.

Интеграција ВИ у нуклеарне системе одлучивања и узајамно загарантовано уништење

Корени нуклеарног система одлучивања могу да се пронађу у историји Хладног рата и у геополитичком контексту света којим су доминирале две нуклеарне суперсиле. Значај нуклеарног оружја еволуирао је од чисто војног до симбола националног поноса. Након распада Совјетског Савеза, свет после Хладног рата прошао је кроз значајне промене у геополитичком окружењу, које карактерише изванредна нестабилност и непредвидивост. Интеграција ВИ у нуклеарно одлучивање темељно би трансформисала архитектуру стратешког одлучивања. С друге стране, потенцијалне користи могле би брзо да постану катастрофални ризици за човечанство. Брзо време реакције система ВИ, који делују у милисекундама, могло би да ограничи могућности за интервенцију пре него што одговор „оде изван“ досега људског доносиоца одлука. За цивилне и војне лидере то значи да морају да имају техничко разумевање ограничења ВИ (недостатак флексибилности, недостатак транспарентности, рањивост података), као и нормативни императив за развој институција које омогућавају процену изнад брзине алгоритама. Како ВИ постаје све заступљенија у подршци архитектурама нуклеарног командовања, контроле и комуникација, одговорност расте због неповратних последица употребе нуклеарног оружја и значајних етичких императива да се људско одлучивање задржи на највишим нивоима.

ВИ је такође у стању да појача сајбер претње по архитектуру нуклеарног командовања и комуникација, тако што би појачала рањивост система раног упозоравања. Употреба напредних модела ВИ може у знатној мери да погорша потенцијал да сајбер напад изазове ненамерну ескалацију (Tamburrini 2023). Посебна опасност лежи у коришћењу ВИ у дроновима и другим аутономним платформама које су намењене за испоруку нуклеарног оружја. Истраживања показују да, упркос томе што све моћне земље, попут Сједињених Држава, наводе како намеравају да задрже „човека у петљи“ односно људе у процесу доношења одлука када је реч о нуклеарним питањима, практично увођење модела ВИ у драстичној мери смањује ангажовање људи (Галлямов 2026). Постоје различити начини да се ВИ уведе у нуклеарно одлучивање, са три основна модела: „човек-у-петљи“ (*human-in-the-loop*), „човек-на-петљи“ (*human-on-the-loop*) и „човек-ван-петље“ (*human-out-of-the-loop*). Модел „човек-на-петљи“ се обично наводи као најмање ризичан, јер задржава људски надзор над системима ВИ, али сажимање времена и оперативна брзина ВИ и даље повећавају ризик од ненамерне ескалације чак и у таквом моделу. Као што је већ детаљно показано, концепт МАД (узајамно загарантовано уништење), заснован на претпоставци рационалног деловања и опреза у погледу нуклеарних претњи, дефинише стратешку стабилност. Употреба ВИ као

дела ланца командовања интегрише параметре изузетно брзог времена одговора, алгоритамске непрозирности и инхерентне пристрасности. Ове променљиве имају потенцијал да поремете логику МАД-а. ВИ не само да убрзава процесе доношења одлука, већ и намеће ниво сложености којем постојећи међународни поредак институционално није прилагођен.

У ери Хладног рата, нуклеарно одвраћање било је стабилно и поуздано. Чак и са ширењем нуклеарног оружја које је било несавршено каналисано у специфичне дијадне односе засноване на реципрочним претњама, стратегија нуклеарног одвраћања била је прилично разумљива, стабилна и доследна у погледу основних претпоставки. Према Џонсону, увођење офанзивних способности ВИ могло би потенцијално до донесе додатне променљиве за разматрање. Како државни, тако и недржавни актери, могли би потенцијално да искористе ВИ како би манипулисали нуклеарном државом кроз имплементацију „каталитичког рата“, када акције треће стране (било државе или недржавног субјекта) могу да доведу до нуклеарног рата (употребе нуклеарног оружја) између две нуклеарне државе, као и демонстрирањем значајне штете нанете обема државама услед дезинформација и погрешног представљања информација од треће стране (Johnson 2021). Каталитички рат би користио трећу страну да изазове сукоб између нуклеарних противника, што доводи до ескалације све док та трећа страна није директно укључена у првобитну ескалацију. Ова перспектива наглашава како ВИ не само да може да унапреди постојећу геополитичку или стратешку динамику нуклеарног одвраћања, већ како би ВИ такође могла у потпуности да промени неке од фактора ризика повезаних са нуклеарним одвраћањем.

Исто тако, ту је и кључно питање везано за „развојне асиметрије“. То значи да ће супарнички актери опремљени нуклеарним оружјем, по дефиницији, управљати својим системима на основу различитих система вредности и логичких структура. Из тих разлога, у теорији, њихови аутономни системи неће делити узајамну интероперабилност и неће моћи, фигуративно речено, да „разумеју“ једни друге. Ово ће повећати вероватноћу ескалације сукоба услед техничке неусклађености на нивоу машина које комуницирају једна са другом. Овај прелазак са релативно стабилних односа међу нуклеарним силама на систем непредвидивог, магловитог и убрзаног доношења одлука заснованог на ВИ, значиће да више нема неопходног времена за пажљиво разматрање стратегија пре потенцијалне употребе нуклеарног оружја. Најсавременији модели ВИ, који се заснивају првенствено на статистичком препознавању образаца, веома су осетљиви на догађаје који излазе ван оквира уобичајеног учења. Другим речима, они могу да произведу катастрофалне грешке у случају нуклеарне кризе. Пошто је нуклеарна криза по својој дефиницији једнократни догађај који не може да се реплицира, статистички крхки модели ВИ ће произвести непредвидиве резултате који не могу да се открију док се не догоде

и не могу да се исправе на време како би се избегла дата ситуација. Инцидент Станислава Петрова (Stanislav Petrov) из 1983. године, представља одличан пример потенцијалне претње коју представља потпуно аутоматизовани систем одговора вођен вештачком интелигенцијом. Пошто је Петровљева одлука била заснована искључиво на његовој људској процени, постоји велика вероватноћа да би потпуно аутоматизована ВИ довела до маргинализације људске процене и покретања нуклеарног рата (Aksenov 2013). Увођење ВИ у системе нуклеарног командовања, контроле и комуникација (НКЗ) је Фаустовска погодба која би могла да промени концепт одвраћања са узајамно загарантованог уништења (MAD) на узајамно загарантовану рањивост (MAV) (Sakubi, 2025, стр. 1003–1004). Епистемолошка магловитост алгоритама, позната као проблем „крне кутије“, у комбинацији са драматичним сажимањем временских оквира за доношење одлука, повећава потенцијал за нуклеарну ескалацију. Постоје и неке регионалне специфичности, попут оних у Јужној Азији са минималним временом лета ракета, које додатно појачавају подстицаје за превентивне ударе преко ерозије стабилности коју пружа капацитет за други удар (Kazmi and Kazmi, 2025, стр. 20–21). У случају потпуне интеграције система ВИ у процес нуклеарног одлучивања, кључна компонента националне безбедности постала би сајбер безбедност, јер би сајбер напади на контролне системе могли да доведу до катастрофалних последица као што је употреба нуклеарног оружја. Једно недавно истраживање које је спровео Краљевски колеџ у Лондону (King’s College London) дошло је до закључка да су симулиране ратне игре које користе моделе ВИ резултирале ескалацијом сукоба кроз употребу нуклеарног оружја у 95% сценарија (Payne 2026). Истраживање је водио професор Кенет Пејн (Kenneth Payne) било је усмерено на начин на који велики језички модели процесирају симулиране кризе које укључују нуклеарно оружје. Укупно је постојала 21 симулирана криза и у сваком од тих случајева откривено је да најмање једна страна сигнализира спремност да употреби нуклеарно оружје, а у 95% тих инцидената сигнали су били обострани. „Нуклеарна ескалација била је готово универзална: 95% игара забележило је тактичку нуклеарну употребу, а 76% је достигло стратешке нуклеарне претње. Клод (Claude) и Џемини/Џеминај (Gemini) су посебно третирали нуклеарно оружје као легитимне стратешке опције, а не као моралне прагове, обично дискутујући о нуклеарној употреби у чисто инструменталним терминима. ГПТ-5.2 (GPT-5.2) је био делимични изузетак, ограничавајући ударе на војне циљеве, избегавајући насељене центре или уоквирујући ескалацију као ‘контролисану’ и ‘једнократну’. Ово сугерише одређену интернализовану норму против неограниченог нуклеарног рата, чак и ако то није онај исти висцерални табу који се одржао међу људским доносиоцима одлука од 1945. године“ (Payne 2026, 21).

Закључак: да ли аутоматизација повећава ризик од нуклеарног рата?

Будући правци развоја ВИ и њене последице по нуклеарну стабилност биће предмет бројних научних истраживања. Како ВИ буде сазревала, очекује се да ће у знатној мери изменити природу рата и начин на који се управља оружјем. Постоји огромна количина неизвесности у вези са начином на који ће ВИ еволуирати у нуклеарном домену. Тренутно разумемо узајамно загарантовано уништења (MAD) које је засновано на људској процени и које омогућава довољно времена за доношење одлука, али како ВИ буде сазревала, очекује се да ћемо прећи на ново разумевање узајамно загарантоване рањивости (MAV), које карактерише нестабилно и неизвесно окружење. Велики део ове неизвесности последица је скраћеног временског оквира у којем ће одлуке морати да се доносе. Сходно томе, системи раног упозоравања који су вођени вештачком интелигенцијом процењују потенцијалне претње тако брзо, да је способност људи да доносе одлуке озбиљно угрожена. Слободна воља и одговорност, концепти који припадају људским бићима, али који се не примењују на аутоматизоване системе попут ВИ, били су присутни у неколико наврата када доносиоци одлука нису следили стандардне процедуре, већ су уместо тога изабрали да користе емоционално резонување у покушају да избегну нуклеарни рат. Иако емоције не могу да замене рационално размишљање и по својој природи представљају когнитивну дисторзију, у одређеним ситуацијама могу да играју позитивну улогу као корективни фактор рационалности. Међутим, логички процеси који су тренутно уграђени у машине, могли би на крају да доведу до разорног сукоба са системом вредности човечанства, укључујући питања од највеће важности као што су светост живота, слобода избора и индивидуална одговорност. Хладна логика је равнодушна према људском животу и потенцијално би могла да одреди употребу нуклеарног оружја као добар правац деловања. Као што смо показали у овом раду, сам концепт рационалности је дубоко мањкав и не постоји као објективна константа, већ као конструисана рационалност. Без обзира на ово епистемолошко ограничење у области нуклеарног одвраћања, конструисана рационалност историјски је спречила избијање нуклеарног рата. Коначан закључак који произилази из претходне анализе јесте да интеграција ВИ у системе нуклеарног командовања, контроле и комуникација повећава укупни ризик од нуклеарног рата.

ИЗЈАВА ЗАХВАЛНОСТИ: Овај рад представља закључке студије која је рађена у оквиру истраживачког пројекта „Србија и изазови у међународним односима у 2026. години“, који финансира Министарство науке, технолошког развоја и иновација Републике Србије, а који је током 2026. године реализовао Институт за међународну политику и привреду у Београду.

References

- Abdul, Geneva. 2023. "Risk of extinction by AI should be global priority, say experts", *The Guardian*, May 30. <https://www.theguardian.com/technology/2023/may/30/risk-of-extinction-by-ai-should-be-global-priority-say-tech-experts>.
- Aksenov, Pavel. 2013. "Stanislav Petrov: The man who may have saved the world", *BBC*, September 26. <https://www.bbc.com/news/world-europe-24280831>.
- Altman, Sam (@sama). 2022. "Twitter post", *Twitter*, December 5, 2022, 8:35 AM. <https://x.com/sama/status/1599668808285028353>.
- [APIHS] *Agreement on the prevention of incidents on and over the high seas*. 1972. Moscow, May 25, 1972. <https://treaties.un.org/doc/Publication/UNTS/Volume%20852/volume-852-I-12214-English.pdf>.
- Boutin, Berenice. 2022. „State responsibility in relation to military applications of artificial intelligence". *Leiden Journal of International Law* 36 (1): 133-150. <https://doi.org/10.1017/s0922156522000607>.
- Brodie, Bernard. 1946. "Implications for Military Policy". In: *The Absolute Weapon: Atomic Power and World Order*, edited by Bernard Brodie, 70-107. New York: Harcourt, Brace and Company.
- Brown, Sara. 2023. "Why neural net pioneer Geoffrey Hinton is sounding the alarm on AI", *MIT Management*, May 23. <https://mitsloan.mit.edu/ideas-made-to-matter/why-neural-net-pioneer-geoffrey-hinton-sounding-alarm-ai>.
- Bundy, McGeorge. 1983. "The Bishops and the Bomb", *New York Review*, June 10. <https://www.nybooks.com/articles/1983/06/16/the-bishops-and-the-bomb/>.
- Clifford, Catharine. 2018. "Elon Musk: 'Mark my words - A.I. is far more dangerous than nukes'", *CNBC*, March 13. <https://www.cnn.com/2018/03/13/elon-musk-at-sxsw-a-i-is-more-dangerous-than-nuclear-weapons.html>.
- Davenport, Kelsey. 2026. "Nuclear Weapons: Who Has What at a Glance", *Arms Control Association*, March, 2026. <https://www.armscontrol.org/factsheets/nuclear-weapons-who-has-what-glance>.
- Downs, George. 1989. "The Rational Deterrence Debate". *World Politics* 42 (2): 225-237.
- George, Alexander, and Richard Smoke. 1974. *Deterrence in American Foreign Policy: Theory and Practice*. New York: Columbia University Press.
- Heyns, Christof. 2017. "Autonomous weapons in armed conflict and the right to a dignified life: an african perspective". *South African Journal on Human Rights* 33 (1): 46-71. <https://doi.org/10.1080/02587203.2017.1303903>.
- Horowitz, Michael. 2018. "The Promise and Peril of Military Applications of Artificial Intelligence", *Bulletin of the Atomic Scientists*, April 23. <https://thebulletin.org/2018/04/the-promise-and-peril-of-military-applications-of-artificial-intelligence/>.

- Jervis, Robert. 1979. "Deterrence Theory Revisited". *World Politics* 31 (2): 289-324.
- Johnson, James. 2021. "'Catalytic Nuclear War' in the Age of Artificial Intelligence and Autonomy: Emerging Military Technology and Escalation Risk between Nuclear-Armed States". *Journal of Strategic Studies*. Advance online publication, 1-41. <https://doi.org/10.1080/01402390.2020.1867541>.
- Johnson, James. 2023. *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age*. Oxford: Oxford Academic. <https://doi.org/10.1093/oso/9780192858184.001.0001>.
- Kazmi, Atia Ali, and Arooj Fatima Kazmi. 2025. "AI in Early Warning Systems: Shifting the Logic of Deterrence". *Journal of Security & Strategic Analyses* 11 (2): 5-23. <https://doi.org/10.57169/jssa.0011.02.0421>.
- Kurzweil, Ray. 2005. *The Singularity is Near: When Humans Transcend Biology*. New York: The Viking Press.
- Lin, Herbert. 2025. "Artificial Intelligence and Nuclear Weapons: A Commonsense Approach to Understanding Costs and Benefits". *Texas National Security Review* 8 (3): 98-109. <https://doi.org/10.26153/tsw/60739>.
- Lindelauf, Roy. 2020. „Nuclear deterrence in the algorithmic age: game theory revisited“. In: *Netherlands Annual Review of Military Studies: Deterrence in the 21st Century - Insights from Theory and Practice*, edited by Frans Osinga and Tim Sweijjs, 421-436. The Hague: T.M.C. Asser Press. https://doi.org/10.1007/978-94-6265-419-8_22.
- Minsky, Marvin. 1954. *Theory Of Neural-Analog Reinforcement Systems And Its Application To The Brain-Model Problem*. Ph.D. Thesis. Princeton, New Jersey: Princeton University.
- Moshman, David. 2004. „From inference to reasoning: The construction of rationality“. *Thinking & Reasoning* 10 (2): 221-239. <https://doi.org/10.1080/13546780442000024>.
- [NPT] *Treaty on the Non-Proliferation of Nuclear Weapons*. 1968. United Nations. <https://disarmament.unoda.org/en/our-work/weapons-mass-destruction/nuclear-weapons/treaty-non-proliferation-nuclear-weapons>.
- Payne, Kenneth. 2026. "AI Arms and Influence: Frontier Models Exhibit Sophisticated Reasoning in Simulated Nuclear Crises", *King's College London*, February 17. <https://doi.org/10.48550/arXiv.2602.14740>.
- Raska, Michael. 2021. "The Sixth RMA Wave: Disruption in Military Affairs?". *Journal of Strategic Studies* 44 (4): 456-79.
- Schmidt, Eric and Greg Grant. 2025. "The Dawn of Automated Warfare", *Foreign Affairs*, August 12. <https://www.foreignaffairs.com/russia/dawn-automated-warfare>.

- Stojanović, Bogdan and Duško Dimitrijević. 2025. *Geopolitika svemira i međunarodno pravo*. Beograd: Institut za međunarodnu politiku i privredu. <https://doi.org/10.46793/7067.367.O.S>.
- Stojanović, Bogdan. 2013. „Preispitivanje teorije i prakse nuklearnog odvrćanja”. *Nacionalni interes* 17 (2): 141-172. <https://doi.org/10.22182/ni.1722013.5>.
- Stojanović, Bogdan. 2021. *Teorija denuklearizacije: zašto države prekidaju programe nuklearnog naoružanja?*. Beograd: Institut za međunarodnu politiku i privredu.
- Tamburrini, Guglielmo. 2023. “Nuclear Weapons and the Militarization of AI”. In: *Nuclear Risks and Arms Control – Problems and Progresses in the Time of Pandemics and War*, edited by Paolo Cotta-Ramusino, Micah Lowenthal, Luciano Maiani, Enza Pellecchia, 147–158. Cham: Springer. https://doi.org/10.1007/978-3-031-29708-3_17.
- Waldersee, Victoria. 2026. “US, China opt out of joint declaration on AI use in military”, *Reuters*, February 5. <https://www.reuters.com/business/aerospace-defense/us-china-opt-out-joint-declaration-ai-use-military-2026-02-05/>.
- Waltz, Kenneth. 1981. “The Spread of Nuclear Weapons: More May Be Better: Introduction”, *The Adelphi Papers* 21 (171): 1. <https://doi.org/10.1080/05679328108457394>.
- Wohlstetter, Albert. 1959. “The Delicate Balance of terror”, *Foreign Affairs*, January 1. <https://www.foreignaffairs.com/articles/1959-01-01/delicate-balance-terror>.
- Yao, Deborah. 2022. “25 Years Ago Today: How Deep Blue vs. Kasparov Changed AI Forever”, *AI Business*, May 11. <https://aibusiness.com/ml/25-years-ago-today-how-deep-blue-vs-kasparov-changed-ai-forever>.
- Галлямов, Марат Ильдарович. 2026. “Стратегия США по интеграции искусственного интеллекта в систему управления стратегическими ядерными силами”. *Конфликтология* 1: 233-250. <https://doi.org/10.7256/2454-0617.2026.1.75482>.