ISSN 2466-4693
UDC/UDK: 005:62

University “Union – Nikola Tesla “
School of Engineering Management

Univerzitet „Union – Nikola Tesla “
Fakultet za inženjerski menadžment

Serbian Journal of Engineering Management

Vol. 11, No. 1, 2026

Belgrade, January 2026

ISSN 2466-4693
UDC/UDK: 005:62

University “Union – Nikola Tesla “
School of Engineering Management

Univerzitet „Union – Nikola Tesla “
Fakultet za inženjerski menadžment

Serbian Journal of Engineering Management

Vol. 11, No. 1, 2026

Belgrade, January, 2026
Beograd, januar, 2026

Serbian Journal of Engineering Management
Vol. 11, No. 1, 2026

Publisher/Izdavač:
University "Union – Nikola Tesla ", School for Engineering Management, Belgrade
Univerzitet „Union – Nikola Tesla ", Fakultet za inženjerski menadžment, Beograd

For publisher/Za izdavača:
Prof. dr Vladimir Tomašević

Editor-in-Chief/Glavni i odgovorni urednik: Prof. dr Vladimir Tomašević

Editor: Prof. dr. Katarina Štrbac

Press/Štampa: Black and White, Belgrade

Circulation/Tiraž: 300

ISSN: 2466-4693

Contact/Kontakt:
Serbian Journal of Engineering Management
Editorial Board/Uredništvo
School of Engineering Management/Fakultet za inženjerski menadžment
Bulevar vojvode Mišića 43
11000 Beograd
casopis@fim.rs
Tel. +381 11 41 40 425

CONTENT/SADRŽAJ

A Message from the Editor-in-Chief

The Serbian Journal of Engineering Management is a scientific publication issued by the School of Engineering Management and the Society of Engineering Management of Serbia. The Ministry of Science, Technological Development, and Innovation of the Republic of Serbia classifies the journal. Since 2020, it has been indexed in EBSCO databases. The journal has been included in the ERIH Plus list since 2023. This international journal covers a wide range of topics in engineering management and industrial engineering and is published twice a year. Articles are written in English.
This special issue of the Serbian Journal of Engineering Management explores the intersection of artificial intelligence and security through several interconnected themes: the geopolitical aspects of AI competition between major powers, notably the US-China technological rivalry and its influence on global order and digital sovereignty; military applications including autonomous weapons systems, AI-enabled warfare, and defence capabilities; risk assessment and governance frameworks for AI deployment in security infrastructures; economic and legal dimensions of AI regulation and international cooperation; the role of AI in amplifying hybrid threats and information warfare, particularly in regional contexts like the Western Balkans; ethical considerations and regulatory challenges of autonomous systems for military and environmental security; industrial safety applications using machine learning in critical processes; and organisational governance issues such as safeguarding technical documentation from unauthorised disclosure to large language models and managing AI systems within transnational institutional frameworks.
The editorial board comprises distinguished academics from various countries, dedicated to establishing the highest academic standards and promoting engineering management principles in Serbia.

Information about the journal in English and Serbian is available on its webpage: https://sjem.fim.edu.rs/index.php/sjem.

Prof. Dr. Vladimir Tomašević, FRSA

Reč glavnog urednika

Serbian Journal of Engineering Management je naučno-stručni časopis, koji izdaje Fakultet za inženjerski menadžment i Društvo inženjerskog menadžmenta Srbije. Časopis je kategorisan od strane Ministarstva nauke, tehnološkog razvoja i inovacija. Časopis je takođe od 2020. indeksiran u EBSCO bazama. Časopis je indeksiran na ERIH plus listi od 2023. Ovaj međunarodni časopis je posvećen temama povezanim sa inženjerskim menadžmentom i industrijskim inženjerstvom i izlazi dva puta godišnje (u januaru i julu). Zastupljeni jezik za članke je engleski.
Ovo izdanje Serbian Journal of Engineering Management obrađuje presek veštačke inteligencije i bezbednosti kroz nekoliko međusobno povezanih tema: geopolitičke dimenzije AI konkurencije između velikih sila, posebno tehnološkog rivalstva SAD-a i Kine i njihovog uticaja na globalni poredak i digitalni suverenitet; vojne primene uključujući autonomne sisteme naoružanja, ratovanje omogućeno AI-jem i odbrambene sposobnosti; procenu rizika i okvire upravljanja za primenu AI-ja u bezbednosnim infrastrukturama; ekonomske i pravne aspekte regulacije AI-ja i međunarodne saradnje; ulogu AI-ja u jačanju hibridnih pretnji i informacionog ratovanja, posebno u regionalnim kontekstima kao što je Zapadni Balkan; etičke implikacije i regulatorne izazove autonomnih sistema kako za vojne svrhe tako i za bezbednost životne sredine; primene u industrijskoj bezbednosti kroz mašinsko učenje u kritičnim procesima; i izazove organizacionog upravljanja uključujući zaštitu tehničke dokumentacije od neovlašćenog otkrivanja velikim jezičkim modelima i upravljanje AI sistemima u transnacionalnim institucionalnim okruženjima.
Uredništvo časopisa čine istaknuti naučnici iz različitih zemalja sveta koji su posvećeni postavljanju visokog akademskog standarda i promocije principa inženjerskog menadžmenta u Srbiji.

Informacije o časopisu i poziv za autore, na srpskom i engleskom jeziku, nalaze se na web stranici časopisa : https://sjem.fim.edu.rs/index.php/sjem.

Prof. dr Vladimir Tomašević, FRSA

A Message from the Editor

Dear Readers,

It is with great pleasure that I present this volume of the Serbian Journal of Engineering Management, focusing on a crucial and transformative challenge of our time: the intersection of artificial intelligence and security in the 21st century. This collection originates from the international scientific conference “Artificial Intelligence and Security in the 21st Century,” held in November 2025. The conference gathered scholars, practitioners, and policymakers to explore how AI is fundamentally changing the landscape of global security.

The integration of artificial intelligence and security studies signifies more than a mere technological advance; it indicates a fundamental shift in our comprehension of power, governance, conflict, and human agency. As AI systems become progressively capable of autonomous decision-making, predictive analysis, and extensive information processing, they present unprecedented opportunities alongside substantial risks across military, economic, social, and political domains. This issue explores these complexities through careful interdisciplinary research that connects computer science, international relations, ethics, law, and engineering management.

This volume of the Serbian Journal of Engineering Management explores the intersection of artificial intelligence and security through several interconnected themes: the geopolitical aspects of AI rivalry among major powers, particularly the US-China technological competition and its influence on global order and digital sovereignty; military applications including autonomous weapon systems, AI-enabled warfare, and defence capabilities; risk assessment and governance frameworks for AI deployment in security infrastructure; economic and legal aspects of AI regulation and international collaboration; the role of AI in amplifying hybrid threats and information warfare, especially in regional contexts such as the Western Balkans; ethical considerations and regulatory challenges posed by autonomous systems for both military and environmental security; industrial safety applications involving machine learning in critical processes; and organisational governance challenges like safeguarding technical documentation from unauthorised access by large language models and managing AI systems within transnational institutional frameworks.

The selection process for this issue was especially thorough and competitive. Each manuscript underwent double-blind peer review by esteemed international experts in security studies, computer science, international relations, and engineering management. Reviewers assessed submissions not only for methodological rigour and theoretical contribution but also for their practical relevance to policymakers, security practitioners, and technology developers working within the complex landscape of AI-enabled security systems. The papers featured in this volume exemplify the highest standards of scholarship, providing both analytical depth and practical insights.

What sets this collection apart is its balanced and nuanced approach to exploring AI's dual role in security contexts. The articles clearly show that AI's impact largely depends on governance frameworks, ethical standards, regulatory mechanisms, and strategic decisions made by states, international organisations, technology developers, and civil society groups.

The geographical and institutional diversity of our contributors, spanning Europe, North America, the Middle East, and East Asia, ensures a variety of perspectives on AI security challenges. This diversity is especially valuable because AI governance cannot be viewed through a single cultural, political, or economic lens. The Western Balkans perspective, well-represented in this volume, provides crucial insights for medium and smaller states navigating between rival technological blocs while striving to maintain strategic autonomy, safeguard national interests, and ensure that AI development upholds democratic values and human rights.

Several key themes emerge across the contributions to this issue:

Firstly, the geopolitical dimension of AI rivalry is reshaping the global order, as major powers compete for technological dominance in AI. This rivalry has significant effects on international stability, alliance formation,

technological standards, and the future distribution of power. Our contributors explore how nations can manage these dynamics while preventing an AI arms race that might threaten global security.

Secondly, the military uses of AI, ranging from autonomous weapons systems to AI-powered intelligence analysis and cyber operations, raise significant ethical, legal, and strategic concerns. The papers in this volume critically assess the promises and dangers of military AI, focusing on accountability, human control, adherence to international humanitarian law, and the potential for AI-driven escalation in crises.

Third, governance and regulatory frameworks for AI in security applications remain fragmented and underdeveloped. Contributors to this issue analyse existing regulatory approaches at national, regional, and international levels, identifying best practices while highlighting critical gaps that require urgent attention from policymakers and international organisations.

Fourth, the role of AI in hybrid threats, including disinformation campaigns, election interference, and information warfare, presents an increasing challenge to democratic societies and regional stability. Several papers examine how AI amplifies these threats and explore how AI tools can be used defensively to detect and counter malicious information operations.

Fifth, ethical considerations influence all aspects of AI deployment in security settings. From algorithmic bias and surveillance issues to questions of human dignity and autonomy, our contributors address the core ethical dilemmas that arise when powerful AI systems are used in security decisions that affect human lives and societal well-being.

Finally, organisational and technical challenges, including cybersecurity vulnerabilities, the protection of sensitive information from AI systems, talent management in AI-intensive security organisations, and the integration of AI into existing institutional structures, require careful attention from both researchers and practitioners.

As we stand on the brink of an era in which AI will increasingly influence security outcomes across fields, the articles in this volume offer both sobering warnings and constructive pathways forward. The future of AI in security is not fixed; it will be shaped by the choices we make today regarding governance, ethics, regulation, and international cooperation. I hope this collection will serve not only as a valuable academic resource but also as a catalyst for ongoing dialogue among academia, policy communities, technology sectors, and civil society on one of the most significant challenges of our time.

The intersection of artificial intelligence and security will remain a vital area of investigation for years to come. This volume marks an important contribution to our understanding of these complex matters. Still, they are only one step in a longer journey towards ensuring that AI serves humanity's security needs while upholding our core values of human dignity, justice, and peace.

Sincerely,

Prof. dr Katarina Štrbac

Reč urednice

Poštovani čitaoci,

Sa velikim zadovoljstvom predstavljam izdanje Serbian Journal of Engineering Management, posvećeno jednom od najznačajnijih izazova našeg vremena: vezi izmedju veštačke inteligencije i bezbednosti u 21. veku. Ova kolekcija radova nastala je nakon međunarodne naučne konferencije ‘’Veštačka inteligencija i bezbednost u 21. veku’’, održane u novembru 2025. godine, koja je okupila naučnike, stručnjake i kreatore politika kako bi ispitali na koji način AI fundamentalno oblikuje arhitekturu globalne bezbednosti.

Integracija veštačke inteligencije i studija bezbednosti predstavlja daleko više od tehnološke evolucije, ona označava paradigmatsku promenu u načinu na koji razumemo moć, upravljanje, konflikt i ljudsku delatnost. Kako AI sistemi postaju sve sposobniji za autonomno donošenje odluka, prediktivnu analizu i obradu podataka u velikom obimu, oni uvode neviđene mogućnosti i podjednako značajne rizike u vojnim, ekonomskim, društvenim i političkim domenima. Ovo izdanje obrađuju kompleksne teme kroz interdisciplinarnu naučnu analizu koja prevazilazi računarske nauke, međunarodne odnose, etiku, pravo i inženjerski menadžment.

Ovo izdanje Serbian Journal of Engineering Management obrađuje presek veštačke inteligencije i bezbednosti kroz nekoliko međusobno povezanih tema: geopolitičke dimenzije AI konkurencije između velikih sila, posebno tehnološkog rivalstva SAD-a i Kine, i njihovog uticaja na globalni poredak i digitalni suverenitet; vojne primene uključujući autonomne sisteme naoružanja, ratovanje omogućeno AI, i odbrambene sposobnosti; procenu rizika i okvire upravljanja za primenu AI u bezbednosnim infrastrukturama; ekonomske i pravne aspekte regulacije AI i međunarodne saradnje; ulogu AI u jačanju hibridnih pretnji i informacionog ratovanja, posebno u regionalnim kontekstima, kao što je Zapadni Balkan; etičke implikacije i regulatorne izazove autonomnih sistema, kako za vojne svrhe, tako i za bezbednost životne sredine; primene u industrijskoj bezbednosti kroz mašinsko učenje u kritičnim procesima; i izazove organizacionog upravljanja, uključujući zaštitu tehničke dokumentacije od neovlašćenog otkrivanja velikim jezičkim modelima i upravljanje AI sistemima u transnacionalnim institucionalnim okruženjima.

Proces selekcije za izdanje bio je posebno rigorozan i kompetitivan. Svaki rukopis prošao je dvostruku recenziju koju su sproveli ugledni međunarodni stručnjaci iz oblasti studija bezbednosti, računarskih nauka, međunarodnih odnosa i inženjerskog menadžmenta. Recenzenti su ocenjivali radove ne samo prema metodološkoj rigoroznosti i teorijskom doprinosu, već i prema praktičnoj relevantnosti za kreatore politika, stručnjake za bezbednost i IT stručnjake koji se snalaze u kompleksnom pejzažu bezbednosnih sistema omogućenih AI. Ono što izdvaja ovu kolekciju radova je njen uravnotežen i nijansiran pristup ispitivanju dvostruke prirode AI u bezbednosnim kontekstima. Radovi ubedljivo pokazuju da uticaj AI fundamentalno zavisi od okvira upravljanja, etičkih smernica, regulatornih mehanizama i strateških izbora koje donose države, međunarodne organizacije, programeri tehnologija i akteri civilnog društva.

Geografska i institucionalna raznolikost naših autora koja obuhvata Evropu, Severnu Ameriku, Bliski istok i Istočnu Aziju, obezbeđuje višestruke perspektive na bezbednosne izazove AI. Ova raznolikost je posebno vredna s obzirom na to da se upravljanju AI ne može pristupiti kroz jednu kulturnu, političku ili ekonomsku prizmu. Perspektiva Zapadnog Balkana, dobro zastupljena u ovom broju, nudi ključne uvide za srednje i manje države koje se snalaze između konkurentskih tehnoloških blokova dok nastoje da održe stratešku autonomiju, zaštite nacionalne interese i obezbede da razvoj AI služi demokratskim vrednostima i ljudskim pravima.

Nekoliko kritičnih tema ističe se kroz doprinose ovom posebnom izdanju.

Prvo, geopolitička dimenzija AI konkurencije menja međunarodni poredak, sa velikim silama koje se utrkuju da postignu tehnološku nadmoć u AI sposobnostima. Ova konkurencija ima značajne implikacije za globalnu stabilnost, strukture saveza, tehnološke standarde i buduću ravnotežu moći. Naši autori ispituju kako države mogu delovati u ovoj dinamici izbegavajući AI trku u naoružanju koja bi mogla destabilizovati međunarodnu bezbednost.

Drugo, vojne primene AI - od autonomnih sistema naoružanja do obaveštajne analize omogućene AI i sajber operacija - postavljaju duboka etička, pravna i strateška pitanja. Radovi u ovom delu kritički ispituju prednosti i nedostatke upotrebe AI u vojne svrhe, obrađujući pitanja odgovornosti, ljudske kontrole, usklađenosti sa međunarodnim humanitarnim pravom i rizika od eskalacije u kriznim situacijama.

Treće, okviri upravljanja i pravne regulative za AI u bezbednosnim primenama ostaju fragmentisani i nedovoljno razvijeni. Autori u ovom izdanju analiziraju postojeće pristupe na nacionalnom, regionalnom i međunarodnom nivou, identifikujući najbolje prakse dok naglašavaju ključne praznine koje zahtevaju urgentnu pažnju kreatora politika i međunarodnih organizacija.

Četvrto, uloga AI u hibridnim pretnjama - uključujući dezinformacione kampanje, mešanje u izbore i sajber ratovanje - predstavlja rastući izazov za demokratska društva i regionalnu stabilnost. Nekoliko radova ispituje kako AI pojačava ove pretnje, dok takođe istražuje kako AI alati mogu biti primenjeni u odbrambene svrhe za otkrivanje i suprotstavljanje zlonamernim sajber operacijama.

Peto, etička razmatranja prožimaju svaki aspekt primene AI u bezbednosnim kontekstima. Od algoritamske pristrasnosti i zabrinutosti oko nadzora, do pitanja ljudskog dostojanstva i autonomije, naši autori se bore sa fundamentalnim etičkim dilemama koje nastaju kada se moćni AI sistemi primenjuju na bezbednosno ključne odluke koje utiču na ljudske živote i društveno blagostanje.

Gledajući unapred, ovo izdanje identifikuje nekoliko ključnih istraživačkih praznina i političkih izazova koji zaslužuju pažnju i u budućnosti. To uključuje potrebu za standardizovanim metodologijama procene rizika za AI u bezbednosnim primenama, razvoj međunarodnih normi i ugovora za autonomne sisteme naoružanja koji balansiraju humanitarnu brigu sa legitimnim odbrambenim potrebama, mehanizme za sprečavanje eskalacije i pogrešnih procena vođenih AI u kriznim situacijama, okvire za doprinos globalnoj stabilnosti i ljudskoj bezbednosti, pojačanu međunarodnu saradnju na istraživanju AI, i obrazovne inicijative za pripremu generacija profesionalaca za AI operativno okruženje.

Dok stojimo na pragu ere u kojoj će AI sve više oblikovati bezbednosne ishode u različitim oblastima, naučna analiza predstavljena u ovom izdanju nudi i ozbiljna upozorenja i konstruktivne predloge. Budućnost AI u bezbednosti nije unapred određena; ona će biti oblikovana izborima koje donosimo danas o upravljanju, etici, regulaciji i međunarodnoj saradnji. Nadam se da će ova kolekcija članaka služiti ne samo kao vredan naučni resurs već i kao katalizator za nastavak dijaloga između akademskih institucija, kreatora politika, tehnoloških sektora i civilnog društva o jednom od odlučujućih izazova našeg doba.

Presek veštačke inteligencije i bezbednosti ostaće ključna oblast istraživanja u godinama koje dolaze. Ovo izdanje predstavlja značajan doprinos našem razumevanju složenih pitanja današnjice, ali je samo jedan korak u dužem putovanju ka obezbeđivanju da AI služi bezbednosnim potrebama čovečanstva, dok podržava naše fundamentalne vrednosti ljudskog dostojanstva, pravde i mira.

S poštovanjem,

Prof dr Katarina Štrbac

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:355.45
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601001M

Odnos između veštačke inteligencije i nacionalne bezbednosti: geopolitičke dimenzije

Branislav Milosavljević[1], Slađan Milosavljević[2]

[1] Faculty of Business Studies and Law, "Union-Nikola Tesla" University, Belgrade, Serbia, branislav.milosavljevic@fpsp.edu.rs
[2] School of Engineering Management, "Union-Nikola Tesla" University, Belgrade, Serbia, sladjan.milosavljevic@fim.rs

Apstrakt: Razvoj veštačke inteligencije (AI) predstavlja jedan od najznačajnijih fenomena 21. veka, koji duboko menja koncepte moći, suvereniteta i bezbednosti u savremenom međunarodnom sistemu. U dobu sveobuhvatne globalne digitalizacije i rastuće tehnološke povezanosti, države ubrzano prepoznaju AI kao suštinski stratešku resurs za održavanje i očuvanje vitalnih nacionalnih interesa, kao i za efikasno projekciju geopolitičkog uticaja na globalnoj sceni. Rad analizira odnos između AI i nacionalne bezbednosti kroz geopolitičku prizmu, sa posebnim naglaskom na ulogu Sjedinjenih Američkih Država i Kine u oblikovanju novog tehnološkog poretka. Autori ispituju kako AI transformiše koncept bezbednosti, kako se koristi u vojnim, obaveštajnim i sajber strukturama, i koje rizike i etičke izazove donosi. Cilj rada je da argumentuje tezu da konkurencija za tehnološko liderstvo u oblasti AI predstavlja centralnu arenu savremenog međunarodnog nadmetanja. Štaviše, konkurencija između velikih sila nije samo tehnološka trka, već ima implikacije za uspostavljanje novih globalnih standarda, pravila angažovanja i ekonomske dominacije, čime direktno oblikuje budućnost globalnog poretka i arhitekturu međunarodne bezbednosti.

Ključne reči: veštačka inteligencija, nacionalna bezbednost, geopolitika, digitalni suverenitet, globalni poredak.

The Relationship Between Artificial Intelligence and National Security: Geopolitical Dimensions

Abstract: The development of Artificial Intelligence (AI) represents one of the most significant phenomena of the 21st century, profoundly altering the concepts of power, sovereignty, and security in the contemporary international system. In the age of comprehensive global digitalisation and growing technological interconnectedness, states are increasingly recognising AI as an essential strategic resource for maintaining and preserving vital national interests and for effectively projecting geopolitical influence on the global stage. The paper analyses the relationship between AI and national security through a geopolitical lens, with particular emphasis on the role of the United States and China in shaping the new technological order. The authors examine how AI transforms the concept of security, how it is used across military, intelligence, and cyber structures, and the risks and ethical challenges it poses. The work aims to argue that competition for technological leadership in AI constitutes the central arena of contemporary international rivalry. Moreover, the competition between major powers is not merely a technological race; it also has implications for establishing new global standards, rules of engagement, and economic dominance, thereby directly shaping the future of the global order and the architecture of international security.

Keywords: artificial intelligence, national security, geopolitics, digital sovereignty, global order.

1. Introduction

The emergence of artificial intelligence (AI) marks the beginning of a new technological era in which information, data, and algorithms become the most valuable resources of the contemporary age. In the context of national security, AI is no longer merely an instrument of technological progress, but also a means of geopolitical competition. Contemporary states are entering a new form of an arms race, not centred on nuclear warheads but on advanced algorithms, data-processing infrastructure, and predictive analytics capabilities. For this reason, artificial intelligence is increasingly becoming a strategic factor that shapes the balance of power in international

politics. The United States and China lead in the development and deployment of machine-learning-based technologies. At the same time, the European Union seeks to build its own model grounded in ethical and legal principles of digital sovereignty. However, the question that arises is not only who will achieve technological leadership, but also how that superiority will be used—not only within the sphere of national security, but also in the arena of international security (Buchanan, 2020: 16–17).

It is increasingly evident that technology is no longer merely a tool in the service of power, but a source of power in its own right. Control over the development and application of AI is becoming a prerequisite for preserving state stability and decision-making autonomy. For this reason, states are investing more heavily in their own research centres, national AI strategies, and cyber-defence infrastructures. Digital sovereignty refers to a state's capacity to control, govern, and protect its digital resources—data, infrastructure, networks, and technological systems. In the age of artificial intelligence, this concept acquires geopolitical significance equal to that of traditional territorial sovereignty. States that possess the capacity to develop and deploy artificial intelligence not only protect their information systems, but also gain political power over those states that depend on foreign technologies (Bradford, 2022, pp. 12–15).

Such dynamics have triggered an intense "artificial intelligence arms race" among the great powers, particularly between the United States and China, fundamentally redefining global power dynamics. The current trajectory of artificial intelligence is characterised by rapid advancement, emerging trends, and intensified global competition, which are reshaping both technological and geopolitical power dynamics. Understanding these elements is essential for grasping the broader impact of artificial intelligence. As a subset of artificial intelligence, machine learning focuses on enabling programs to adapt and improve when exposed to new information, without explicit reprogramming. Machine learning software can identify new, more efficient decision-making methods by training itself through data analysis. Its primary functions revolve around prediction and pattern recognition within data.

2. Artificial Intelligence as an Instrument of National Power

The development of artificial intelligence in the contemporary era represents one of the most significant civilizational shifts, increasingly reshaping the global balance of power. AI has become an instrument of strategic influence, operating simultaneously as a tool of hard power within military and security capabilities, and as a tool of soft power through cultural, technological, and informational influence over other states and societies. Within the concept of hard power, artificial intelligence is applied across a range of security and defence domains, including autonomous weapons systems, cyber defence, advanced surveillance systems, as well as intelligence algorithms for analysing large volumes of data (big data) to predict terrorist activities or enemy movements (Horowitz, 2018, pp. 5–8). The accelerated development of artificial intelligence (AI) in recent decades has opened a new chapter in geopolitical relations. States no longer compete solely on economic strength and military capabilities, but also on technological advancement, which carries with it soft power—the ability to influence others through attraction, ideas, and standards rather than coercion. It is well known that the concept of soft power was introduced by the political scientist Joseph Nye, defining it as a state's ability to shape the preferences of others by making them "want what it wants," in contrast to hard power, which is based on force and coercion (Kamarck, 2018).

Traditional sources of soft power include culture, political values, and foreign policy—that is, those societal attributes that generate admiration or consent among others (Shahla, 2025). The development of digital technologies has dramatically expanded the reach of soft power, as the internet and social media enable governments to engage global audiences in real time while bypassing traditional media channels (Shahla, 2025). Artificial intelligence represents the latest instrument in this progression—from advanced algorithms embedded in social media platforms to generative models such as contemporary chatbots. Through artificial intelligence, states can automate and personalise their media and cultural influence, shaping the narratives that circulate globally. In this way, power in the field of artificial intelligence has become a strategic resource, which is why competition between the United States and China for AI dominance is frequently emphasised. At the same time, Europe positions itself as a global regulator. Indeed, in 2024, researchers reported that institutions in the United States produced 40 significant AI models—far more than China (15 models) and the entirety of Europe (three models) (Stanford University, 2025, p. 3). This advantage in technological development grants the United States considerable influence; however, other actors are also striving to leverage artificial intelligence to project their cultural, informational, and technological standing.

However, in the digital era, the boundary between different forms of influence has become increasingly blurred. In addition to legitimate and open influence (soft power in its strictest sense), authoritarian regimes also employ covert or manipulative methods of information warfare, which some authors describe as "sharp power." Whereas

soft power is characterised by transparency, legality, and the absence of coercion, sharp power refers to concealed and harmful activities that "pierce or penetrate" the information environments of other countries to undermine institutions and societal trust. Examples of sharp power include coordinated disinformation campaigns, cyber operations aimed at interfering in elections, or algorithmic censorship targeting foreign public opinion—activities such as those documented in Russia's interference in the 2016 U.S. elections and China's covert campaigns within diaspora communities (Fu & Dirks, 2024).

In addition to the above, it is necessary to distinguish clearly between the two forms of influence. First, artificial intelligence can serve both benign diplomatic purposes and malicious, manipulative practices. Soft power in the domain of artificial intelligence entails the transparent use of AI to enhance a country's attractiveness and credibility. In contrast, its misuse—among other things for disinformation or surveillance—can generate resistance and produce counterproductive effects. In considering the use of artificial intelligence, it is evident that soft power has expanded into new dimensions, including cultural, informational, and technological influence. The cultural dimension is reflected in the fact that artificial intelligence enables the creation of new forms of content and communication (e.g., real-time language translation, personalised educational programs, virtual cultural events, and similar applications) that can contribute to the global promotion of a state's culture and values. The informational dimension encompasses artificial intelligence's capacity to shape the flow of information—for example, social media algorithms determine which news and narratives reach the broader public. In this way, algorithmic recommendations (such as those on platforms like YouTube, Facebook, or TikTok) become part of the struggle for "narrative power." States seek to increase the visibility of their perspectives, either through their own media outlets or by exerting influence over other platforms. The technological (normative) dimension of soft power in the context of artificial intelligence refers to the establishment of standards and rules governing AI itself. Those who take the lead in shaping ethical norms, technical standards, and regulatory frameworks related to artificial intelligence acquire a form of "normative power," serving as models to which others adapt. As researchers have observed, strategic influence in the sphere of artificial intelligence largely depends on who succeeds in setting the rules that others voluntarily adopt (Ariyoruk, 2025).

Generally speaking, if a state succeeds in embedding its values (such as transparency, privacy, human rights, and similar principles) into global standards for artificial intelligence, it has effectively transformed this technology into an instrument of soft power. Moreover, it should be emphasised that soft power is not exercised in a vacuum, but rather in competition among different narratives. The digital sphere has become a "battleground" where democratic and authoritarian models compete for dominant influence. For example, as social media platforms have become global channels of diplomacy, governments such as China's have developed their own alternative ecosystems (e.g., WeChat, Weibo) to promote their vision and control narratives, while simultaneously limiting Western influence (SAIS, 2025).

In addition to the above, artificial intelligence significantly enhances a state's capacity to exercise hard power by enabling new tools and methods across military, intelligence, and cyber operations. In the military sphere, AI is used to improve nearly all aspects of warfare. Autonomous combat systems are being developed, ranging from intelligent unmanned aerial vehicles and robotic ground vehicles to AI-enabled "drone swarms" capable of independently locating and engaging targets. Beyond weapons systems, AI enhances command and control through decision-support algorithms, enabling the processing of vast amounts of battlefield data in real time, the recognition of patterns, and the rapid suggestion of optimal tactical decisions that human operators could not achieve. In this way, AI has become a development priority for defence systems in many countries precisely because it allows for the automation of decision-making processes and increases the efficiency of military operations. For example, Israeli forces reportedly employed AI as a support tool to coordinate strikes and analyse intelligence data during operations against Iranian targets in 2025, thereby gaining a technological advantage over the adversary (Koppel & Parkhomchuk, 2025, pp. 2–4). Although the details of such operations remain classified, it is believed that algorithms processed reconnaissance imagery, identified key targets, and thus coordinated strikes, illustrating how AI can enhance a state's strike capability.

In the domain of intelligence operations, artificial intelligence is revolutionising data collection and analysis. Machine learning tools are already being used for intelligence analytics—from the automated recognition of objects in satellite imagery and video footage (for example, Project Maven, launched by the U.S. Department of Defence in 2017 to analyse drone imagery using the TensorFlow AI system) to predictive modelling that can indicate potential threats (Gibbs, 2018).

AI systems can sift through vast datasets (such as intercepted communications or publicly available social media posts) and identify patterns or anomalies that point to security risks—something that would be impossible for human analysts to accomplish in real time. For example, AI algorithms are used to detect terrorist plots in advance

and to predict potential conflict hotspots by processing diverse data sources far more rapidly than humans can. At the same time, advances in natural language processing enable the automated analysis of textual sources used by intelligence services. In short, AI enables intelligence communities to "see" more deeply and faster within vast seas of data, thereby providing decision-makers with richer, more timely information (Koppel & Parkhomchuk, 2025, pp. 5–8).

In the field of cybersecurity, AI operates dually, simultaneously strengthening defences and creating new offensive capabilities. From a defensive perspective, AI tools are used for intrusion detection and anomaly identification within information systems, as they can learn normal traffic patterns and instantly flag suspicious activities. Advanced AI models can predict cyber threats based on previous attacks and intrusion attempts, thereby enabling proactive action by security teams (Koppel & Parkhomchuk, 2025, pp. 8–9).

From an offensive perspective, the same technologies enable the development of new attack types through the automated discovery of vulnerabilities in adversarial software, the generation of highly convincing phishing messages, and even the use of so-called "deepfake" content (fabricated video and audio recordings generated by AI) in targeted deception campaigns. It is particularly concerning that AI can accelerate and scale cyberattacks to a level that exceeds human defensive capacities. According to some analyses, there are already cases in which AI assists attacks on critical infrastructure, as certain simulations indicate that coordinated AI-driven attacks could disrupt power grids or communication systems far more effectively than traditional cyberattacks (NSCA, 2021).

On the other hand, methods for applying AI to cyber deterrence are also being explored, including systems that automatically identify attack sources and respond with counterattacks or by deceiving the adversary. Beyond military and security applications, AI also enhances the economic instruments of hard power. States use advanced algorithms to analyse financial markets and energy price movements to optimise the application of economic sanctions and trade pressure measures. AI can predict the effects of sanctions on the targeted state's economy, detect complex schemes of sanctions evasion and money laundering, and automate the monitoring of financial flows to prevent the financing of undesirable activities. In this way, by combining rapid market data analysis with the autonomous enforcement of measures (such as the automatic freezing of suspicious transactions), states can use AI to employ economic coercion more effectively as an instrument of foreign policy.

In conclusion, AI expands the scope of hard power by elevating the speed, precision, and scale of state instruments of coercion to unprecedented levels. Naturally, the use of AI in these domains also raises new ethical and legal questions—from determining responsibility for decisions made by autonomous systems to preventing the escalation of conflict driven by the rapid operation of AI systems beyond human control. Artificial intelligence challenges traditional categories of power analysis because it does not fit neatly into the classical division among military, economic, and political power. It permeates all three domains, yet relates to them in a manner that transcends their conventional boundaries. In this sense, AI becomes a form of meta-power, enabling the optimisation and accumulation of all other forms of power. The United States and China recognised at a very early stage that a relationship of direct proportionality exists between technological advancement in AI and their position within the international order. The more successful a state is in developing AI, the more it strengthens its strategic autonomy and reduces its dependence on external actors.

3. Artificial Intelligence and the U.S.–China Rivalry

The final decades of the twenty-first century are marked by an accelerated transformation in international relations, in which technological innovations—particularly artificial intelligence—are emerging not only as sources of economic or military superiority but also as instruments for the deeper shaping of the global order. Artificial intelligence today does not function merely as a technical tool, nor as just another resource in the long history of innovation, but rather as a new axis of structural power that permeates all levels of social functioning from the production of economic value, through symbolic dominance in the information space, to the reorganization of states' military and intelligence capabilities (Kania, 2017, pp. 22–29).

The rivalry between the United States and China does not take the form of a classical geopolitical conflict in the sense of Cold War–style bipolar dynamics (Campbell & Ratner, 2018, pp. 60–63), nor can it be reduced to economic competition between the world's two largest economies. At its core, it manifests as a struggle for control over the future (Rolland, 2020, pp. 47–48), that is, for the ability to use technological means to shape the form of political institutions, market structures, data-governance models, and normative standards that will determine the position of states and individuals in the decades to come.

The United States was among the first to recognise the strategic importance of autonomous systems and algorithmic command-and-control platforms (Gentile et al., 2021, pp. 14–17), including the integration of AI into

intelligence analysis, predictive analytics, and surveillance. As early as the beginning of the 2010s, the Pentagon initiated a doctrinal shift positioning artificial intelligence as the core of the U.S. military's future capabilities. China, by contrast, is developing its own model of intelligent warfare. Chinese doctrine emphasises the integration of AI across all levels of military operations—from autonomous combat platforms to logistics, electronic warfare, and cyber offensives. China views AI as a means of achieving long-term strategic parity with the United States, a perspective reflected in massive investments in civil–military fusion, a strategic practice that enables scientific advances from the civilian sector to be directly transferred into the military apparatus (Rolland, 2020, pp. 49–51).

Semiconductors represent the most critical resource of the digital age (Statista, 2025). Without them, it is impossible to develop AI models, process large volumes of data, or produce modern weapons systems. The United States dominates chip design, while Taiwan (TSMC) and South Korea (Samsung) control the most advanced manufacturing processes. China, which depends on imports of technologically advanced chips, has identified semiconductors as its greatest strategic vulnerability. This has led to a series of U.S. sanctions in 2022–2023 that have denied Beijing access to cutting-edge equipment and lithography machines. At the same time, China has increased investment in domestic semiconductor production through programs totalling more than USD 40 billion. All of this clearly indicates that the world is moving toward deep technological bipolarization (China Daily, 2024).

As the United States and China compete for control over materials critical to artificial intelligence, export restrictions have become a central instrument of geopolitical leverage. In August 2023, China introduced licensing requirements and quantitative limits on the export of gallium and germanium, leading to price increases of more than 70 per cent, in response to U.S.-led export controls on chips implemented in 2022 and 2023 (MOFCOM, 2023).

By the end of 2024, the United States further expanded these controls, restricting semiconductor manufacturing equipment and sanctioning 140 Chinese companies (U.S. BIS, 2024), which prompted Beijing to halt exports of critical minerals to the United States. A short-lived truce in mid-2025 briefly restored some of these exports, but the agreement was subsequently breached as U.S. companies continued to suffer shortages of rare-earth elements.

In parallel with these measures, Washington exerted pressure on its allies to follow suit by restricting the export of materials and technologies relevant to artificial intelligence to China. The United States, the European Union, and allied countries, including Australia, Canada, and Japan, established the Minerals Security Partnership to reduce dependence on minerals under Chinese control (U.S. Department of State, 2025). Although African, Latin American, and Eastern European countries are not members of this Partnership, it seeks partnerships and engagement with selected countries that possess minerals and rare elements of strategic interest, including those critical to the development of artificial intelligence. Partnerships with African countries are particularly at odds with China's "Digital Silk Road" initiative and its broader "Belt and Road" Initiative, thereby making the development of artificial intelligence a new dimension of geopolitical competition in Africa.

As these tensions intensify, parallel strategies for minerals critical to artificial intelligence may divide the world into new technology-driven political and economic blocs. Divergent supply chains could give rise to distinct AI ecosystems, in which countries align with either Chinese- or Western-centred networks. Such fragmentation risks undermining international cooperation on emerging technologies and complicating interoperability, standardisation, and the scalability of artificial intelligence infrastructure across different geopolitical spheres. China represents a paradigmatic example of such an approach, particularly through the implementation of the "Belt and Road" Initiative, within which digital infrastructure and surveillance tools—such as facial recognition systems and smart city technologies—are exported to numerous developing countries. This process has resulted in the emergence of a phenomenon increasingly described in the academic literature as "algorithmic dependence," a form of technological dependency that enables Beijing, through control over critical digital infrastructure, to exert political influence over the national policies of recipient states, especially in Southeast Asia, Sub-Saharan Africa, and parts of Latin America. Such a strategy allows China to shape technological norms and security protocols in these regions, thereby further consolidating its position within the global geopolitical order.

By contrast, the United States positions itself as a defender of a liberal and democratic technological order, insisting on transparency, privacy, and the ethical use of AI. In line with this approach, the United States seeks to limit Chinese technological influence in strategic sectors such as telecommunications (particularly 5G networks), defence artificial intelligence, biotechnology, and semiconductor manufacturing. At the same time, the U.S. strategy includes economic instruments, such as sanctions and restrictions on the export of high-technology goods (especially advanced chips and the equipment required for their production), as well as the formation of technological alliances and partnerships with like-minded states. These measures are aimed at preserving Western technological superiority and limiting the spread of authoritarian technological models based on mass surveillance and state control of data. The competition between the United States and China in the field of artificial intelligence

is not exclusively technological in nature, but rather represents a profound clash between two normative models. The American model is grounded in liberal values of transparency, data privacy, and individual rights. In contrast, the Chinese model is based on the concept of digital sovereignty, in which the state exercises control over data infrastructure and algorithmic systems (Creemers, 2018, pp. 1–7).

China promotes AI as a tool for political stability and social governance; this is most clearly reflected in its social credit system. The United States, by contrast, seeks to establish global standards for ethical AI through OECD documents (OECD, 2025), the National AI Strategy, and cooperation with technology corporations. However, because these corporations possess enormous power, the American model contains internal tensions between democratic oversight and corporate interests. China uses the Digital Silk Road to export its algorithmic and infrastructural standards to Asia, Africa, and Eastern Europe, thereby gradually constructing a parallel digital order. This process generates a competition of legitimacy: which civilizational vision of the future will come to dominate the Global South (Hillman, 2021, pp. 3–9). The rivalry between the two technological superpowers is leading to the formation of digital blocs. While the United States is building a coalition of technological democracies, China is creating a network of states that adopt an authoritarian techno-model (Scharre, 2023, pp. 1–2). The consequence is the weakening of the universality of the international order and the emergence of what may be described as "algorithmic geopolitics." The conflict also has significant economic repercussions. While the United States seeks to restrict China's access to advanced chips, China is striving to achieve full technological self-sufficiency. This dynamic disrupts supply chains, raises production costs, and deepens the technological divide between competing blocs. The risks of escalation are significant, ranging from cyberattacks and technological arms races to the possibility that autonomous systems could lead to unintended military confrontation. Without the establishment of global governance mechanisms for high-risk technologies, this conflict may become a destabilising factor of the twenty-first century (Wu, 2020, pp. 101–114).

4. Consequences of Global Rivalry and Artificial Intelligence

The geopolitical competition between the United States and China in the domain of artificial intelligence increasingly demonstrates that its consequences extend far beyond bilateral relations between the two powers and are evolving into a global transformation of the international order. Every historical epoch has had its dominant source of power (Lee, 2018, pp. 1–5).

In one period, this resource was the control of territory, in another, industrial production, and in the twentieth century, nuclear technology. Today, it is the control of information and algorithms, as well as states' ability to translate technology into political, economic, and military advantage. In this sense, the rivalry between the United States and China is not merely a competition between two technological models, but a struggle over the shape of the future world order. The first global consequence of this rivalry is reflected in a security transformation unfolding across multiple levels. Whereas during the Cold War, strategic stability was based on the predictability of nuclear capabilities, today the world is entering an era in which autonomous systems, algorithms for anticipating adversary behaviour, and automated platforms for cyber offence and defence become central sources of insecurity. While the nuclear arms race could not be widely proliferated due to its high costs and technical complexity, the race in artificial intelligence can be pursued with far lower barriers to entry, meaning that the number of actors capable of threatening global stability is greater than ever before (Scharre, 2018, pp. 18–22).

As the United States and China develop their military AI systems, they enter a dynamic that generates a classic security dilemma: each advance by one side is interpreted as a threat by the other, thereby fueling further escalation. However, in the era of autonomous technologies, the consequences of such a dilemma may unfold far more rapidly and with less control, as defensive and offensive systems increasingly rely on algorithms whose decisions are not always transparent or fully understood, even by their creators (Scharre, 2018, pp. 112–119). This opens space for a new type of conflict—"high-speed conflict"—in which decisions about escalation could be partially delegated to machines rather than humans (Horowitz et al., 2018, pp. 14–18).

The second major consequence is the economic recomposition of the world. Artificial intelligence creates new forms of dependency and new global hierarchies. Countries that control big-data infrastructure, semiconductors, supercomputing capabilities, and the talent needed for AI development have the potential to become new economic superpowers (Cohen, 2019, pp. 1–5). The United States and China, as the two states leading in most of these areas, set a framework that compels third countries to align with one of the two technological systems (Farrell & Newman, 2019, pp. 55–61). This process is gradually pushing the world toward a bipolar digital structure, in which the United States leads a bloc based on liberal markets, multinational corporations, and interoperable technological architectures, while China is building a bloc grounded in state interventionism, vertically integrated technological systems, and political control of the digital space (Segal, 2018, pp. 8–16). This

division is neither formal nor institutionalised, unlike the Cold War division. Still, it is technologically far deeper, extending beyond the military and ideological spheres to encompass the very structure of digital life itself (Zuboff, 2019, pp. 107–112).

The third global consequence concerns the emergence of new forms of alliances. Instead of traditional security pacts, states today are building digital alliances that involve access to cloud infrastructure, data sharing, common standards for AI development, and the harmonisation of regulations on technological risks (Farrell & Newman, 2020, pp. 112–120). The United States is actively building coalitions within the G7, the OECD, and NATO aimed at defining democratic standards for the safe use of artificial intelligence, while China, through initiatives such as the "Digital Silk Road" and BRICS technological cooperation, is creating its own zone of digital influence (Segal, 2022, pp. 153–154). In this way, global politics is gradually shifting from the traditional diplomatic arena to a technological one, in which standards and protocols become new instruments of international power (Kissinger, Schmidt, & Huttenlocher, 2021, pp. 207–213).

The fourth consequence is the international order's increased vulnerability. Whereas in the Cold War model a conflict between two superpowers was governed by clear mechanisms of deterrence, today there is no consensus on the rules that should regulate the use of artificial intelligence for military or intelligence purposes (UN, 2019, pp. 4–9). On the other hand, international law lags behind technological change, and institutions such as the United Nations lack effective mechanisms to control the development of autonomous weapons or transnational surveillance systems (Scharre & Horowitz, 2015, pp. 2–4). In addition, tensions surrounding Taiwan—the centre of global production of the most advanced microchips—represent a potential flashpoint for a global crisis that could paralyse the entire world economy.

The fifth consequence—perhaps the most profound—concerns the transformation of the nature of power in the twenty-first century. Artificial intelligence is reshaping the very essence of what it means to be a great power. States no longer need to control territory to possess power; it suffices to control data flows, algorithmic infrastructures, and technological standards (Nye, 2011, pp. 115–121). In this context, a state that succeeds in institutionalising its own technological model at the global level gains not only economic and military advantages, but also a form of civilizational superiority. It is precisely for this reason that the U.S.–China rivalry is increasingly transforming into a struggle for normative leadership, since the power that establishes AI standards determines the rules by which the future world will operate. These consequences demonstrate that the rivalry between the United States and China is far more than a technological race. It is becoming a contest over the kind of world we wish to create: a world dominated by a liberal, market-based paradigm in which technology serves individual freedom, or a world in which technology serves state power and collective security as the foundation of stability and a hierarchical order (Farrell & Newman, 2019, pp. 42–79).

The changes brought about by the development of artificial intelligence leave little room for maintaining the existing international order in its current form. Just as the Industrial Revolution reshaped the structure of power in the nineteenth century and nuclear weapons did so in the twentieth, AI in the twenty-first century is a driver of broad systemic transformation that may lead to an entirely new architecture of global relations. This transformation is not merely a technological shift but a profound political reorganisation of the world, as AI affects the very foundations of international power: the nature of sovereignty, security, the economy, and global cooperation. Given that the rivalry between the United States and China lies at the centre of this transformation, it is possible to identify several scenarios for the evolution of the global order, each arising from different dynamics of technological development and political interaction.

5. Conclusion

The development of artificial intelligence in the first decades of the twenty-first century has become more than a technological revolution; it has emerged as a profound force of transformation in global politics, economics, and security. Whereas in previous historical eras, state power could be defined by control over material resources, territory, or military capabilities, today the true source of power is increasingly shifting to intangible infrastructure such as algorithms, data, and digital architectures. In this new reality, the rivalry between the United States and China represents not only a struggle between the world's two largest powers, but also a clash between two visions of future civilisation. It embodies, among other things, a conflict between a liberal technological model that emphasises innovation, private initiative, and global interconnectedness, and an authoritarian digital model that prioritises stability, national security, and centralisation. This indicates that the differences are not merely economic or political in nature, but lie in the very conception of the digital individual and digital society.

For this reason, the U.S.–China rivalry in artificial intelligence cannot be understood as a transient confrontation, but rather as a long-term process that will shape the structure of the international order in the decades to come.

What makes this competition particularly significant is the fact that AI technology has systemic implications: whoever controls chips controls algorithms; whoever controls algorithms controls data. And whoever controls data controls the economy, military power, and political stability. The United States enters this process as a power seeking to preserve the existing order in which its own technological and institutional dominance lies at the centre. For the United States, AI therefore serves as a means of defending the global liberal system that America shaped after 1945 and that enabled its long-term hegemony. China, by contrast, enters this process from an opposite historical position—as a power that has only recently acquired the capacity to challenge Western technological superiority and that views AI as a tool of national rejuvenation, geo-economic ascent, and civilizational emancipation from Western dominance. These two approaches cannot be dominant simultaneously, which is why the struggle over technological standards, regulation, data control, and semiconductor supply chains is so intense. The state that first succeeds in institutionalising its own model at the global level would gain a strategic advantage comparable to that once held by empires that controlled sea lanes or key energy resources.

Another dimension concerns the inevitable fragmentation of the world. Artificial intelligence does not unite the international community; it divides it. The reason lies in the fact that technology is deeply political. Autonomous systems, algorithmic governance, digital surveillance, cyber operations, and data infrastructures have become instruments of political control and international power projection. In such a world, states have little interest in open technological cooperation; instead, they seek to protect their own digital sovereignty and prevent competitors from gaining dominance. The result is the emergence of technological blocs in which digital systems, standards, and algorithmic infrastructures are incompatible. This is already evident in telecommunications, semiconductors, AI in the public sector, and data regulation.

Despite the growing confrontation, a segment of the scholarly community emphasises that the United States and China, aware of the scale of the risks involved, will be compelled at some point to establish a minimal level of cooperation in areas with existential implications, such as the development of autonomous weapons or the regulation of generative models that can influence political processes. Even if such cooperation does occur, however, it is likely to be limited, fragmented, and conducted in an atmosphere of deep mistrust. In the broadest sense, the U.S.–China rivalry in artificial intelligence is accelerating the transition toward a new form of international structure in which technological infrastructure becomes a new border, digital standards become new laws, and algorithms become new instruments of power. The question that arises is not whether the world will change, but what form that change will take: whether artificial intelligence will become a means of liberating human potential or a tool of political control; whether the digital order will be open or fragmented; and whether the future will be defined by cooperation or confrontation. At present, there is no clear answer to these questions. What is certain, however, is that the outcome of the U.S.–China rivalry will be a decisive factor in shaping that answer. In this sense, AI is not merely a technological innovation, it is the axis around which a new history of the world will be constructed.

References

1. Ariyoruk, A. (2025). From Bias to Influence: AI Governance as Soft Power, TRENDS Research, https://trendsresearch.org/insight/from-bias-to-influence-ai-governance-as-soft-power/?srsltid=AfmBOoqsAKqRTOJmVa0vX9tPJnpTvzW8bNwJWNFfbh0DYJ9EQRSwIgKy
2. Bradford, A. (2022). Digital Empires: The Global Battle to Regulate Technology. Oxford University Press.
3. Buchanan, B. (2020). The Hacker and the State: Cyber Attacks and the New Normal of Geopolitics. Harvard University Press
4. Gentile, G., Shurkin, M., Evans, T. A., Grisé, M., Hvizda, M.& Jensen, R. (2021). A History of the Third Offset, 2014–2018. RAND Corporation. Santa Monica
5. Gibbs, S. (2018). Google's AI is being used by US military drone programme (Project Maven). The Guardian, 7. март 2018, dostupno na: theguardian.comtheguardian.com
6. Zuboff, S. (2019). The Age of Surveillance Capitalism. PublicAffairs.
7. Kissinger, H., Schmidt, E., & Huttenlocher, D. (2021). The age of AI and our human future. Little, Brown and Company
8. Kamarck, E. (2018). Malevolent Soft Power, AI, and the Threat to Democracy. Brookings Institution, доступно на: https://www.brookings.edu/articles/malevolent-soft-power-ai-and-the-threat-to-democracy/
9. Kania, M. (2017). Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power, Center for a New American Security

10. Koppel, O. & Parkhomchuk, A. (2025). Artificial Intelligence as a Tool of Power in International Relations, Actual problems of international relations 1(164)
11. Lee, K. F. (2018). AI Superpowers: China, Silicon Valley, and the New World Order. Houghton Mifflin Harcourt.
12. Ministry of Commerce of the People's Republic of China (MOFCOM). (2023, August). Announcement on export control measures on gallium and germanium-related items, https://english.mofcom.gov.cn/News/PressConference/art/2023/art_36fb2d80e4b4453891bb8fc83e2b3c4e.html
13. National Security Commission on Artificial Intelligence (NSCAI). (2021). Final Report. Вашингтон: NSCAI. reports.nscai.govreports.nscai.gov
14. Nye, J. S. (2011). The Future of Power. PublicAffairs.
15. OECD, 2025. AI Principles overview. https://oecd.ai/en/ai-principles
16. Rolland, S. (2020). China's World Order Vision. The National Bureau of Asian Research. , Washington
17. Rolland, N. (2020). China's vision for a new world order, The National Bureau of Asian Research, Special Report, #83
18. SAIS Review of International Affairs – The Role of Soft Power in the Digital Age, https://saisreview.sais.jhu.edu/the-role-of-soft-power-in-the-digital-age/
19. State Council of the People's Republic of China. (2017). New Generation Artificial Intelligence Development Plan.
20. Stanford University (2025). AI Index Report
21. Scharre, P. (2023). The Dangers of the Global Spread of China's Digital Authoritarianism. Center for a New American Security
22. Scharre, P. (2018). Army of None: Autonomous Weapons and the Future of War. W. W. Norton.
23. Segal, A. (2018). The Hacked World Order: How Nations Fight, Trade, Maneuver, and Manipulate in the Digital Age. PublicAffairs.
24. Segal, A. (2022). China's alternative cyber governance regime. Council on Foreign Relations.
25. Shahla, A. (2025). The Role of Soft Power in the Digital Age. SAIS Review of International Affairs, https://saisreview.sais.jhu.edu/the-role-of-soft-power-in-the-digital-age/
26. Scharre, P., & Horowitz, M. C. (2015). An introduction to autonomy in weapon systems. Center for a New American Security
27. US. Department of state 2025, Minerals Security Partnership, https://www.state.gov/minerals-security-partnership
28. UN Group of Governmental Experts on Lethal Autonomous Weapons Systems. (2019). Report of the GGE on LAWS pp. 4-8
29. U.S. Department of Commerce, Bureau of Industry and Security (BIS). (2024). Entity List additions and semiconductor manufacturing equipment controls.),
30. https://sanctionsnews.bakermckenzie.com/us-department-of-commerce-significantly-expands-controls-targeting-indigenous-production-of-advanced-semiconductors-in-china/
31. Fu, D.& Dirks, E. (2024). The TikTok debacle: Distinguishing between foreign influence and interference. Brookings Institution, https://www.brookings.edu/articles/the-tiktok-debacle-distinguishing-between-foreign-influence-and-interference/brookings.edu.
32. Farrell, H., & Newman, A. L. (2019). Weaponized interdependence: How global economic networks shape state coercion. International Security, 44(1), 42–79.
33. Farrell, H., & Newman, A. L. (2020). Will the AI revolution cause a great power war? Foreign Affairs, 99(2), 112–120.
34. Farrell, H., & Newman, A. L. (2019). Weaponized interdependence: How global economic networks shape state coercion. International Security, 44(1), 42–79.
35. Horowitz, M.C., et al. (2018). Artificial Intelligence and International Security. Center for a New American Security.
36. Horowitz, M. C., Allen, G. C., Kania, E. B., & Scharre, P. (2018). Strategic competition in an era of artificial intelligence. Center for a New American Security
37. Hillman, E. J. (2021). The Digital Silk Road China's Quest to Wire the World and Win the Future. Profile Books Ltd, London, pp. 3-10
38. Campbell, K. M., & Ely R.(2018). The China reckoning: how Beijing defied American expectations. Foreign Aff. 97: 60-70
39. Creemers, R. (2018). China's Social Credit System: An Evolving Practice of Control. AARN: Science & Technology Studies (Sub-Topic), pp. 1-32

40. Cohen, J. E. (2019). Between Truth and Power: The Legal Constructions of Informational Capitalism. Oxford University Press.
41. China Daily, 2024: Six banks to invest in big way in IC fund, https://english.www.gov.cn/news/202405/29/content_WS66569746c6d0868f4e8e7987.html
42. Wu, X.(2020). Technology, power, and uncontrolled great power strategic competition between China and the United States. China Int Strategy Rev. 2: 99–119

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:355.45
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601011M

Значај вештачке интелигенције у заштити националне безбедности

Dejan Milenković[1], Katarina Štrbac[2], Jelena Mitić[3]

[1] Ministry of Interior, dejanmilenkovic1979@yahoo.com,
[2] School of Engineering Management, "Union-Nikola Tesla" University, Belgrade, Serbia, katarina.strbac@fim.rs,
[3] School of Engineering Management, "Union-Nikola Tesla" University, Belgrade, Serbia, jelena.mitic@fim.rs,

Апстракт: Вештачка интелигенција (ВИ) представља један од кључних технолошких иновативних алата у домену националне безбедности, омогућавајући значајна унапређења у обради података, анализи безбедносних претњи и подршци процесу доношења одлука. Њена примена у безбедносним структурама Републике Србије потенцијално доприноси већој ефикасности и прецизности у превенцији и откривању терористичких активности, организованог криминала, сајбер напада и других облика угрожавања безбедности. Посебан значај добија у контексту анализе великих података (big data), где омогућава идентификацију обрасца и индикатора потенцијалних претњи, чиме се омогућава проактиван и благовремен одговор. У оквиру специјалних истражних мера – попут електронског надзора, тајног праћења, видео-надзора, симулираних трансакција, контролисаних испорука и тајних истрага – примена ВИ може значајно повећати ефикасност и смањити оперативни ризик. Рад анализира кључне области примене ВИ у контексту националне безбедности, технолошке и институционалне изазове у њеној имплементацији, као и релевантне етичке и правне аспекте. Циљ рада је да се укаже на значај интеграције савремених ВИ решења у постојеће безбедносне системе и да се допринесе стратешком планирању развоја националне безбедности у дигиталном добу.

Кључне речи: национална безбедност, вештачка интелигенција, специјалне истражне мере, организовани криминал, тероризам, сајбер-безбедност.

The Significance of Artificial Intelligence in National Security Protection

Abstract: Artificial Intelligence (AI) is a key technological tool in national security, enabling significant improvements in data processing, threat analysis, and decision support. Its application to the security structures of the Republic of Serbia potentially contributes to greater efficiency and precision in preventing and detecting terrorist activities, organised crime, cyber-attacks, and other security threats. It gains particular significance in big data analysis, where it enables the identification of patterns and indicators of potential threats, thereby facilitating a proactive, timely response. Within special investigative measures – such as electronic surveillance, covert tracking, video surveillance, simulated transactions, controlled deliveries, and clandestine investigations – the application of AI can significantly increase efficiency and reduce operational risk. The paper analyses key areas of AI application in the context of national security, the technological and institutional challenges in its implementation, and the relevant ethical and legal aspects. The work aims to highlight the importance of integrating contemporary AI solutions into existing security systems and to contribute to the strategic planning of national security development in the digital age.

Keywords: national security, artificial intelligence, special investigative measures, organised crime, terrorism, cybersecurity.

1. Introduction

The national security of the Republic of Serbia faces a growing number of complex, multidimensional threats, including terrorism, organised crime, cyberattacks, and hybrid security challenges. In the contemporary security environment, the rapid development of advanced technologies creates the preconditions for applying new tools

and methods to counter these challenges, with artificial intelligence (AI) increasingly emerging as a key national security resource.

Protecting national security is the primary responsibility of the national security system, defined as a subsystem of the state and society composed of state and non-state actors, civilian and military sectors, tasked with preserving national interests and values from military and non-military security challenges, risks, and threats (Mijalkovic, 2007).

In the global context, artificial intelligence already plays a significant role in various aspects of national security. Its application encompasses cybersecurity, intelligence and counterintelligence activities, big-data analysis, and operational support in combating organised crime and terrorism, particularly through specialised investigative methods.

Here's an overview of the key ways artificial intelligence contributes to protecting national security:

1. **Detection and prevention of cyber-attacks:** Modern algorithms enable automated analysis of network traffic and identification of anomalies indicating potential cyber threats. Applying AI in this domain enables early detection and rapid response, which is critically important for protecting critical infrastructure (Chen et al., 2021).
2. **Processing and analysis of data from intelligence sources:** AI facilitates the systematic organisation and analytical processing of intelligence data across multiple domains, including human intelligence (HUMINT), open-source intelligence (OSINT), technical intelligence (TECHINT), and signals intelligence (SIGINT), alongside diplomatic communications, media content, scientific publications, and other information streams. Automated processes accelerate the production of intelligence assessments and recommendations for decision-makers.
3. **Security-intelligence analysis and trend identification:** Through applying machine learning and statistical models, AI can identify behavioural patterns indicating the emergence of new security risks. This enables proactive action and incident prevention (Gonzalez et al., 2020).
4. **Management of unmanned aerial vehicles and drones:** Using AI in navigation and operational engagement of unmanned aerial vehicles enables the collection of precise intelligence data in real-time, reducing the need for direct engagement of human resources in high-risk zones (Walsh, 2019).
5. **Combating organised crime:** AI algorithms enable analysis of financial flows, video and audio surveillance, communication patterns, and movements of suspected individuals, thereby improving identification of crime networks and their activities. AI is used for risk modelling and predicting criminal phenomena (UNODC, 2021).
6. **Fighting terrorism and radicalisation: Tools for analysing** internet content (textual, visual, and audio material) enable identification of narratives and activities connected to extremist ideologies. AI can recognise communication patterns indicating the radicalisation process (Sandler, 2022).
7. **Predicting security events:** Applying historical data and trends enables building analytical models for forecasting future security challenges, contributing to long-term planning and preventive strategies.
8. **Biometric identification and border security:** AI is integrated into systems for facial recognition, voice recognition, fingerprints, and other biometric characteristics. These technologies are applied at checkpoints, airports, and border crossings to prevent illegal activities and identify individuals of security interest.

With further technological development and increasing algorithmic sophistication, AI's significance in protecting national security is expected to continue to grow. However, alongside technological progress, it's necessary to establish robust ethical and legal frameworks to regulate its application, aiming to prevent potential abuses and preserve citizens' fundamental rights and freedoms. While AI represents revolutionary progress across various social and technological sectors, including the security domain, it also raises numerous questions about privacy protection, freedom of movement, communication, and non-discrimination. Applying AI in security mechanisms—particularly in surveillance, predictive policing, and biometric data processing—must align with existing legal and ethical standards (Jobin, Ienca, & Vayena, 2019; Taddeo & Floridi, 2018). In this context, education and raising awareness of the ethical aspects of AI applications represent key prerequisites for its responsible and legitimate use in the security sector. Educational programs and training within security agencies that incorporate analysis of ethical dilemmas, legal regulations, and international standards can significantly contribute to building a system in which AI applications are transparent, proportional, and subject to democratic oversight (Cath et al., 2018). Only through an interdisciplinary approach—involving experts from law, ethics,

security, technological sciences, and public policy—is it possible to develop an applicable and effective model for responsible use of artificial intelligence in the context of protecting national security.

2. Artificial Intelligence and National Security

Artificial intelligence encompasses the capacity of computational systems to perform tasks that conventionally require human cognitive abilities, such as experiential learning, speech recognition, decision-making, and natural language comprehension. Contemporary national security extends beyond the traditional preservation of state sovereignty and territorial integrity to encompass comprehensive societal protection, incorporating public health, economic stability, environmental sustainability, and social welfare. Modern national security frameworks additionally presuppose active state engagement in international and global security architectures. In this context, numerous international conventions and national strategies constitute the normative framework shaping the application of artificial intelligence to strengthen security. The Budapest Convention on Cybercrime (2001) establishes foundations for harmonising national legislation, improving investigations, and strengthening international cooperation in combating cybercrime (Council of Europe, 2001). Its implementation enables the application of AI to prevent, detect, and prosecute cyber offences. Similarly, the UN Convention against Transnational Organised Crime (2000) encourages the use of modern technologies, including AI, to discover and dismantle crime networks (United Nations, 2000). The Convention on the Rights of the Child and the Optional Protocol on the Sale of Children, Child Prostitution and Child Pornography (2000) directly encourage applying AI tools in protecting children in digital space (United Nations, 2000a). At the strategic document level, the U.S. National Security Strategy (2017) incorporates AI applications to maintain technological superiority, protect intellectual property, and strengthen cyber defence (The White House, 2017). The European Union's Artificial Intelligence Strategy (2018) emphasises an ethical approach and promotes the application of AI to improve cybersecurity and combat terrorism (European Commission, 2022). The National Strategy for Artificial Intelligence of the Republic of Serbia (2019) aims to position the country as a regional leader in this field. The document encourages the development of sophisticated technological solutions for security, including surveillance, data analysis, and the prevention of terrorist and criminal activities (Ministry of Education, 2019). Additionally, the EU Artificial Intelligence Act (2024) supports global cooperation in research and development, emphasising the responsible use of AI in line with the UN Sustainable Development Goals.

Examples of artificial intelligence applications in security include:

- **Crime forecasting systems**, which predict locations and timeframes of potential criminal activities based on analysing large quantities of data, enabling more efficient use of police resources.
- **Border control systems**, employing facial recognition tools and behaviour analysis to prevent illegal crossings and terrorist activities.
- **Cybersecurity systems**, where AI is used for real-time detection and neutralisation of cyber threats, particularly in protecting critical infrastructure.
- **Crisis management**, where AI enables rapid data processing and analysis to support decision-makers during terrorist attacks or natural disasters.

These international conventions, strategies, and examples of good practice underscore the importance of a coordinated, ethically responsible, and technologically advanced approach to the application of AI in national and international security contexts.

At the institutional level, establishing inter-agency coordination among relevant bodies, such as the Ministry of Interior, the Security Information Agency, the Ministry of Defence, the Office for Information Technologies and e-Government, and the academic and research sector, is necessary. Establishing strategic working groups to assess the ethical and technological risks of projects involving AI in security represents an important step toward responsible implementation. From a technological perspective, developing sophisticated ICT infrastructure and big-data processing capabilities (data lakes, cloud-based analysis, real-time monitoring) is a prerequisite for effective AI applications. Also, strengthening domestic scientific research institutions and the security technology start-up ecosystem is essential to ensure technological independence and resilience against external risks.

3. Artificial Intelligence and Anti-Crime Intelligence Work

The application of artificial intelligence (AI) in anti-crime intelligence operations constitutes a substantial advancement in enhancing investigative efficiency, accuracy, and responsiveness, while simultaneously strengthening preventive capabilities. These technologically advanced systems enable real-time analysis of intelligence data, improving decision-making in combating crime and terrorism.

AI is used in the context of fighting crime through several complementary approaches:

- Crime intelligence analysis. AI is used to analyse data on behavioural patterns, demographic characteristics, and members of organised crime and terrorist groups, as well as the spatial and temporal patterns of criminal activity. This enables the identification of hotspots of future crime activity, allowing the proactive deployment of police forces and other resources.
- Crime network analysis. Using advanced tools for visualisation and data processing, AI enables the discovery of hidden connections within complex criminal structures. This is particularly significant in organised crime and terrorism cases, where mapping the structure and roles of individuals is crucial for dismantling the network.
- Surveillance and border security. Applying AI to border control systems involves using facial recognition, behavioural analysis, and automated identification of suspicious individuals. This significantly improves border services' capabilities in preventing illegal activities such as human trafficking and smuggling.
- Cybersecurity. In combating cybercrime, AI systems are used to identify and respond to attacks in real time. Applying machine learning algorithms enables the detection of anomalies in network activity, the identification of malicious activity, and the prevention of compromise of critical infrastructure.
- Forensics and digital evidence analysis. AI is applied in analysing digital traces, including video and audio recordings, electronic communications, metadata, and textual documents. AI-based tools can identify relevant evidence faster and more precisely, thereby accelerating investigations and improving prospects for successful judicial outcomes.
- Psychological profiling. By analysing digital communications and behaviour, AI can help create psychological profiles of potential perpetrators. These profiles can be crucially helpful in recognising motives and action patterns.
- Geospatial analysis. Using GIS systems combined with machine learning algorithms, AI enables the analysis of spatial and temporal patterns of criminal activities. Such analyses serve both operational and strategic planning of police and intelligence agency activities.
- Methodology, tactics, and techniques in criminalistics. Methodologically, AI enables the development of advanced investigative methods, particularly in combating organised crime, terrorism, and high-tech crime. In tactics, it enables more precise planning of covert operations and evidence processing. In the technical sphere, applying AI in analysing digital devices enables automated searching, filtering, and classification of large volumes of data.

Despite numerous advantages, applying AI in fighting crime carries significant challenges. Key among them are privacy concerns, the risk of algorithmic bias, and the potential for technology abuse for mass surveillance. Therefore, an AI application must align with existing legal and ethical standards and ensure clear democratic oversight and transparency in institutional operations.

4. Artificial Intelligence and Special Investigative Methods

Applying artificial intelligence (AI) to special investigative methods represents significant progress in combating organised crime, terrorism, and high-tech crime, enabling faster, more precise, and more systematic data collection and analysis. AI encompasses a range of tools and techniques that enable the processing of large volumes of data, the identification of behavioural patterns, the prediction of criminal activity, and the support of operational actions.

- Electronic surveillance and covert tracking. AI systems enable the integration and processing of data from various sources—including telephone communications, financial transactions, and video surveillance—in real-time. Applying machine learning algorithms enables the identification of suspicious activities and behavioural patterns, significantly improving surveillance efficiency for suspected individuals (Strohmeier, 2020).
- Covert tracking, recording, and communications surveillance leverage advanced biometric recognition technologies and intelligent data extraction algorithms to analyse substantial volumes of audio-visual evidence. These capabilities enable accelerated identification of relevant subjects and their interaction patterns while materially reducing the analytical demands placed on investigative personnel.
- Simulated operations and operational-technical means. Applying AI to simulated actions and GPS device installation enables tracking suspects' movements and predicting routes and contact points. Systems can analyse data in real-time and signal operationally significant patterns (Bachner, 2017).

- Controlled delivery. AI-based analytical modules can optimise controlled delivery plans, assess risk across scenarios, and suggest optimal tactical options, thereby increasing the effectiveness of such measures (Perols, 2011).
- Electronic data processing. AI tools enable processing textual content—messages, emails, social networks—using natural language processing (NLP) to discover threats and communication patterns (Chowdhury, 2010).

AI systems that enable continuous monitoring of public spaces and real-time recognition of suspicious activities are increasingly being utilised in urban environments. These systems can automatically alert relevant authorities, thereby enhancing preventive action (Buil-Gil, 2021). The use of AI in special investigative techniques must be fully aligned with legal and ethical frameworks. In the Republic of Serbia, these methods are regulated by the Criminal Procedure Code, the Law on Personal Data Protection, the Law on the BIA, and other regulations. Special attention must be paid to privacy protection, limitations on the application of technical means, and the right to due process. In addition to legal regulation, it is crucial to ensure systematic, continuous education of investigators to guarantee the professional, legally valid, and ethically acceptable application of AI in special investigative techniques. In this regard, cooperation between state authorities and scientific institutions is important, as well as monitoring best practices and comparative models from other countries.

The application of artificial intelligence (AI) in combating organised crime, terrorism, high-tech and economic crime represents a key turning point in the evolution of the state's security capabilities. This involves transforming traditional investigative approaches into more sophisticated, data-driven models that enable deeper, faster, and more precise analysis of crime phenomena. AI systems provide support through advanced crime-intelligence analysis, the identification of latent behavioural patterns, predictive analytics, and the rapid processing of large datasets, thereby significantly enhancing the operational and strategic capabilities of security structures.

In the field of organised crime, the application of specialised investigative techniques, combined with AI, enables the analysis of social networks and structures within criminal organisations. Through sophisticated network analysis tools, it is possible to identify key actors, intermediaries, and the operational mechanisms of criminal networks (Xu & Chen, 2005). By combining these methods with predictive algorithms, tools are obtained to anticipate the activities and movements of organised groups (Chen et al., 2004).

In the field of counterterrorism, AI contributes to the analysis of digital traces, particularly on social networks, where radicalisation processes and the dissemination of extremist content can be detected. The application of special methods, such as covert surveillance, undercover operations, and open-source intelligence analysis (OSINT), supported by AI, enables timely detection of communication patterns and early warning of potential terrorist threats (Agarwal & Sureka, 2015).

When it comes to high-tech crime, AI enables advanced network traffic analysis, anomaly identification, and automated response to cyber threats in real time (Nguyen et al., 2020). Additionally, machine learning algorithms serve for the classification and identification of malicious software, thereby enabling faster and more precise protection of critical infrastructure (Buczak & Guven, 2016). In the domain of economic crime, special investigative techniques, such as financial monitoring, forensic auditing, and money flow tracking, gain a new dimension through AI integration. By analysing large datasets of transactional data, it is possible to detect irregularities indicative of money laundering, tax evasion, or financial manipulation (Perols, 2011; Jans et al., 2010). AI thereby accelerates the detection and prosecution of complex economic criminal offences. Overall, the synergy between artificial intelligence and special investigative techniques represents an essential instrument in the modern response to increasingly dynamic, technologically advanced, and transnational forms of crime. Its application contributes not only to more efficient crime suppression but also to enhancing the preventive and strategic capabilities of the security system.

5. Conclusion

In conditions of accelerated technological development and an increasingly complex security environment, the application of artificial intelligence to national security functions represents a necessary step toward modernising and enhancing the state's capacity to respond to contemporary threats. Artificial intelligence is becoming a key component of the security-intelligence system, which is based on institutional coordination, legal regulation, and technological infrastructure.

Institutions responsible for national security, as implementers of AI technologies, represent integral parts of the broader security-intelligence system. Their role encompasses the systematic collection, processing, and interpretation of data, using advanced machine learning algorithms and analysis of large volumes of information.

The use of AI in this context enables deeper situational awareness, faster response, and more efficient protection of the vital interests of the state and its citizens.

The key areas in which AI enhances the security system include:

- **Data collection** – from multiple sources, including open-source (OSINT), closed (HUMINT, SIGINT), and technical methods;
- **Processing and analysis** – through the application of machine learning algorithms for threat identification, pattern recognition, and anomaly detection;
- **Interpretation and presentation of results** – through the generation of analytical reports and support for decision-makers;
- **Multi-agency cooperation** – with AI support in connecting and exchanging information between institutions for improved coordination.

The use of artificial intelligence significantly contributes not only to operational efficiency but also to the strategic planning of national security. Its integration enables a more flexible, predictive, and evidence-based approach to security challenges.

However, alongside all the advantages, it is necessary to consistently emphasise the need for the balanced application of AI in accordance with the principles of the rule of law, the protection of human rights, transparency, and democratic oversight. Challenges such as privacy protection, algorithmic bias, and legal uncertainty must be at the centre of future normative and institutional development. Overall, the application of artificial intelligence to protect national security represents not only technological progress but also a new paradigm for shaping contemporary security policies. Its effective and responsible application can significantly enhance the state's capacity to protect its interests and respond to the complex challenges of the 21st century.

Literature

1. Agarwal, S., & Sureka, A. (2015). Using K-means clustering for detecting coordinated cyber-attack activities in a highly interactive online social network. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 777–782.
2. Artificial Intelligence Act (Regulation (EU) 2024/1689). (2024, June 13). Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
3. Bachner, J. (2017). Predictive policing: Forecasting crime for law enforcement. IEEE Technology and Society Magazine, 36(4), 34–42.
4. Bandyopadhyay, S., & Sandler, T. (2022). Effects of defensive and proactive measures on competition between terrorist groups. Journal of Conflict Resolution, 66(10), 1797–1825. https://doi.org/10.1177/00220027221108432
5. Brantingham, P. J., Valasik, M., & Mohler, G. O. (2018). Does predictive policing lead to biased arrests? Statistics and Public Policy, 5(1), 1–6.
6. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cybersecurity intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176.
7. Buil-Gil, D. (2021). Automated surveillance systems: Ethical and legal issues. Surveillance & Society, 19(1), 56–71.
8. Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the "good society": The US, EU, and UK approach. Science and Engineering Ethics, 24(2), 505–528.
9. Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: A general framework and some examples. Computer, 37(4), 50–56.
10. Chowdhury, G. (2010). Natural language processing. Annual Review of Information Science and Technology, 45(1), 101–134.
11. Council of Europe. (2001). Convention on Cybercrime (ETS No. 185). https://rm.coe.int/1680081561
12. European Commission. (2018). Artificial Intelligence for Europe (COM(2018) 237 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN
13. Ferguson, A. G. (2017). The rise of big data policing: Surveillance, race, and the future of law enforcement. NYU Press.
14. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144.

15. González Fuster, G. (2020). Artificial intelligence and law enforcement: Impact on fundamental rights. European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/656295/IPOL_STU(2020)656295_EN.pdf
16. Government of the Republic of Serbia. (2019). Strategy for the development of artificial intelligence in the Republic of Serbia. https://www.srbija.gov.rs/tekst/en/149169/strategy-for-the-development-of-artificial-intelligence-in-the-republic-of-serbia.php
17. Jans, M., Lybaert, N., & Vanhoof, K. (2010). A framework for internal fraud risk reduction at IT integrating business processes: The IFR2 framework. Information Management & Computer Security, 18(2), 111–127.
18. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.
19. Li, Z. (2019). Facial recognition technology in criminal investigations. Criminal Justice Review, 44(2), 205–221.
20. Мијалковић, С. (2007). О кризи националног система безбедности Републике Србије [On the crisis of the national security system of the Republic of Serbia]. Ревија за безбедност, 5, 41–45.
21. Министарство просвете, науке и технолошког развоја Републике Србије. (2019). Национална стратегија за вештачку интелигенцију за период 2020–2025 [National strategy for artificial intelligence for the period 2020–2025]. https://www.mpn.gov.rs
22. Ng, K. C., So, M. K. P., & Tam, K. Y. (2021). A latent space modeling approach to interfirm relationship analysis. ACM Transactions on Management Information Systems, 12(2), Article 10. https://doi.org/10.1145/3424240
23. Nguyen, T. T., Reddi, V. J., Yosinski, J., & Hooker, S. (2020). Machine learning for cybersecurity: A comprehensive survey. Journal of Information Security and Applications, 53, 102–124.
24. Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.
25. Perols, J. L. (2011). Financial statement fraud detection using a cross-sectional approach. Journal of Forensic Accounting Research, 12(1), 15–31.
26. Strohmeier, M. (2020). Big data analytics for crime prevention: Applications and challenges. Journal of Data Intelligence, 2(3), 123–136.
27. Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. Science, 361(6404), 751–752.
28. United Nations. (2000a). United Nations Convention against Transnational Organized Crime and the Protocols Thereto. https://www.unodc.org/unodc/en/organized-crime/intro/UNTOC.html
29. United Nations. (2000b). Optional Protocol to the Convention on the Rights of the Child on the Sale of Children, Child Prostitution and Child Pornography. https://www.ohchr.org/en/instruments-mechanisms/instruments/optional-protocol-convention-rights-child-sale-children-child
30. United Nations Office on Drugs and Crime. (2021). Darknet cybercrime threats to Southeast Asia. https://www.unodc.org/roseap/uploads/documents/Publications/2021/Darknet_Cybercrime_Threats_to_Southeast_Asia_report.pdf
31. Walsh, T. (2019). 2062: The world that AI made. Black Inc.
32. White House. (2017). National security strategy of the United States of America. https://www.whitehouse.gov
33. Xu, J., & Chen, H. (2005). Criminal network analysis and visualization. Communications of the ACM, 48(6), 100–107. https://doi.org/10.1145/1064830.1064834

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8: 341.35(497.11)
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601018P

Улога вештачке интелигенције у војној неутралности Републике Србије: изазови и перспективе

Aleksandar Pavić[1], Hatidža Beriša[2]

[1.] Military Academy, University of Defence, Belgrade, aleksandar.pavic@mod.gov.rs
[2.] Military Academy, University of Defence, Belgrade, hatidza.berisa@mod.gov.rs

Апстракт: Политика војне неутралности представља један од темељних принципа савремене спољне и безбедносне политике Републике Србије. У епохи вештачке интелигенције, овај концепт добија нову димензију, технолошку неутралност као услов стратешке аутономије. Развој и примена вештачке интелигенције не само да утичу на војне способности, већ обликују односе моћи, економску зависност и когнитивну контролу унутар глобалног система. У таквом контексту, питање неутралности више се не односи искључиво на војно савезништво, већ на способност државе да управља сопственим дигиталним суверенитетом и одржи равнотежу између различитих технолошких сфера утицаја. Војно савезништво се може и треба разматрати у контексту неутралности, али не као чланство, већ као референтни оквир стратегијске оријентације и ограничења.

Рад полази од претпоставке да вештачка интелигенција постаје важан фактор у дефинисању нових облика стратешке зависности и отпорности. Србија, позиционирана између западних и источних технолошких система, суочава се са изазовом да своју неутралност прошири у домен алгоритамске политике и дигиталне инфраструктуре. Уместо блоковског сврставања, вештачка интелигенција може постати инструмент интелигентног балансирања, средство којим држава јача сопствене аналитичке, комуникационе и безбедносне капацитете, задржавајући контролу над подацима и процесима од националног значаја.

Закључно, у раду се предлаже концепт „интелигентне неутралности“, који подразумева интеграцију принципа дигиталног суверенитета, етичке одговорности и технолошке независности Србије. Тај модел може послужити као теоријска и практична основа за развој новог типа неутралности у дигиталној ери, неутралности која не почива на изолацији, већ на активној контроли алгоритамских токова моћи. Србија тиме добија прилику да постане пример адаптивне државе која не само одолева утицајима великих сила, већ обликује сопствени дигитални идентитет унутар мултиполарног поретка.

Кључне речи: интелигентна неутралност, војна неутралност, технолошка диверсификација

The role of artificial intelligence in the military neutrality of the Republic of Serbia: challenges and perspectives

Abstract: The policy of military neutrality is one of the fundamental principles of the contemporary foreign and security policy of the Republic of Serbia. In the era of artificial intelligence, this concept gains a new dimension, technological neutrality as a condition for strategic autonomy. The development and application of artificial intelligence not only affect military capabilities, but also shape power relations, economic dependence and cognitive control within the global system. In such a context, the issue of neutrality no longer refers exclusively to military alliances, but to the ability of a state to manage its own digital sovereignty and maintain a balance between different technological spheres of influence. Military alliances can and should be considered in the context of neutrality, but not as membership, but as a reference framework of strategic orientation and limitations.

The paper starts from the assumption that artificial intelligence is becoming an important factor in defining new forms of strategic dependence and resilience. Serbia, positioned between Western and Eastern technological systems, faces the challenge of extending its neutrality into the domain of algorithmic policy and digital infrastructure. Instead of bloc alignment, artificial intelligence can become an instrument of intelligent balancing, a means by which the state strengthens its own analytical, communication and security capacities, while maintaining control over data and processes of national importance.

In conclusion, the paper proposes the concept of "intelligent neutrality", which implies the integration of the principles of digital sovereignty, ethical responsibility and technological independence of Serbia. This model can serve as a theoretical and practical basis for the development of a new type of neutrality in the digital era, a neutrality that does not rest on isolation, but on the active control of algorithmic power flows. Serbia thus has the opportunity to become an example of an adaptive state that not only resists the influences of great powers, but also shapes its own digital identity within a multipolar order.

Keywords: intelligent neutrality, military neutrality, technological diversification.

1. Introduction

The Republic of Serbia, faced with great challenges after gaining independence in 2006, the separation of Montenegro, i.e. the disappearance of the State Union of Serbia and Montenegro, was for the first time since 1918 in a position to independently build its position in international relations. The complex geopolitical situation, the regional environment that was significantly unfavorable to the national interests of independent Serbia, the legacy of the post-Cold War period that was marked by the so-called disintegration of a large state, wars, economic crisis and transition, caused the army to orient itself in accordance with the security policy of the newly established state. Thus, in 2007, military neutrality was declared and the entire military apparatus was redefined and transformed in accordance with the requirements and ideas of the then state administration.

The policy of military neutrality is still one of the basic postulates on which the Republic of Serbia builds its security policy. After major organizational changes, a complete strategic-doctrinal transformation and the establishment of a system that is complementary to the challenges, risks and threats to security, the Serbian Armed Forces were created in their current form. Projections of potential instability on European soil, after the changes in Ukraine in 2014 and the annexation of Crimea by the Russian Federation were a signal that the military instrument of power must be improved even in militarily neutral states. The policy of not joining military alliances, which is a legacy of military neutrality, contributed to the fact that the Republic of Serbia had to significantly strengthen its military potential both by developing its own military industry, but also by equipping it with modern weapons and military equipment purchased from abroad. In addition to changes in organization, doctrine and other segments, the segment of applying the most modern technologies stood out in particular. The technological aspect of military power has come to the fore in particular following the launch of the Russian Federation's special military operation in Ukraine in 2022. Modern technologies have contributed to the conflict having a special technological character, in addition to its hybrid, special and conventional character, as the most modern achievements in the field of technology have also been applied in the context of intelligence gathering, command, and the conduct of conventional operations. One of the technologies is artificial intelligence, the use of which has significantly transformed the use of military forces in modern military operations (Pavić, Beriša, & Jonev Ćiraković, 2024).

In the context of geopolitical instability, technological progress and global competition, classic models of state security are undergoing a fundamental transformation. For states that seek to maintain a policy of military neutrality, the issue is no longer simply one of avoiding agreements and alliances, but of the ability to strengthen their strategic autonomy in all dimensions, including the digital one. This paper argues that the advent of the age of artificial intelligence creates a new framework in which neutrality can be reinterpreted as "intelligent neutrality": A model in which a state not only does not enter into military alliances, but purposefully builds its own capacities in the field of artificial intelligence, maintains control over data, infrastructure and algorithms, and thus ensures digital sovereignty and strategic independence.

This change in the nature of neutrality is not only technological; it has profound implications for positioning in international relations, the distribution of power, economic dependence and security policy. By viewing the state as an entity that must manage its own digital identity and algorithmic flows, the paper explores whether "smart neutrality" will constitute a sustainable and meaningful model for states such as the Republic of Serbia in a multipolar, technologically shaped world. The aim of the paper is twofold, on the one hand, to theoretically define and conceptualize "smart neutrality", and on the other, to offer an assessment of the challenges and opportunities that artificial intelligence brings as part of modern military neutrality through an analysis of Serbia's position.

Methodologically, the paper will combine theoretical analysis of works from public sources related to the literature on international relations, law, sovereignty theory and military doctrine, conceptual synthesis and empirical, comparative analysis, as well as an assessment of contemporary international trends in the development of artificial intelligence and digital sovereignty.

2. Theoretical framework

Historically, military neutrality represented a clearly defined legal status of a state that refrained from participating in armed conflicts and did not join military alliances. The traditional model, which developed from the Congress of Vienna to the Hague Conventions, was based on three fundamental principles: non-aggression, non-cooperation, and non-prejudice. Neutrality was then primarily a legal concept, and only secondarily a political one. However, after the Cold War and the disintegration of strict bloc divisions, the classical legal definition was no longer sufficient to explain the place of states that sought to avoid structural alignments, but at the same time actively participated in international missions, economic integration, or political partnerships. Researchers such as Gardener and Karsh indicate that contemporary neutrality has transformed from a static status into a dynamic risk management strategy, in which the state balances between multiple centers of power, but without formal membership in alliances (Gärtner, 2020; Karsh, 2011). In the 21st century, especially after 2014 and the conflict in Ukraine, neutrality enters a postmodern phase in which technological, informational and cognitive factors play a key role, and not only military ones. Artificial intelligence, cyberspace and digital infrastructure become new terrains of alignment, which leads to the emergence of a new concept of technological neutrality as a prerequisite for strategic autonomy.

Consideration of the theoretical framework of military neutrality cannot be separated from a traditional approach that includes normative legal foundations, case studies of military neutrality that are internationally recognized and accepted, consideration of digital sovereignty as an inseparable part of the existence of a modern state, and consideration of the possibilities of artificial intelligence in this concept.

2.1. Military neutrality between traditional and modern

Military neutrality (hereinafter: neutrality) in theoretical frameworks can be studied from the standpoint of classical and contemporary theories. Therefore, if we look at classical theories, the first aspect of the theoretical framework is a legal category. Traditionally understood, military neutrality is the institutionalized status of a state that does not participate in armed conflicts and does not join military-political alliances. This model is based on the Hague Conventions of 1907, which guarantee neutral states certain rights, but also impose obligations such as the prohibition of ceding territory to warring parties or participating in military operations. However, the Hague framework was created in the era of territorial wars and is limited to the physical dimension of conflict, while modern forms of power such as cyber operations, information warfare, artificial intelligence, go beyond the traditional understanding of neutrality.

The changes that are an integral part of international relations have not bypassed the states that have proclaimed neutrality as a framework for the realization of their interests. Modern views on the status of neutrality speak in favor of the fact that previously neutral states can always and often must discriminate against a state that is waging an illegal war (Haque, 2022). An example is military aid to Ukraine. Namely, some authors believe that in this way the obligations of neutrality towards Russia are not violated, that is, that they do not exist and have never existed because the classical law of neutrality assumes the legal equality of warring states. A neutral state could abandon its “obligation” of impartiality between the warring parties at any time, simply by using its discretionary right to declare war on any party (Haque, 2022). The above examples are supported by the fact that the classical law of neutrality was abolished, first by the Pact of Paris, and later by the Charter of the United Nations, i.e. the prohibition of aggression was accepted and recognized by the international community of states as a whole as an imperative norm of general international law.

On the other hand, the authors Kolb and Meret in their analysis present how in contemporary conflicts the concept of neutrality is increasingly being broken down into different statuses that are not clearly defined in classical international law. The authors show that the traditional neutrality from the Hague Conventions, which refers to a strictly impartial and restrained position, is today often inadequate for complex situations such as the conflict in Ukraine. In response to the new changes in the interpretation of international frameworks, new forms of neutrality are emerging in practice, such as differential neutrality, which is aligned with the UN collective security system. Then there is qualified neutrality, where states claim to be neutral but support one side, which essentially destroys the classical concept. Finally, non-belligerence as a central position. Nominally, a state does not participate in hostilities, but can support one side, although this status is legally insufficiently defined (Kolb & Meret, 2025). The fact is that changes in the understanding of the concept of neutrality lead to the expansion of vague categories that can completely blur the distinction between neutrality and actual participation in the conflict.

If we consider military neutrality as a strategic orientation, we gain a new dimension in the application of this concept. Contemporary literature indicates that neutrality today is primarily a strategic practice, not just a legal status. It can encompass selective participation in various peacekeeping missions, political balancing, maintaining maneuvering space, avoiding long - term commitments that would limit state autonomy (Lottaz & Reginbogin, 2018; Radoman, 2021) This perspective opens up space for a more dynamic interpretation of neutrality, which is important for understanding the position of Serbia, which defines its neutrality politically, rather than formally and legally.

As an example of good practice, Ireland's policy of military neutrality is a long-standing and deeply rooted element of its foreign policy, based on non-membership in military alliances and avoidance of mutual defense obligations. This position allows Ireland to maintain an active and credible role in peace support operations, crisis management and peace diplomacy, as well as to promote human rights, development and global disarmament (European Commission, 2025). Neutrality expands the scope of its foreign policy effectiveness, particularly within the UN and the EU, while its special status within European integration is formally recognized and guaranteed by the Protocol to the Lisbon Treaty.

In addition to the previously mentioned facts and positions related to neutrality, it is important to take into account the strategy that a state resorts to in order to achieve its interests in complex international relations without violating the status of military neutrality. Thus, we can single out several approaches that are completely relevant from the point of view of Serbia. Small states try to balance powerful actors by balancing in relations with other international actors, without directly aligning. On the other hand, they try to align themselves with a more powerful side for the sake of benefit and protection, the so-called bandwagoning concept. Of particular interest is the concept *of hedging,* i.e. the strategy of simultaneous limitation and engagement, i.e. avoiding choices while maintaining *relations* with multiple actors, which is recognized as part of Serbia's strategic pragmatism (Pavić & Beriša, 2025). The hedging concept is particularly relevant to this work because it more closely describes the position of a state that faces two technological blocs, on the one hand the USA/EU and on the other hand the People's Republic of China/Russian Federation, and wants to maintain autonomy in the choice of technologies.

2.2. Intelligent neutrality as a concept

If we look at military neutrality from the perspective of contemporary threats, which are multidimensional, non-transparent and technologically advanced, we can consider the same concept through non-traditional frameworks. Thus, Čaloud points out that military neutrality in the 21st century is seriously challenged by hybrid threats, cyberattacks and wars that no longer respect classical legal frameworks. The author emphasizes that neutrality is not protection in itself, but requires a clear strategy, defense capacities and political determination to preserve it. A comparison of different models, especially with the example of Switzerland, shows that successful neutrality implies active and well-funded defense, and not just a declarative stance (Čaloud, 2025). It can be concluded that neutrality can be an advantage, but only if it is realistically designed and aligned with modern security threats.

The threats previously identified as a framework that diverges from the classical approach of military neutrality imply that this paper introduces a theoretical model of “intelligent neutrality”. This model is based on both traditional and contemporary postulates. First, military Neutrality must exist in the classical sense, i.e. the principle of non-alignment with military alliances must be respected. The second postulate is reflected in digital independence, i.e. that the state must have control over the infrastructure and data, which it should provide algorithmic autonomy and cognitive security. The response to technologically advanced threats leads us to the third postulate, which is technological pluralism, which is reflected in the diversification of partners in the technological domain.

The previous chapter explained the theoretical concept of military neutrality, where it is evident that despite changes in the international system, traditional sources of military neutrality must still be relevant, primarily due to the normative and legal foundation of the concept. Another separate postulate of intelligent neutrality is reflected in digital sovereignty. Digital sovereignty, according to Baldoni and Di Luna, implies the ability of the state to controls data generated on its territory, ensure independence from critical digital infrastructures, manages technological dependencies that can threaten political autonomy (Baldoni & Di Luna, 2025). This expansion of sovereignty becomes fundamental in a world where algorithms, models and data become the new points of power. In addition to digital sovereignty, a necessary condition must be the control of algorithmic sovereignty. Algorithmic sovereignty is a narrower and more specific concept that refers to the control over key models of artificial intelligence, software supply chains, critical decision-making algorithms, the possibility of modifying and revising model structures (Martin, 2023). If a state does not control the algorithms that govern its security or

economic sectors, it loses autonomy, even if it formally maintains military neutrality. If we implement artificial intelligence into the system, such systems are inherently dependent on data, computer infrastructure, model, but also human resources. These dependencies are becoming a new form of strategic pressure, which puts countries like Serbia in a position where they must carefully manage their technological partnerships, that is, diversify their technological partners.

3. Artificial intelligence as a transformer of military neutrality

The position of neutral states in the modern security environment increasingly depends on their ability to understand, control and integrate artificial intelligence into their own defense and security architecture. It is increasingly accepted that artificial intelligence is becoming a key instrument of power, a means of strategic influence and a source of new geopolitical dependence, thereby changing the traditional meaning of neutrality as the absence of military alliances. Instead of the classic dilemma of "alliance or independence", states are faced with the question of technological dependence or digital autonomy. For this reason, artificial intelligence is increasingly becoming an instrument for the implementation of defined strategies and as such can be a factor that defines the real limits of military neutrality. In this chapter, artificial intelligence will be considered in the context of the necessity of diversifying technological sources from multiple sides, as well as the transformative power of artificial intelligence in military forces, which is a fundamental factor in the achievement of military neutrality.

If we imagine that instead of the previous bloc divisions from the time of cold war, we can today promote technological ecosystems as new security blocs, then surely a state or alliance that has a more developed application of artificial intelligence can hope for supremacy. If we take, for example, today's situation where in the Eurasian region, instead of the classic division into NATO, opposing and non-aligned states are forming technological blocs (Winkler, 2025). The Western system consisting of the USA and EU members, the Chinese model of digital infrastructure and the partially independent systems of India, Israel and South Korea are becoming new centers of strategic influence. Neutral states are not exempt from these processes. Although they are not formally in alliances, their digital sovereignty directly depends on whose technology they use, especially in the areas of information and telecommunications technologies, which are of vital importance for the development of artificial intelligence. Control over technological architecture means control over defense, information flows, and national security. For states that strive for neutrality, it is no longer crucial that they do not belong to a military alliance, but rather that they do not become dependent on a single technological bloc. It is precisely the diversification of a neutral state's technological dependence on a single source that is one of the most important factors of strategic resilience. The fact is that no neutral state that has no alternatives in vital areas, such as energy, technology, and weapons and military equipment, can formally ensure its neutrality.

Artificial intelligence generates new forms of power ranging from surveillance, data analysis and predictive intelligence analytics to cyber offensive capabilities and cognitive influence. Depending on who controls critical systems, artificial intelligence can strengthen state autonomy or become a tool for political pressure (Schneier & Sanders, 2025). For neutral states, which do not have guarantees of collective defense, this dilemma is particularly pronounced because the choice of technological partners can enhance security independence or permanently undermine it.

In addition to the importance of technological diversification in the concept of artificial intelligence-related technologies, another factor that stands out is its impact on military forces, or rather on defense capabilities. Artificial intelligence represents the greatest revolution in military capabilities since nuclear weapons, reshaping the domains of command and control, intelligence and reconnaissance systems, cyber defense, logistics, autonomous platforms, and communications. (Bin Rashid et al, 2023)The militaries of major states and alliances are integrating artificial intelligence into multi-domain doctrines and concepts such as "mosaic warfare" (Clark, Patt, & Schramm, 2020), which changes the rules of modern operations and strategic action. In such an environment, artificial intelligence ceases to be exclusively technological, but becomes a geopolitical issue. The choice of a technological partner becomes the choice of a strategic support, which is especially pronounced in states that strive for military neutrality (Taddeo & Floridi, 2018).

Artificial intelligence also enables the development of situational awareness that was previously available mainly to members of large defense alliances, thanks to the ability to analyze large amounts of data, early detection of threats, and predictive risk assessment. Balmfort illustrates in his article how dependence on foreign allies can limit the strategic autonomy of a state in conflict. The US threat to suspend intelligence support and arms deliveries if Ukraine does not accept the US peace plan shows that military support can become an instrument of political and strategic pressure (Balmforth, 2025). For states seeking to maintain military neutrality, this example highlights

a key challenge: even formally, independent states can become dependent on the capabilities provided by major patrons, which affects their ability to make independent decisions on security and defense issues. In this sense, the situation in Ukraine serves as a warning that true neutrality implies the development of domestic and diversified capabilities, including technology and intelligence resources, in order to minimize external dependence and preserve strategic independence. For states that do not rely on allied intelligence infrastructure, these capabilities can partially compensate for structural resource gaps and raise the level of independent strategic assessment. Such models allow neutral states to develop their own risk assessment and security analysis mechanisms, which directly strengthens their autonomy in defense decision- making.

Although autonomous systems significantly improve operational capabilities, their value depends on software architecture, cloud infrastructure, cryptography and algorithms, which are often owned by foreign actors. Such dependence can generate serious security risks, including the possibility of surveillance, manipulation or political pressure (Timmers, 2019). In addition, due to the limited resources for the development of new high-tech systems, most small and militarily neutral states have to purchase them on a turnkey basis, often leaving them without control over critical parts of the technology, thus creating a new form of strategic dependence on larger states. Therefore, it is crucial to establish hybrid, partially sovereign artificial intelligence systems that combine domestic capabilities with carefully controlled external partnerships, in order to minimize dependence in critical segments.

Finally, the application of artificial intelligence in the defense structures of militarily neutral states raises complex ethical, normative and political issues related to accountability, privacy, transparency, control of algorithms and compliance with international law (Cath et al, 2018). Militarily neutral states must be particularly cautious because they do not have the protection of alliance mechanisms and institutions that could cushion the consequences of possible mistakes or disputes. For Serbia, which seeks to preserve both security autonomy and political credibility, it is crucial to develop a normatively aligned, transparent and ethically oriented strategy for the application of artificial intelligence that will ensure that technological development serves to strengthen, rather than erode, national independence.

4. Artificial intelligence and strengthening Serbia's military neutrality

In the modern multipolar system, the concept of military neutrality is undergoing a significant transformation under the influence of digital technologies, and in particular artificial intelligence. The traditional meaning of neutrality as the absence of formal alliances is no longer sufficient, as states, striving for independence must simultaneously manage technological dependencies, digital sovereignty, and algorithmic power flows. Artificial intelligence provides opportunities for increasing independent security capacities, reducing dependence on great powers, building digital sovereignty, and improving negotiating positions, but at the same time carries the risks of undermining neutrality through dependence on a single technological partner, data control by foreign actors, and integration into other people's algorithmic systems. Analyzing the possibilities of artificial intelligence for militarily neutral states in the context of Serbia, it is concluded that it can act in two directions, namely to be stronger, but also to be a factor undermining military neutrality. Therefore, artificial intelligence for Serbia's military neutrality becomes a platform of strategic choice. Whether technological development will increase or decrease the degree of neutrality depends on the model of implementation, partnership, and control.

Serbia formally defined a policy of military neutrality in 2007, but its essence is being built through three parallel processes, the historical experience of non-alignment and resistance to large alliances, on the line of contact between East and West, and the need to preserve strategic flexibility in conditions of multiple levels of pressure. Instead of formal neutrality modeled after Austria or Switzerland, Serbia is developing a pragmatic and geopolitically adaptive neutrality. Through the application of the concept of strategic pragmatism, Serbia is trying to achieve its national interests in a rational and long-term sustainable manner. This approach is in line with the geopolitical reality of Serbia as a small state at the crossroads of global interests (Pavić & Beriša, 2025). Instead of formal and static neutrality, Serbia is developing a multidimensional, adaptive strategy that includes military cooperation with NATO through the Partnership for Peace, strategic reliance on the Russian Federation in the field of energy and defense technology, economic integration with the EU, and technological partnership with the People's Republic of China. This structural multi-vectored allows Serbia to use artificial intelligence as a platform for strategic choice, where maintaining a balance between different global technological systems is crucial.

A particular challenge for Serbia is its positioning between the Western and Eastern technological systems (Mulvihill, 2025). In the modern digital environment, neutrality is no longer a matter of military power alone, but also of control over information flows, algorithms and critical infrastructure. Artificial intelligence enables advanced detection of cyber-attacks, automated response, management of cognitive space and prevention of

disinformation, which is especially important for states that cannot count on allied support in the information and cyber spheres (Brundage et al, 2018). For Serbia, which is exposed to the influence of various technological and political actors, maintaining digital sovereignty becomes a prerequisite for functional military neutrality. Strategic autonomy in this context implies control of critical data, infrastructure, algorithms and integration with global systems in a way that does not threaten national independence.

The strategic model that Serbia can implement is based on algorithmic neutrality, which involves diversifying technological partners, developing domestic AI-based systems for critical functions, and integrating cognitive and ethical resilience. This allows minimizing the risk of technological dependence and preserving room for maneuver in the conditions of competition between great powers from the East and the West. Digital sovereignty and controlled algorithmic flows allow the state to manage its own strategic position and make neutrality an active, rather than a passive, political choice.

Establishing intelligent neutrality in Serbia requires a combination of technological, institutional and cognitive capacities. This primarily involves the development of domestic solutions for intelligence analytics, cyber defense and protection of critical infrastructure. The secondary segment involves the implementation of mixed architectures with foreign partners with control of critical components. The tertiary segment is viewed through compliance with EU regulations without loss of autonomy; strategic communications for the prevention of disinformation and the formation of an expert base for artificial intelligence and cybersecurity. Intelligent neutrality can enable Serbia to preserve strategic autonomy, reduce dependence, strengthen national capacities and prevent cognitive and information threats, representing a model of adaptive, flexible policy in a multipolar and digitally transformed world.

5. Conclusion

In conclusion, the analysis shows that the age of artificial intelligence fundamentally changes the nature of military neutrality, expanding it from a legal-doctrinal framework to a technological, algorithmic and cognitive dimension. Classical military neutrality is no longer sufficient to ensure the strategic autonomy of a state in the conditions of global technological competition, where the control of power is determined not only by armed capabilities, but also by data, infrastructure, algorithms and digital information flows. In such an environment, the need for the concept of “intelligent neutrality” arises, which integrates the traditional principles of aversion to military alliances with the new demands of digital sovereignty, algorithmic control and technological diversification.

Serbia, as a country between major geopolitical and technological systems, is in a particularly sensitive position. Artificial intelligence depending on the implementation model can become an instrument for strengthening its neutrality or a factor in its undermining. If developed through strategic diversification of partners, control of critical data, building domestic capacities, and an ethically and legally harmonized normative framework, artificial intelligence becomes a mechanism for strengthening strategic autonomy and reducing dependence on great powers. In contrast, uncritical or one-sided technological dependence, especially in the areas of cybersecurity, intelligence analytics, and infrastructure, can lead to the erosion of neutrality and the transfer of part of the state's decision-making capacity to external actors.

Achieving intelligent neutrality requires a multi-layered approach. The development of domestic intelligence and analytical platforms based on artificial intelligence; control and a hybrid management model for critical digital infrastructure; building a human resource base in the field of artificial intelligence and cybersecurity; alignment with minimum European and international ethical and legal standards; and strategic communications aimed at increasing the cognitive resilience of society. As a result, neutrality ceases to be a static legal status and becomes a dynamic risk management strategy in a multipolar digital order.

The key implication of this paper is that Serbia has a real opportunity to transform the existing model of military neutrality into the concept of intelligent, algorithmically anchored neutrality, based on digital sovereignty, technological diversification and institutional adaptability. If this strategy is consistently developed, Serbia can become an example of a state that in the era of artificial intelligence not only preserves neutrality, but also turns it into a source of strategic strength and resilience.

Literature

1. Baldoni, R., & Di Luna, G. (2025). Sovereignty in the Digital Era: The Quest for Continuous Access to Dependable Technological Capabilities. *IEEE Security & Privacy*, *23*(1), 91-96. doi:https://doi.org/10.1109/MSEC.2024.3500192

2. Balmforth, T. (2025, 11 22). *Exclusive: US threatens to cut intel, weapons to press Ukraine into peace deal, sources say*. Retrieved 12 07, 2025, from https://www.reuters.com/world/europe/us-threatens-cut-intel-weapons-press-ukraine-into-peace-deal-sources-2025-11-21/
3. Bin Rashid, A., Kausik, A., Hassan, A., Bappy, A., & Mehedy, H. (2023). Artificial Intelligence in the Military: An Overview of the Capabilities, Applications, and Challenges. *International Journal of Intelligent Systems, 2023*(1), 31. doi:https://doi.org/10.1155/2023/8676366
4. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., . . . Scharre, P. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.* arXiv preprint arXiv. doi:https://doi.org/10.48550/arXiv.1802.07228
5. Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and engineering ethics, 24*(2), 505–528. doi:https://doi.org/10.1007/s11948-017-9901-7
6. Clark, B., Patt, D., & Schramm, H. (2020). *Mosaic Warfare: Exploiting Artificial Intelligence and Autonomous Systems to Implement Decision-Centric Operations.* Washington, DC: Center for Strategic and Budgetary Assessments (CSBA).
7. Čaloud, A. (2025, 05 16). *Military neutrality in the 21st century: an advantage or a strategic illusion?* Retrieved 12 07, 2025, from https://www.czdefence.com/article/military-neutrality-in-the-21st-century-an-advantage-or-a-strategic-illusion
8. European Commission. (2025, 12 04). *Ireland and EU defence & security*. Retrieved 12 07, 2025, from https://www.ireland.ie/en/dfa/role-policies/international-priorities/peace-and-security/neutrality/
9. Gärtner, H. (2020, 03 21). *FRIENDS WITH ENEMIES: NEUTRALITY AND NONALIGNMENT THEN AND NOW*. Retrieved 12 07, 2025, from https://homepage.univie.ac.at/heinz.gaertner/?p=2599&utm_source=chatgpt.com
10. Haque, A. (2022). An Unlawful War. *AJIL Unbound, 116*, 155–519. doi:https://doi.org/10.1017/aju.2022.23
11. Karsh, E. (2011). *Neutrality and Small States.* New York: Routleadge.
12. Kolb, R., & Meret, B. (2025, 03 19). *Clarifying Neutrality: The Rise of Different Statuses?* Retrieved 12 07, 2025, from https://lieber.westpoint.edu/clarifying-neutrality-rise-different-statuses/
13. Lottaz, P., & Reginbogin, H. (2018). *Notions of Neutralities.* Lanham: Lexington. doi:https://doi.org/10.5771/9781498582278
14. Martin, M. (2023). Algorithmic sovereignty: Machine learning, ground truth, and the state of exception. *Philosophy & Social Criticism, 51*(7), 1044-1074. doi:https://doi.org/10.1177/01914537231222885
15. Mulvihill, C. (2025, 09 23). *Assessing Serbia's ground forces procurement efforts*. Retrieved 12 07, 2025, from https://euro-sd.com/2025/09/articles/armament/46589/assessing-serbias-ground-forces-procurement-efforts/
16. Pavić, A., & Beriša, H. (2025). STRATEGIC PRAGMATICITY AS A FACTOR IN ACHIEVING SERBIA'S NATIONAL INTERESTS. *Serbian Political Thought, 92*(4), 25-49. doi:https://doi.org/10.5937/spm92-59186
17. Pavić, A., Beriša, H., & Jonev Ćiraković, K. (2024). Artificial Intelligence as a Factor of Change in the Use of Military Forces. *Međunarodna politika, 75*(1192), 557-574. doi:http://dx.doi.org/10.18485/iipe_mp.2024.75.1192.4
18. Radoman, J. (2021). *Military Neutrality of Small States in the Twenty-First Century.* Cham: Palgrave Macmillan. doi:https://doi.org/10.1007/978-3-030-80595-1
19. Schneier, B., & Sanders, N. (2025, 10 21). *AI Will Be Your Personal Political Proxy*. Retrieved 12 07, 2025, from https://ai-frontiers.org/articles/ai-will-be-your-personal-political-proxy
20. Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752. doi:https://doi.org/10.1126/science.aat5991
21. Timmers, P. (2019). Ethics of AI and Cybersecurity When Sovereignty is at Stake. *Minds & Machines, 29*, 635–645. doi:https://doi.org/10.1007/s11023-019-09508-4
22. Winkler, S. (2025). New and Old Cold Wars: The Tech War and the Role of Technology in Great Power Politics. *Global Studies Quarterly, 5*(2), 1-13. doi:https://doi.org/10.1093/isagsq/ksaf038

Original Scientific Paper/Original naučni rad UDC/UDK: 005.334:004.8
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601026M

Konceptualizacija sadržaja procene rizika primene veštačke inteligencije (AI) u sistemima bezbednosti

Milica Mladenović[1], Katarina Janković[2], Nenad Komazec[3], Zoran Vučinić[4]

[1] Regional Association for Security and Crisis Management, Belgrade, Republic of Serbia mladenovicmilica21@yahoo.com,
[2] Technical test center, Ministry of Defense, Belgrade, Republic of Serbia Jankovickatarina95@gmail.com,
[3] Military academy, Ministry of Defense, Belgrade, Republic of Serbia nenadkomazec@yahoo.com,
[4] Karlovac University of Applied Sciences, Karlovac, Republic of Croatia zoranvucinic5@gmail.com

Apstrakt: Primena veštačke inteligencije u svakodnevnom životu ljudi je postala neizbežna, posebno u poslovnom okruženju. Međutim, većina organizacija potvrđuje činjenicu da je upotreba AI Sistema neophodna, ali su neophodne i smernice za prilagođavanje ubrzanoj ekspanziji veštačke inteligencije. Sistemi bezbednosti predstavljaju posebno osetljivu kategoriju kada je u pitanju primena veštačke inteligencije s obzirom na ogromne nepoznanice o rizicima koje ovi sistemi sa sobom nose. Sadržajni okvir za procenu rizika u ovoj oblasti se nameće kao neophodan, posebno, jer postojeće metode za procenu rizika uglavnom ne prepoznaju AI rizike, čime se onemogućava njihovo sveobuhvatno sagledavanje. U ovom radu je dat sveobuhvatan konceptualni sadržaj procene rizika primene AI u sistemima bezbednosti koji obuhvata rizike tokom celokupnog životnog ciklusa AI sistema čime se omogućava identifikacija, analiza, ocena i tretman prepoznatih rizika i praćenje i kontrola njihovog uticaja na sistem.

Ključne reči: procena rizika, AI, sistem bezbednosti

Risk Assessment Content Conceptualization for the Application of Artificial Intelligence in Security Systems

Abstract: The application of artificial intelligence in people's daily life has become inevitable, especially in the business environment. However, most organizations acknowledge the fact that the use of AI Systems is necessary, but guidelines are also necessary to adapt to the accelerated expansion of artificial intelligence. Security systems represent a particularly sensitive category when it comes to the application of artificial intelligence, given the huge unknowns about the risks that these systems carry with them. A substantive framework for risk assessment in this area is imposed as necessary because the existing risk assessment methods generally do not recognize AI risks, which prevents their comprehensive overview. This paper provides a comprehensive conceptual content of the risk assessment of the application of AI in security systems, which includes risks during the entire life cycle of the AI system, which enables the identification, analysis, assessment and treatment of recognized risks and the monitoring and control of their impact on the system.

Keywords: risk assessment, AI, security system

1. Introduction

Artificial intelligence is gaining more importance in security systems considering their complexity and the growing need for these systems to respond to growing risks in their practice as soon as possible. The application of artificial intelligence in security systems facilitates the performance of many tasks and enables faster and more efficient decision-making. Application of AI in alarm systems e.g. reduced the huge number of false alarms and increased the level of their accuracy, while for security managers it significantly accelerated the process of finding the best security solutions and quickly selecting the tools needed to implement security measures. Machine and deep learning in security practice can be extremely useful, because their mechanisms are based on self-learning, which means that they are able to quickly recognize threats or incidents and transfer what they have learned to other elements, such as robots, which leads to significant savings in the long term.

The fact is that the application of AI is a growing trend in security systems and its expansion in the future is inevitable, as are the risks it brings with it, which impose consideration of the role of assessing those risks, the consequences of which may not be known now. The risks of applying AI technologies in security systems can be political, social, economic, social, technological, legal, ethical, environmental risks and many others, which means that their consequences can be extensive and require such protection measures. To see all that, it is necessary to have a precise methodology for assessing all those risks, or at least those that can be seen now so that the security system can be reliable, efficient and effective.

The aim of this paper is to conceptualize the content of the risk assessment of the application of AI in security systems, based on the existing normative basis and international and national standards, by defining the key elements of the assessment. The existence of this content enables the management of risks arising from the application of AI in security systems, provides a practical framework based on existing normative and standardized rules applicable in security practice and creates space for further research in finding the most effective solutions.

2. Concept of risk and risk assessment in using ai technology

In theory, the concept of risk has always been very complex, multidimensional and ambiguous. In modern security systems, risk means uncertainties in relation to the outcomes that may arise and may originate from different sources. Risk means "potential danger that is predictable, inherent in a situation or activity". It is the possibility of the occurrence of some future event, of uncertain or indefinite duration, which may cause loss or some other consequences (Ineris, n.d.). Risk represents the effect of uncertainty on goals (International Organization for Standardization, 2018). Risk is a concept that affects decision-making at all organizational levels, and therefore it is necessary to comprehensively understand all its elements: risk exposure, system vulnerability, probability of occurrence, damage it can cause, criticality of the system and the consequences it can leave. The entire process requires effective management of all identified risks to ensure the highest possible level of system resilience. To successfully manage risks, a key step is risk assessment, which enables the implementation of a systematic process of identifying, analyzing and evaluating risks (Institute for Standardization of Serbia, 2025), and then implementing reasonable control measures to eliminate or reduce them.

In modern security systems, the integration of artificial intelligence (AI) is becoming a standard practice and defines the processes in these systems, which implies numerous new risks that are not yet well defined or adequately identified and therefore require a methodological framework for risk assessment that will enable this. The content of the risk assessment in the use of artificial intelligence should be practical, adaptable to the development of AI and applicable to most security systems. The application of artificial intelligence in modern security systems brings specific risks such as performance failures, unintentional behavior, impact on human rights, bias, misuse of data, independent decision-making, violations of regulations and ethical problems. The management of those risks should lead to minimizing the potential negative consequences of using AI technology, while at the same time providing opportunities to increase all its positive impacts. Effective management of AI risks should lead to the creation of more reliable security systems (Mladenović et al., 2025).

The use of artificial intelligence in security systems has opened numerous legal issues, given the lack of a universal definition of AI and the possibility of autonomous system behavior, which led to the first comprehensive law regulating the use of AI - the EU AI Act (EU Artificial Intelligence Act), which was adopted in July 2024. The aim of the Act is to ensure the use of AI systems in a safe and ethical manner. This law classifies all risks in AI systems into 4 categories (European Union, 2024):

- Prohibited risks – systems that are completely prohibited:
- High-risk AI - allowed, but under strict conditions:
- Limited risk – basic transparency is required:
- Minimal risk - no special restrictions.

The EU AI Act mandated manufacturers to establish, implement, document and maintain a risk assessment system for high-risk AI systems as a process that must last throughout the system's entire lifecycle — from development, testing, commissioning, and beyond — with regular and systematic upgrades. The assessment should be regularly reviewed and updated as new information or changes in system use become available. Article 9.2 a–d defines the stages of the Risk Assessment (European Union, 2024):

- Identification and analysis: Disclosure of all known and foreseeable risks, considering intended use and potential misuse.

- Assessment and evaluation: Quantifying the probability and level of risk, by analyzing the occurrence of risk in different scenarios.
- Assessment based on post-market monitoring: Risk analysis after system commissioning, using data from post-market monitoring.
- Application of risk management measures: Defining measures to reduce/eliminate risk, in accordance with user requirements and system functionalities, which means that risk should be eliminated or reduced as much as possible through technical design and development. If a residual (remaining) risk occurs, it must be acceptable for each hazard, and information about it must be communicated to all users.

The EU AI Act represents the first step towards the responsible use of AI. It does not prohibit the use of AI but sets rules to prevent abuses.

The UNESCO Recommendations on the Ethics of Artificial Intelligence (2021) represent the first global framework that addresses ethical principles and values in the development and use of artificial intelligence. The recommendations were adopted with the aim of promoting ethical principles in AI systems: human rights, sustainability, privacy, diversity and transparency. The UNESCO recommendations on the ethics of artificial intelligence contain guidelines for assessing the ethical risks of AI systems - Ethical Impact Assessment (EIA) - which includes the following stages (UNESCO, 2021):

- Scoping and preliminary analysis
- Risk assessment
- Verification and testing in real conditions
- Monitoring during the life cycle
- Mitigation and control measures.

As part of the UNESCO recommendations, an EIA comprehensive instrument was developed aimed at assessing and managing risks before and after the implementation of AI systems. RAM (Readiness Assessment Methodology), which measures how countries are ready to implement the recommendation of the methodology and encourages public transparency, human oversight and a multidisciplinary approach, is supported (UNESCO, 2021).

Within the framework of national legislation, various normative acts were adopted with the aim of regulating the use of AI technologies within the security system. The Law on the Protection of Personal Data of the Republic of Serbia ("Official Gazette of RS", No. 87/2018) is harmonized with the General Data Protection Regulation (GDPR) of the EU and sets rules on how to collect, store, process and protect personal data of citizens and contains certain provisions related to the application of AI technologies in security systems, especially those that use automatic decision-making, profiling, biometrics or the processing of large amounts of data. According to the Personal Data Protection Act and the GDPR, a Data Protection Risk Assessment (DPIA) is mandatory before starting data processing if the AI system uses new technologies, biometric systems in public spaces, including profiling or processes large amounts of data.

The Republic of Serbia also adopted the Conclusion on the Adoption of Ethical Guidelines for the Development, Application and Use of Reliable and Responsible Artificial Intelligence ("Official Gazette of RS", No. 23/2023). The new Strategy for the Development of Artificial Intelligence in the Republic of Serbia 2025-2030, adopted by the Government on January 10, 2025, represents a continuation and specifies the institutional, legal, educational and infrastructural framework for the reliable and responsible application of AI. All this shows that great efforts are being made in the comprehensive legal regulation of the use of AI in security systems, which confirms that defining the content of the risk assessment is an inevitable step in this process.

3. Methodological basis of risk assessment: international and national standards

In security practice, different risk assessment models are applied depending on the specifics required by each security area. The application of artificial intelligence in security systems makes the changes in these systems even more dynamic, so the regulations on artificial intelligence must also reflect this phenomenon. Globally, all countries are considering how best to regulate the application of AI technologies without hindering the innovation where it is necessary. Therefore, a responsible and precise normative framework becomes even more important, because the way in which these technologies are approached will play a decisive role in shaping regulations and standards. The existence of a standardized concept of the content of risk assessment of the application of AI in security systems would significantly contribute to the regulation of this area.

Methodologies for risk assessment in security systems are most often found in different groups of ISO standards, and taking into account the sensitivity of the area itself, some assessment methods require updating and pre-definition in order to arrive at a comprehensively applicable methodology. Standards such as ISO 31000:2018 - Principles of risk management, EN 17640:22, SRPS AL.2.003:2025 regulate certain elements of risk assessment, while 23894:2023 - Artificial Intelligence - Guidance on risk management is currently the only comprehensive standard that provides basic guidelines on risk management in AI systems.

ISO 31000:2018 - Principles of risk management include guidelines for managing the risks faced by organizations. The standard provides a common approach for managing any type of risk, is used throughout the entire life cycle of an organization but is not specific to AI risks. (Institute for Standardization of Serbia, 2018) According to this standard, the risk assessment process includes:

- Risk identification,
- Risk analysis,
- Risk assessment.

Figure 1. Risk management process

Source: (Institute for Standardization of Serbia, 2018)

The standard itself allows organizations to see and assess the risks they face but does not provide specific instructions (methodology) for risk assessment in AI systems.

EN 17640:2022 - Methodology for evaluation of cyber security in a fixed time for ICT products includes the assessment of security measures of ICT products, which is carried out through various scenarios aimed at vulnerabilities, development, testing and resilience of the system (Institute for Standardization of Serbia, 2022), but its applicability is limited in relation to AI systems, because it deals with classic ICT products, does not follow the entire life cycle of AI models and does not contain a methodology applicable in the field of AI.

ISO/IEC 38507:2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations and the national counterpart SRPS ISO/IEC 38507:2023 Information technologies - IT management - Implications of the use of artificial intelligence on the management of organizations emphasize management, which is carried out by people using AI, and not on the AI systems themselves (Raković, 2024). In addition, the standards ISO/IEC 24028:2020 Security aspects and reliability of AI, ISO/IEC TR 24027:2021, Assessment of bias in AI, ISO/IEC 22989:2022 - Terminology and concepts of AI

contain different types of guidelines for risk assessment in AI systems, but do not propose a methodology for these assessments.

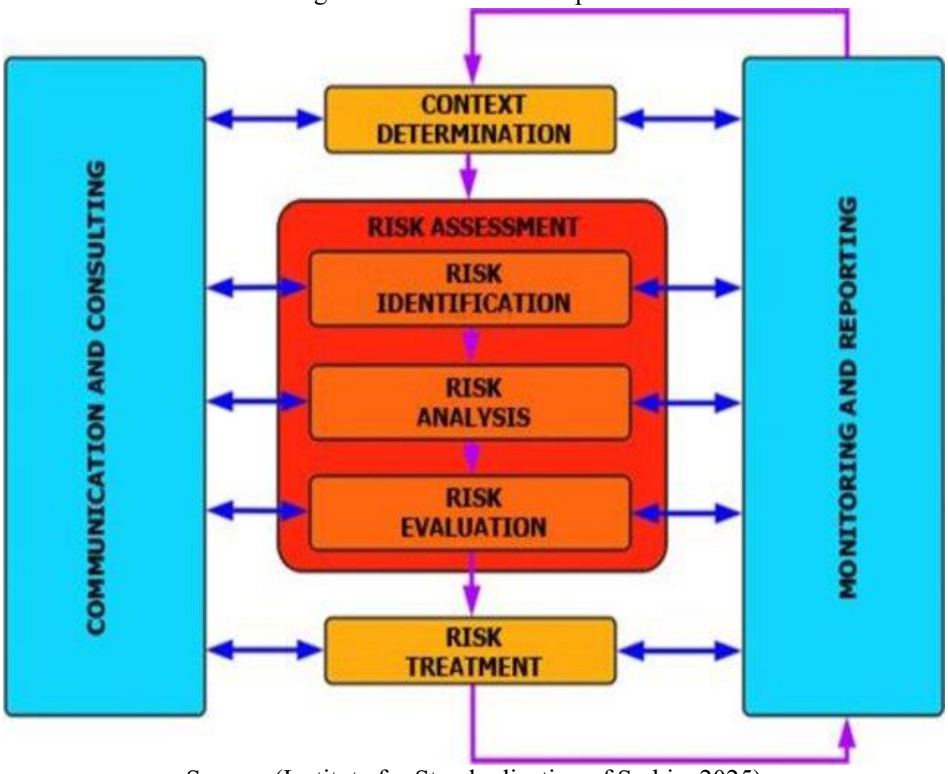
ISO/IEC 42001:2023 - "Artificial intelligence — Management system" is a standard for the management of AI systems that includes provisions on risk management as part of a mandatory policy and process, defines requirements for the assessment and treatment of risks associated with the use of AI and emphasizes the need for compliance with ethical principles, security and data protection in AI systems, but its scope in the area of risk assessment of the application of AI in security systems is significantly limited due to the high level of generality and the lack of security-specific methodological solutions.

The national standard for risk assessment in the protection of persons, property and business SRPS AL.2.003:2025 - Security and resilience - Risk assessment - Requirements and instructions for assessing compliance in the Republic of Serbia contains a methodology for assessing different types of risks in the field of security (Institute for Standardization of Serbia, 2025). According to this standard, the risk assessment process includes:

- Risk identification
- Risk analysis
- Risk assessment

SRPS A.L2.003 does not contain provisions related to the application of AI technologies or risk management in the field of AI but only addresses risks in the field of ICT infrastructure, without defining the requirements regarding artificial intelligence and its application.

Figure 2. Risk assessment process



Source: (Institute for Standardization of Serbia, 2025)

ISO/IEC 27001 Information security, cyber security and privacy protection — Information security management systems — Requirements are an international standard that regulates information security (ISMS) and defines requirements for the establishment, implementation, maintenance and continuous improvement of a security management framework aimed at protecting the confidentiality, integrity and availability of information. This standard focuses on the modernization of risk control in the context of new technological challenges, including the integration of digital services, cloud environments and automated systems such as AI tools. ISO 27001 requires systematic risk management through:

- identification of risks and vulnerabilities,

- probability and impact assessment,
- risk treatment through control and measures.

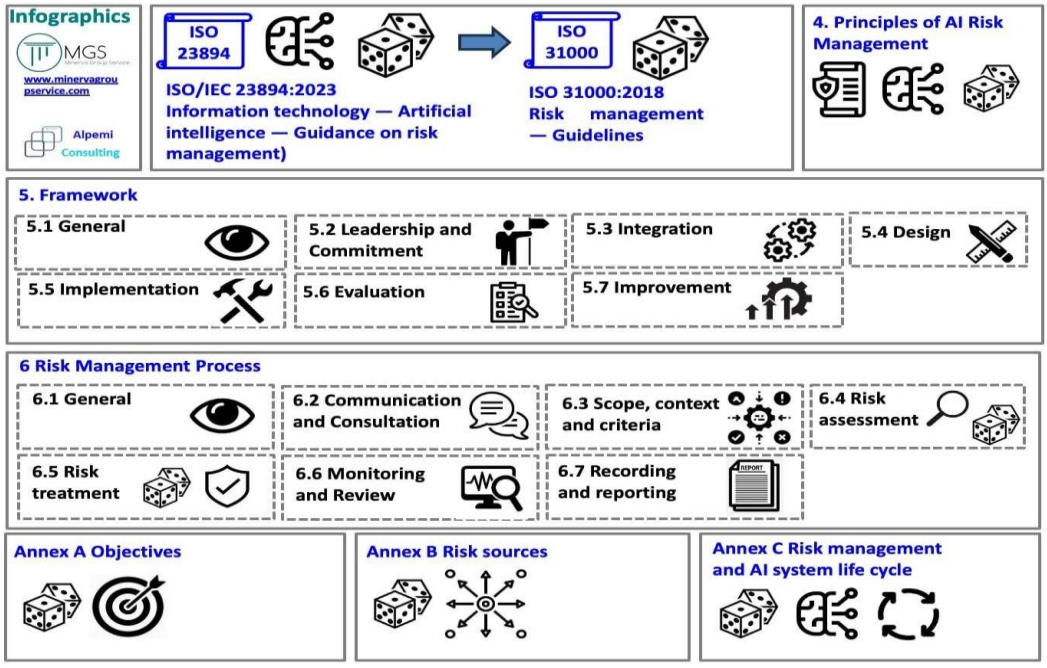
According to this standard, each organization must define risk assessment criteria and methodology that reflect their objectives and context. Risks must be continuously monitored and updated. The ISO 27001 standard does not contain explicit provisions that directly name AI risks, but the risk assessment process within an ISMS is flexible enough to include AI-specific risks (ISO, 2025).

In the last few years, due to the expansion of AI systems, international standards specific to risk assessments in this area have been adopted. ISO/IEC 23894:2023 - Artificial Intelligence - Guidance on risk management is a comprehensive standard that provides basic guidance on risk management in AI systems. The standard enables the assessment of specific AI risks during the life cycle of the AI model in compliance with prescribed ethical principles. Risk assessment according to this standard includes:

- Context of organization and system
- Risk identification
- Risk analysis
- Risk assessment
- Risk treatment
- Supervision and audit
- Communication and consultation (International Organization for Standardization, 2023)

This standard is the first international standard solely dedicated to risk assessment in AI systems and provides guidance on how organizations that develop, produce, implement or use products, systems and services that use artificial intelligence (AI) can manage risks specifically related to these systems. The guidelines also aim to help organizations integrate risk management into their AI-related activities and functions and describe processes for effective implementation and integration of AI risk management (Raković, 2024).

Figure 3. Risk management process



Source: (ISO/IEC 23894:2023)

The application of these guidelines is adaptable to any organization and its context.

On the basis of the presented standards, it can be concluded that there are a small number of developed methodologies for assessing AI risks in security systems, but that the risks that AI carries with it are multiplying daily, so that the existing regulations and standards can serve as the basis on which the methodology for assessing

risks arising as a consequence of the development of artificial intelligence, especially in security systems, will be further developed.

4. Risk assessment content of ai application in security systems

Risk assessment of the application of AI in security systems requires an expanded methodological framework compared to classic security systems, due to the specific characteristics of AI technologies, such as autonomy, adaptability and limited explainability of decisions (International Organization for Standardization [ISO], 2023). Based on the already existing normative framework and valid standards that deal with this topic, a comprehensive content of the risk assessment of the application of AI in security systems can be reached, which is applicable to any type of organization and during the entire life cycle of AI technology. The conceptualization of the content of the risk assessment is obtained by integrating the guidelines from the relevant normative frameworks and international standards, which ensures consistency and verifiability of the results. ISO/IEC 27001 prescribes mandatory elements of the risk assessment and treatment process within an information security management system (ISO, 2022), while ISO/IEC 23894 provides guidelines for adapting this process to the specifics of an AI system (ISO, 2023). In addition, the principles from the ISO 31000 standard provide a conceptual framework for the integration of risk into the wider management system of the organization (ISO, 2018), while the provisions of the EU Act and GDPR allow, among other things, the categorization of risks according to acceptability and oblige them to constantly review, control and assess. An integrated risk assessment model based on the application of multiple standards and normative acts enables a holistic approach to risk management in security systems with AI components. By combining the principles of ISO 31000, the requirements of ISO/IEC 27001 and the guidelines of ISO/IEC 23894 and other standards, organizations can identify the interdependencies between technical, organizational, ethical and other risks, thereby improving the overall resilience of security systems (ISO, 2018; ISO, 2022; ISO, 2023). This approach is in line with the recommendations of the NIST AI Risk Management Framework, which emphasizes the need for continuous and iterative AI risk management (NIST, 2023). Based on existing guidelines, a comprehensive and systematic content of risk assessment in the application of AI technologies in security systems can be reached, which can be applied in any type of organization, is independent of the size of the organization and applicable throughout the entire life cycle of AI.

Risk assessment begins with scoping/preliminary analysis and establishing a context that includes an analysis of the observed organization, leadership and commitment to risk management, the internal and external context of the organization, assignment of organizational roles, powers and responsibilities, provision and distribution of resources, establishment of communication, consultation and mechanisms for monitoring, reporting, implementation, evaluation, adaptation and continuous improvement (1).

After the initial phase of establishing the context, there is a phase of risk assessment. The risk assessment process consists of the identification, analysis and evaluation of risks. (2) Risk identification (2.1) represents the starting point of risk assessment and includes the systematic recognition of risk sources that may negatively affect the functioning of AI security systems. In addition to traditional threats to information security, such as unauthorized access and data compromise, specific AI risks are also identified, including manipulation of input data (adversarial attacks), model bias, performance degradation over time, and dependence on the quality of learning data (ISO, 2022; ISO, 2023). This approach is in accordance with the requirements of the ISO/IEC 27001 standard, which emphasizes the need to identify and document relevant risks within the information security management system (ISO, 2022). In the risk identification phase, it is necessary to identify all possible risks to make a comprehensive assessment. The list of identified risks classified according to the EU AI Act (Article 5 of the EU AI Act) can roughly be:

1. Unacceptable risks (prohibited systems)

    - AI for manipulating human behavior (e.g. subconscious)
    - Credit rating system by governments
    - Real-time biometric identification in public spaces (except in exceptional cases)

2. High-Risk AI Systems

    - AI in critical infrastructure (traffic, energy...)
    - AI in education (e.g. automatic grading)
    - AI for recruitment (e.g. automatic selection of candidates)
    - Biometric identification (under strictly controlled conditions)

3. Limited Risks (Limited Risk AI Systems)
    - Chatbot must clearly inform the user that it is not human
    - AI that generates images, video or text must indicate that the content was generated
4. Minimal risks (Minimal / Low Risk AI Systems)
    - AI in video games, spam filters...

The identified risks can also be categorized according to the stages of the life cycle of the AI system:

- Phase of concept and planning (purpose & scope)
- Phase of system design
- Phase of data collection and preparation
- Phase of model training
- Phase of testing and validation
- Phase of implementation and commissioning
- Phase of operational use
- Monitoring, maintenance and updating phase
- Phase of withdrawal and decommissioning

ISO/IEC 23894:2023 - Artificial Intelligence - Guidance on risk management forms a list of risks that can occur in AI systems. According to those guidelines, a comprehensive list of risks targeting AI systems can be made.

- Technical risks
- Data Risks
- Risks of bias and discrimination
- Risks of non-transparency and inexplicability
- Ethical risks
- Legal and regulatory risks
- Privacy and data protection risks
- Security risks
- Risks of model degradation (concept drift)
- Risks of Improper Use
- Risks of social and social influence
- Risks of Human Supervision and Interaction
- Risks of system trust and acceptability
- Risks related to interoperability and compatibility
- Risks related to the sustainability and resilience of the system.

After the identification of all risks and their detailed description, the size of the danger is determined based on an adequately chosen methodological model, which moves on to the risk analysis (2.2). In the analysis phase, the organization's exposure to those risks and its vulnerability are assessed to determine the probability of realization of the identified risks. By determining the damage and criticality, the size of the consequences for the organization is obtained, and by crossing them with the probability, the level of risk is reached. In the context of AI security systems, the consequences are not only reflected in operational or financial losses, but also in the potential violation of human security, violation of basic rights and loss of trust of stakeholders (Floridi et al., 2018). The assessment of AI risks in security systems, due to their scope and complexity, requires a combination of qualitative and quantitative assessment methods.

After the stage of obtaining the value of the risk level, there follows the stage of their evaluation, i.e. deciding on their admissibility or inadmissibility (2.3). Defining risk acceptability criteria is a key step in deciding on further risk treatment. According to risk management guidelines, acceptance criteria must be clearly defined, documented and aligned with the organization's strategy and applicable regulatory requirements (ISO, 2018). In the case of AI systems in the security domain, acceptance criteria are often more restrictive, especially when the system affects critical decision-making that may have direct consequences for the security of people, property and business (ISO, 2023).

Figure 3. AI Risk assessment

Source: Author

Assessed risk levels and defined acceptance criteria result in risks that we consider acceptable or unacceptable. All risks assessed as unacceptable by the organization require treatment to reduce their level to an acceptable level. Risk treatment includes various measures that organizations undertake in risk management, i.e. reducing the probability of their occurrence or mitigating their consequences (3). Measures can be physical protection measures, technical protection measures, normative-administrative and procedural measures, risk mitigation options, feasibility options (Institute for Standardization of Serbia, 2025). The ISO/IEC 27001 standard requires the application of appropriate technical and organizational measures, while ISO/IEC 23894 additionally emphasizes the need for specific AI measures, such as model validation, performance monitoring, decision explainability and human supervision of system operation (ISO, 2022; ISO, 2023a). These measures are the basis for building reliable and secure security systems in which AI technologies are applied.

In addition to treatment, every organization within the framework of risk assessment must foresee the ways of control and verification of applied measures, recording and reporting and constant review of risk assessment. Based on the entire assessment, within its content, there must be a final report (4) (Institute for Standardization of Serbia, 2025) in which the evaluation and presentation of the level and category of aggregate AI risk of the observed organization and analysis by risk groups with total data: level, category and acceptability can be seen. It is also necessary to draw a conclusion on the current state of protection and review new measures to improve the state to achieve maximum effectiveness and efficiency of protection.

A comprehensive AI risk assessment in security systems makes these systems efficient, effective and reliable, which is of utmost importance for their successful functioning.

5. Conslusion

Security systems in which AI technologies are applied face specific risks, which require continuous updating of risk assessment and an adaptive approach to risk management. The lack of content of this risk assessment further complicates the establishment of an effective and efficient security system. The development of risk assessment content and methodology in AI security systems is aimed at strengthening standardization, developing qualitative and quantitative reliability metrics, and integrating ethical principles into formal risk management models. This allows future regulatory and standardization frameworks to further specify the requirements for assessing and managing AI risks in the security domain. The application of multiple standards in risk assessment enables a comprehensive and systematic approach to risk management in complex security systems where AI technologies are applied. However, this approach can lead to increased complexity of implementation, the need for additional resources and the potential overlap of requirements of different standards, which requires a high level of organizational maturity and a serious approach to risk management.

The presented comparative analysis of the existing normative and standardization framework shows that risk assessment in security systems with integrated artificial intelligence components requires a systematic, structured and standardized approach that goes beyond the scope of traditional risk management models. The identification of exposure and vulnerability, the analysis of probability and consequences, as well as the definition of risk acceptance criteria, are key elements for an objective assessment of the level and category of AI risk, both at the level of the organization as a whole and by individual functional units and locations.

The results of the work confirm that the application of an integrated model of risk assessment content based on a combination of several relevant standards enables a more comprehensive overview of technical, organizational, legal and ethical aspects of AI risks. The need for continuous monitoring of AI system performance, regular review of risk assessment and application of adequate control and mitigation measures, with clearly defined responsibilities and an active role of management, was particularly emphasized. This ensures not only a reduction in unacceptable risks, but also an increase in the overall resilience and reliability of the security system.

The work indicates that effective risk management in security systems with AI components is possible only through the integration of risk assessment into the wider system of corporate security and management, with constant improvement of the process in accordance with the development of technology and the regulatory framework. The proposed approach represents a viable basis for practical application in organizations, as well as a starting point for further research in the field of standardization and quantification of AI risks in security systems.

Literature

1. European Union. (2024). EU Artificial Intelligence Act (EU 2024/865). https://artificialintelligenceact.eu/ai-act-explorer/
2. Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society. Minds and Machines, 28(4), 689–707.
3. Ineris. (n.d.). How can risk be defined? Ineris. https://www.ineris.fr/en/risks/what-risk/how-can-risk-be-defined
4. Institute for Standardization of Serbia. (2018). ISO 31000:2018 Risk management - Guidelines. Belgrade: ISS.
5. Institute for Standardization of Serbia. (2022). SRPS EN 17640:2022 - Security - Risk assessment of security measures. Belgrade: ISS.
6. Institute for Standardization of Serbia. (2025). SRPS A.L2.003 – Safety and resilience – Risk assessment – Requirements and guidance for conformity assessment (III edition). ISS.
7. International Organization for Standardization & International Electrotechnical Commission. (2025). ISO/IEC 27001:2025 Information security, cybersecurity and privacy protection — Information security management systems — Requirements. ISO.
8. International Organization for Standardization. (2018). ISO 31000:2018 – Risk management — Guidelines. ISO.
9. International Organization for Standardization. (2023). ISO/IEC 23894:2023 – Information technology – Artificial intelligence – Guidance on risk management. ISO.
10. Mladenović, M., Janković, K., & Komazec, N. (2025). Risk assessment for AI applications in security systems: Challenges and opportunities. In 11th Scientific-Professional Conference Security and Crisis Management – Theory and Practice (SeCMan), Belgrade. https://doi.org/10.70995/YMSH5950

11. National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). NIST.
12. Raković, R. (2024). Artificial intelligence and ISO standards [Artificial intelligence and ISO standards]. Military Information Bulletin, 27(1), 19–29. https://doi.org/10.5937/VI24019R
13. Republic of Serbia. (2018). Law on Protection of Personal Data ("Official Gazette of RS", No. 87/2018).
14. UNESCO Recommendation on the Ethics of AI - https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Original Scientific Paper/Original naučni rad UDC/UDK: 005.334:[004.8:623
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601037J

Rizici i upravljanje autonomnim oružjem u savremenom ratovanju: Sveobuhvatna analiza

Katarina Janković[1], Milica Mladenović[2], Nenad Komazec[3]

[1] General Staff of the Serbian Army, Directorate for Development and Equipment J-5, Technical Test Centre, Centre for Testing Armaments and Military Equipment, Nikinci, Republic of Serbia, jankovickatarina95@gmail.com
[2] Regional Association for Security and Crisis Management RABEK, Belgrade, Serbia, mladenovicmilica21@yahoo.com
[3] Military Academy, University of Defence, Belgrade, Serbia, nenadkomazec@yahoo.com

Rezime: Brz tehnološki napredak u oblasti veštačke inteligencije (VI) menja savremene vojne sposobnosti i uvodi autonomno oružje kao važno strateško i etičko pitanje. U radu se analiziraju rizici koji prate uvođenje autonomnog oružja u savremene sukobe, kao i njegov tehnološki potencijal i bezbednosni izazovi. Razvoj i širenje autonomnog oružja predstavlja složenu tehnološku inovaciju koja prevazilazi uobičajene obrasce vojnog angažovanja. Napredne VI tehnologije menjaju međunarodne bezbednosne okvire i otvaraju etičke i operativne dileme bez presedana. U radu se sprovodi analiza rizika koja obuhvata tehnološke, geopolitičke, pravne i etičke aspekte primene autonomnog oružja. Metodologija istraživanja obuhvata proces identifikacije, procene i upravljanja rizicima. Kroz analizu relevantne literature, konsultacije sa ekspertima i modeliranje različitih scenarija, razmatraju se rizici poput algoritamske pristrasnosti, nenamerene eskalacije sukoba, izazova u određivanju odgovornosti i mogućnosti da sistemi deluju van ljudske kontrole. Rezultati pokazuju da autonomno oružje pruža određene taktičke prednosti, ali istovremeno stvara ozbiljne izazove za globalnu bezbednost. Nalazi ukazuju da međunarodna zajednica mora da razvije snažne mehanizme upravljanja i efikasne mere kontrole rizika. Završni deo rada predlaže okvir za smanjenje rizika koji podrazumeva saradnju stručnjaka za tehnologiju, donosioca odluka, vojnih planera i etičara. Ovakav pristup doprinosi aktuelnoj raspravi o odgovornoj primeni veštačke inteligencije u vojnim sistemima i naglašava potrebu za proaktivnim upravljanjem rizicima koje usklađuje tehnološke inovacije, etičke principe i zahteve globalne bezbednosti.

Keywords: Rizik, Upravljanje rizicima, Autnomno oružje, Veštačka inteligencija, Vojne tehnologije, Globalna bezbednost

Risks And Management of Autonomous Weapons In Contemporary Warfare: A Comprehensive Analysis

Abstract: The rapid technological advancement in artificial intelligence (AI) has precipitated a paradigm shift in military capabilities, with autonomous weapons emerging as a critical domain of strategic and ethical concern. This research critically examines the multifaceted risks associated with the integration of autonomous weapons systems into military conflict landscapes, exploring their technological potential and inherent security challenges. The proliferation of autonomous weapons represents a complex technological innovation that transcends traditional military engagement strategies. By leveraging advanced AI technologies, these systems challenge established international security frameworks and introduce unprecedented ethical and operational uncertainties. This study conducts a comprehensive risk analysis that encompasses technological, strategic, legal, and humanitarian dimensions of autonomous weapon deployment. The research methodology employs a systematic approach to risk identification, assessment, and management. Through comprehensive literature review, expert consultations, and scenario modelling, the study investigates potential risks such as algorithmic bias, unintended escalation, accountability challenges, and the potential for autonomous systems to operate beyond human control. The analysis reveals that while autonomous weapons offer significant tactical advantages, they simultaneously introduce substantial risks to global security architectures. Key findings underscore the critical need for robust international governance mechanisms and comprehensive risk management strategies. The research proposes a multi-stakeholder framework for mitigating autonomous weapon risks, emphasizing the importance of

interdisciplinary collaboration among technologists, policymakers, military strategists, and ethicists. By providing a nuanced understanding of autonomous weapon risks, this study contributes to the emerging discourse on responsible AI development in military contexts. The findings advocate for proactive risk management approaches that balance technological innovation with ethical considerations and global security imperatives.

Keywords: Risk, Risk Management, Autonomous Weapons, Artificial Intelligence, Military Technology, Global Security

1. Introduction

The rapid development of artificial intelligence is transforming the way modern armed forces plan, decide, and conduct military operations. This process introduces a new generation of combat systems in which autonomous weapons hold a central position. Autonomous systems provide a high level of automation, fast reaction times, and the ability to operate in dynamic environments, while simultaneously creating significant security, ethical, and geopolitical challenges. Contemporary armed conflicts demonstrate the growing use of algorithmically supported systems that identify targets, make decisions, and execute tasks with minimal or no human supervision. Examples from recent conflicts show that artificial intelligence accelerates combat activities but also increases the risk of unintended escalation and target misidentification (Jankovic, Mladenovic & Komazec, 2025). Autonomous weapons thus become a source of tactical advantage, as well as a potential cause of consequences that are difficult to predict and control.

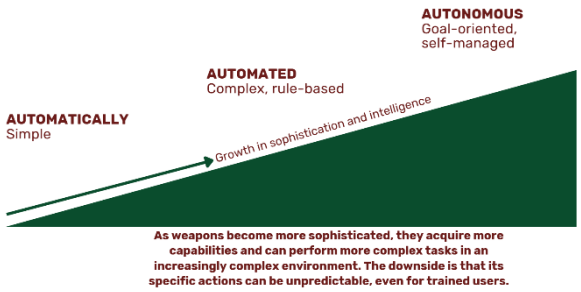
The risks associated with these systems extend beyond traditional military frameworks. Algorithmic bias, errors in pattern recognition, cyber intrusions, and loss of control can lead to incidents with serious consequences for military units, civilian populations, and international stability. The issue of accountability becomes particularly prominent when autonomous systems make decisions faster than humans can intervene.

In such an environment, the need for a systematic approach to managing the risks of autonomous weapons continues to grow. An analysis of technological, legal, geopolitical, and ethical factors enables a clearer understanding of the capabilities and limitations of these systems. The development of risk-management models provides the foundation for their responsible, controlled, and safe use. The aim of the paper is to offer a comprehensive overview of the risks posed by autonomous weapons and to propose a framework for their responsible application in modern warfare.

2. Characteristics of Autonomous Weapons

Modern weapons systems are undergoing significant technological transformation, and the development of autonomous weapons represents one of the fastest-growing innovations. Although the term autonomous weapons is often used as if it has a clear definition, in practice it involves considerable ambiguity. The greatest confusion arises from the distinctions between automatic, automated, and autonomous weapons, which complicates the understanding of this new generation of systems based on artificial intelligence (Figure 1).

Figure 1: Automatic, automated and autonomous weapons



Source: (Scharee, 2020)

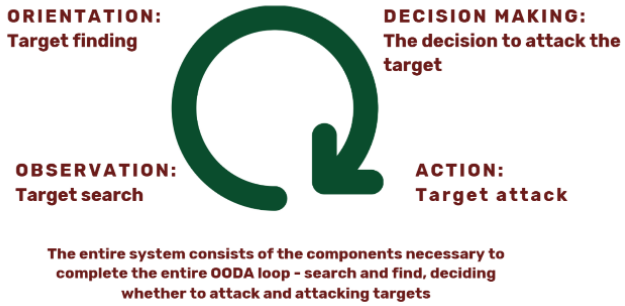
To understand what autonomous weapons represent, it is necessary to distinguish between the basic terms used for modern weapons systems (Janković, Komazec & Erkic, 2023). Three concepts appear in current use—automatic, automated, and autonomous weapons—and the boundaries between them are often fluid because these technologies overlap.

Automatic systems operate in a simple manner and do not include any decision-making process. They react only to the physical act of pulling the trigger and rely on basic mechanical principles. Automated systems represent a more advanced level: they process a greater volume of input data, combine multiple variables, and select a response based on predefined rules.

Autonomous weapons stand above these two levels. They use artificial intelligence algorithms that enable the system to perform tasks independently, such as sensing, locating targets, selecting targets, and executing attacks (Scharee, 2020) (Figure 2). This level of capability surpasses classical automation and introduces decision-making with or without direct human supervision.

Autonomy refers to the ability of a machine to perform a task independently. Although there is no internationally agreed definition of autonomous weapons, one of the definitions used in the working groups of the International Committee of the Red Cross states: „Autonomous weapon systems are weapons that can select and attack targets independently of human intervention, that is, with autonomy in their 'critical functions' of acquiring, tracking, selecting, and attacking targets" (International Committee of the Red Cross, 2014).

Figure 2: Components of Autonomous Weapons



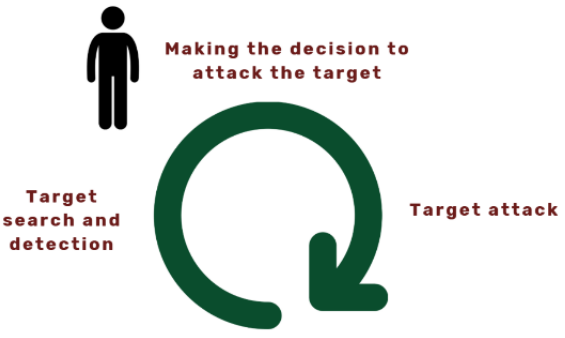
Source: (Scharee, 2020)

Autonomous weapons integrate several technologies into a unified system: sensors, data-processing units, pattern-recognition modules, decision-making algorithms, and executive subsystems for delivering force. Artificial intelligence allows these components to analyze the situation, adjust their reactions, and operate effectively in fast and highly dynamic environments.

Figure 3: Human-supervised autonomous weapons

Source: (Scharee, 2020)

Figure 4: Human-supervised autonomous weapons



Source: (Scharee, 2020)

According to the level of human involvement in decision-making, autonomous weapons fall into two categories (Scharee, 2020):

- Human-supervised autonomous weapons, where the operator remains "in the loop" and makes the final decision to attack (Figure 3).

- Fully autonomous weapons, where the entire engagement cycle—target search, detection, decision, and attack—unfolds without human intervention (Figure 4).

This distinction plays an important role in assessing the tactical value of autonomous weapons, as well as in identifying the risks they create. The level of autonomy determines the speed of decision-making, the possibility of human reaction, and the degree of control that armed forces can maintain over the system.

Understanding these characteristics forms the basis for correctly interpreting the capabilities and limitations of autonomous weapons in modern warfare. The combination of sensor technologies, data processing, and artificial intelligence algorithms reshapes the way weapons systems perceive the environment, make decisions, and execute tasks. For this reason, autonomous weapons have become a central topic in the analysis of modern military technologies and one of the most sensitive issues in evaluating their security, ethical, and tactical implications. Further analysis must examine the risks these systems generate and the conditions under which they can be used in a responsible and controlled manner.

3. Risks of Autonomous Weapons

Autonomous weapons introduce a new generation of military systems based on artificial intelligence and open a broad spectrum of technological, geopolitical, ethical, legal, and economic risks. These systems provide advanced capabilities but simultaneously create consequences that armed forces and international institutions struggle to predict and control. Their risks extend beyond traditional military considerations and include political, social, and economic implications. Autonomous systems operate through algorithms for pattern recognition, decision-making, and the execution of attacks, which makes their risks more complex than those of conventional weapons. Based on relevant literature and available research, this study identifies groups of risks that best illustrate the key challenges associated with autonomous weapons. These categories do not represent an exhaustive list, but they serve as the most significant groups for analysis within the scope of this paper.

Technological risks of autonomous weapons arise from their dependence on artificial intelligence algorithms, digital subsystems, and complex communication networks. A review of available literature highlights several key technological risks, although this list does not encompass all potential challenges. The most commonly identified risks include:

1. **Cyberattacks** – adversaries can manipulate algorithms, sensors, or communication channels, causing incorrect target identification, redirecting attacks, or disabling the system.
2. **Algorithmic errors and bias** – autonomous weapons depend on data quality and the reliability of AI models. Poor, incomplete, or misinterpreted data can lead to incorrect decisions.
3. **Technical failures and operational errors** – failures in sensors, guidance systems, or decision-making modules can trigger unintended reactions or result in a loss of functionality at critical moments.
4. **Loss of control** – the speed of autonomous decision-making can exceed the operator's ability to intervene, increasing the risk of unintended consequences and escalation.
5. **Dependence on complex AI architectures** – software vulnerabilities, attacks on algorithms, or coding errors can lead to unpredictable behavior or complete system failure.

Geopolitical risks of autonomous weapons arise from their potential to shift power balances, accelerate arms races, and increase uncertainty in international relations (Meiches, 2017). The use of AI-enabled autonomous systems affects political stability, security architectures, and the dynamics of conflicts between state and non-state actors. An analysis of relevant literature most commonly highlights the following geopolitical risks:

1. **Arms race** – the development of autonomous weapons motivates states to rapidly invest in new AI-driven military capabilities, which increases global tensions and reduces opportunities for diplomatic conflict resolution.
2. **Access by non-state actors** – increasingly accessible AI technologies raise the likelihood that autonomous systems will reach terrorist organizations or criminal groups, potentially destabilizing entire regions.
3. **Proliferation of technology** – the transfer of autonomous weapons to states with weak institutional capacities or to parties engaged in conflict elevates the risk of regional instability and violent escalation.

4. **Escalation of autonomous conflicts** – interactions between multiple autonomous systems on the battlefield can create rapid and uncontrolled conflict dynamics, as AI systems make decisions much faster than humans can respond.
5. **New competition in space and cyberspace** – autonomous systems used for surveillance, satellite reconnaissance, and cyber operations introduce new arenas of confrontation, with consequences that the global security framework struggles to anticipate.

Ethical and legal risks of autonomous weapons arise from the fact that algorithms take over decision-making processes that previously belonged exclusively to human actors. These risks include a range of dilemmas related to the protection of civilians, compliance with international humanitarian law, and the question of responsibility for decisions made by the system. An analysis of relevant literature most commonly highlights the following ethical and legal challenges:

1. **Collateral damage** – autonomous systems may misidentify targets due to algorithmic errors or insufficiently reliable data. This increases the risk of civilian casualties and the destruction of protected objects (Filipovic, 2023).
2. **Responsibility and accountability** – autonomous weapons raise the dilemma of who is accountable for the system's decisions: the commander, the programmer, the state, or the system itself. Current legal frameworks do not offer a clear answer to this issue.
3. **Limitations within international humanitarian law** – rapid technological development and high levels of automation exceed existing norms of international humanitarian law, creating legal gaps in regulating target selection and the use of force.
4. **Privacy intrusion and surveillance** – autonomous tracking and recognition systems may endanger individual privacy and raise ethical questions about the limits of surveillance in civilian and military environments.
5. **Information manipulation and disinformation** – autonomous systems, especially those relying on large datasets, can become tools for propaganda, psychological operations, or the dissemination of misleading content

Economic risks of autonomous weapons arise from the high costs of developing, producing, maintaining, and modernizing systems based on artificial intelligence. The use of these systems affects national budgets, international economic flows, and the long-term stability of states that invest in AI-driven military capabilities. Relevant literature most commonly highlights the following economic challenges:

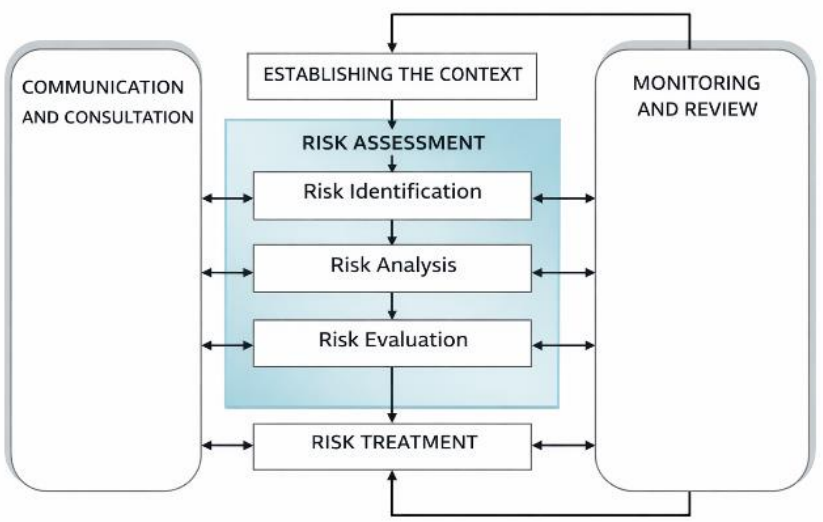
1. **High development and maintenance costs** – the creation and modernization of autonomous weapons require substantial financial investment in research, software modules, sensor equipment, and advanced communication systems, which can place significant pressure on military and state budgets.
2. **Impact on national and global economies** – large allocations for autonomous weapons may reduce funding for key public sectors such as healthcare, education, and infrastructure, affecting long-term economic development.
3. **Economic inequality among states** – wealthier states can invest in advanced AI-enabled military capabilities, while less developed states remain technologically inferior, deepening global security and economic disparities (Center for Strategic and International Studies, 2015).
4. **Market destabilization** – increased demand for AI-based military systems can fuel rapid and unregulated growth in the global arms market, creating opportunities for uncontrolled technology flows and illegal trade.
5. **Humanitarian and societal costs** – economic risks also include the long-term costs of post-conflict recovery, such as displacement of populations, health consequences, and the reconstruction of affected areas.

The analysis of technological, geopolitical, ethical, legal, and economic risks demonstrates that autonomous weapons fundamentally reshape the nature of modern warfare and the broader global security landscape. Although these systems offer significant operational advantages, their reliance on artificial intelligence creates layers of uncertainty that extend far beyond traditional military concerns. The risks identified in this chapter highlight the need for continuous assessment, transparent governance, and clearly defined international norms to prevent unintended escalation, misuse, or loss of control. Addressing these challenges requires coordinated action among states, international organizations, and the scientific community to ensure that the development and deployment of autonomous weapons remain aligned with principles of security, responsibility, and humanitarian protection.

4. Risk Management Model for Autonomous Weapons

The risk management process, defined by ISO standards, represents a systematic and interactive approach that aims to ensure effective and responsible management of risks associated with autonomous weapons (Jankovic & Komazec, 2024). These standards outline key steps that form the foundation of a proactive methodology in risk management. Figure 6 illustrates the risk management process and helps clarify how the different phases connect and contribute to the overall objective of managing risks.

Figure 6: Components of Autonomous Weapons



Source: (ISO 31000:2018)

The risk management process for autonomous weapons continually adapts to changes in technological development, political and social conditions, and the framework of international law. According to ISO 31000, the risk management process for autonomous weapons consists of five key phases:

Communication and Consultation: The process begins with open and continuous communication with all relevant stakeholders (Jankovic & Komazec, 2024; ISO 3100:2018). In the context of autonomous weapons, this phase involves engaging experts from military, political, technological, and other sectors, as well as representatives of governmental and international institutions, to ensure diverse perspectives and accurate information.

Establishing the Context: Clearly defining the external and internal parameters in which autonomous weapons are developed and used is essential for identifying potential risks. This includes considering political, legal, technological, and operational conditions that shape the environment of autonomous weapon deployment.

Risk Assessment: This phase includes Risk Identification, Risk Analysis, and Risk Evaluation. The assessment of risks related to autonomous weapons considers technological, geopolitical, ethical, legal, and economic risks, along with the probability of occurrence, the severity of consequences, and the overall level of risk.

Risk Treatment: In this phase, measures are implemented to manage, reduce, or eliminate identified risks. For autonomous weapons, this includes developing strategies that mitigate technological, geopolitical, ethical, legal, and economic risks. Each risk category requires specific treatment strategies (Table 1).

Monitoring and Review: Monitoring and review in the risk management process for autonomous weapons represent two essential components that ensure the overall effectiveness of the system (Jankovic & Komazec, 2024). Monitoring involves continuous observation and evaluation of the implemented risk management measures, including tracking technological developments, political conditions, and emerging threats. Review refers to the periodic re-examination of the entire risk management framework, which enables the identification of necessary improvements and alignment with new standards, regulations, and changes in the operational environment. This phase also includes assessing the effectiveness of autonomous weapons control measures and compliance with international agreements.

Table 1: Example Strategies for Managing Autonomous Weapon Risks

Risk Group	Example Strategies for Managing Autonomous Weapon Risks
Technological	- Enhance cybersecurity measures and conduct continuous threat monitoring. - Test algorithms, sensors, and decision-making modules before deployment. - Perform regular maintenance and AI model updates to reduce errors. - Implement double-check protocols before system activation. - Collaborate with technological institutions to improve AI robustness.
Geopolitical	- Strengthen diplomatic mechanisms for autonomous weapons control. - Participate in international agreements and oversight regimes. - Support intelligence-sharing to prevent misuse of autonomous systems. - Invest in confidence-building and transparency measures among states.
Ethical and Legal	- Define clear accountability norms for autonomous decision-making. - Align autonomous weapon development with international humanitarian law. - Establish oversight mechanisms for real-world deployment. - Minimize collateral damage and ensure civilian protection. - Increase transparency in development and operational use.
Economic	- Conduct cost-benefit analyses for AI military system adoption. - Plan budgets to avoid over-dependence on expensive AI systems. - Invest in post-conflict recovery programs. - Develop alternative economic sectors to reduce reliance on the defense industry.

Source: (author's elaboration based on the analysis of relevant literature)

The ISO 31000 risk management model for autonomous weapons supports their safe and responsible use. Each phase — from Communication and Consultation to Monitoring and Review — helps identify, analyze, and mitigate the risks associated with the development and deployment of autonomous weapons. Implementing targeted strategies for managing technological, geopolitical, ethical, legal, and economic risks provides the foundation for a sustainable and responsible approach in this field.

5. Conclusion

The analysis presented in this paper demonstrates that the integration of autonomous weapons into contemporary warfare introduces serious challenges that cannot be viewed solely through a technological lens. Autonomous systems, driven by artificial intelligence algorithms, generate risks far more complex than those associated with conventional weapons, as they introduce the possibility of loss of control, legally undefined responsibility, and the potential for unintended conflict escalation. At the same time, the geopolitical context and the ambiguity of international regulations further complicate predictable management of these systems. These risks require systematic understanding and governance, which this study enables through the identification of the most significant categories: technological, geopolitical, ethical, legal, and economic. Establishing such a spectrum of risks allows for a comprehensive assessment of the impact of autonomous weapons on security, civilian protection, and the stability of international relations. Additionally, this classification provides a foundation for practical strategies of control and risk mitigation. The ISO 31000 model, applied to autonomous weapons, offers a systematic framework that integrates all phases of risk management—from communication to monitoring and review. Continuous observation, system auditing, and revising measures in accordance with new insights ensure that risk management remains precise, effective, and aligned with relevant standards.

In conclusion, autonomous weapons cannot be treated merely as technological assets. The implementation of coherent and comprehensive risk-management strategies is a prerequisite for their responsible and safe use. Only through the integration of technical, legal, and ethical control mechanisms can the development and deployment of autonomous weapons be aligned with the principles of international law, humanitarian norms, and global security.

Literature

1. Center for Strategic and International Studies. (2025). Lessons from the Ukraine conflict: Modern warfare in the age of autonomy, information, and resilience. CSIS. Pristupljeno 25.12.2025. u 17:36 https://www.csis.org/analysis/lessons-ukraine-conflict-modern-warfare-age-autonomy-information-and-resilience
2. Filipovic, A. (2023). *Lethal Autonomous Weapon Systems (LAWS) – Towards Global Regulation or Indiscriminate Employment?* Political Review, No. 01/2023, Vol. XXXI(XXIII), p. 75.
3. International Committee of the Red Cross (ICRC). (2014). *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*. Expert Meeting, Geneva, Switzerland, 26–28 March 2014.
4. Janković, K., Komazec, N. & Mladenovic, M. (2025). *Dual Implications of Modern Armament in Contemporary Conflicts Through the Lens of Swot Analysis and The Ahp Method*. In Proceedings of the 11th International Scientific-Professional Conference "Security and Crisis Management – Theory and Practice (SeCMan)" (p. 253). RASEC & S4 Glosec Global Security. ISBN 978-86-80692-12-8
5. Janković, K., & Komazec, N. (2024). *Implications of modern weapons development*. In Proceedings of the 10th International Scientific-Professional Conference "Security and Crisis Management – Theory and Practice (SeCMan*)"* (p. 320). RASEC & S4 Glosec Global Security. ISBN 978-86-80692-11-1
6. Janković, K., & Komazec, N. (2024). *Upravljanje rizicima savremenog oružja*. In Zbornik radova sa Međunarodnog naučnog skupa „Savremeni izazovi i prijetnje bezbjednosti "(p. 161). Univerzitet u Banjoj Luci, Fakultet bezbjednosnih nauka. ISBN 978-99976-805-4-9.
7. Janković, K., Komazec, N., & Erkic, D. (2023). *Review of the risks of autonomous weapons*. In Proceedings of the 9th International Scientific-Professional Conference – Security and Crisis Management: Theory and Practice (SeCMan) (p. 54). Regional Association for Security and Crisis Management – RASEC, S4 GLOSEC Global Security. ISBN 978-86-80692-10-4.
8. Jončić, V. (2015). *International Humanitarian Law*. Faculty of Law, Belgrade.
9. Meiches, B. (2017). *Weapons, desire, and the making of war.* Critical Studies on Security, 5, 27-29. doi:10.1080/21624887.2017.1312149
10. Scharee, P. (2020). Vojska bez vojnika (Original work published as Army of None: Autonomous Weapons and the Future of War). Beograd: Laguna.
11. International Organization for Standardization. (2018). ISO 31000:2018 Risk management — Guidelines. ISO.

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:355/359]:17
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601045R

Upravljanje primenom veštačke inteligencije u vojne svrhe

Vanja Roković[1]

[1] Faculty of Security Studies, University of Belgrade, Serbia, vanjarokvic@fb.bg.ac.rs

Sažetak: Veštačka inteligencija (VI) suštinski menja karakter savremenog ratovanja. Iako nudi značajne prednosti, poput veće preciznosti, operativne efikasnosti, bržeg donošenja odluka i boljeg potencijala za zaštitu civila, ona istovremeno pokreće ozbiljna pitanja odgovornosti, smislene ljudske kontrole i rizika od eskalacije sukoba. Poseban izazov predstavljaju mogućnosti zloupotrebe i neželjene posledice u složenim i asimetričnim operativnim okruženjima. Shodno tome, regulisanje upotrebe veštačke inteligencije u vojne svrhe postalo je prioritetno pitanje na međunarodnoj agendi. U ovom radu razmatraju se aktuelni napori za normativno uređenje ove oblasti na globalnom, regionalnom i nacionalnom nivou. Poseban osvrt dat je na inicijative u okviru Ujedinjenih nacija, stavove ključnih međunarodnih aktera, kao i na nove regulatorne pristupe u Evropskoj uniji, NATO-u i odabranim državama. Okosnicu ovih nastojanja čini rasprava o smrtonosnim autonomnim oružanim sistemima (LAWS), koja oslikava dublje podele u pogledu uloge autonomije pri upotrebi sile i primene međunarodnog humanitarnog prava. Uprkos napretku u normativnim i političkim debatama, još uvek ne postoji pravno obavezujući međunarodni akt specifično posvećen vojnoj primeni veštačke inteligencije. Regulatorni okvir je stoga fragmentisan i uslovljen različitim nacionalnim prioritetima, tehnološkim kapacitetima i pravnim tumačenjima, što ukazuje na to da će buduće upravljanje ovom sferom obeležiti kontinuirane rasprave i suprotstavljeni interesi.

Keywords: veštačka inteligencija, primena veštačke inteligencije u vojne svrhe, smrtonosni autonomni oružani sistemi, međunarodno humanitarno pravo, upravljanje

Governing the Use of Artificial Intelligence for Military Purposess

Abstract: Artificial intelligence (AI) is reshaping the character of modern warfare. While AI offers advantages such as improved precision, efficiency, decision-making, and civilian protection, it also raises concerns related to accountability, meaningful human control, escalation, misuse, and unintended consequences, particularly in asymmetric and complex environments. In response, the regulation of military AI has become an increasingly prominent issue on the international agenda. This paper examines current efforts to regulate the military use of AI at the global, regional, and national levels, with particular attention to recent initiatives within the United Nations (UN), the positions of international actors, and emerging approaches within the European Union (EU), NATO, and individual states. A central theme across these efforts is the debate on lethal autonomous weapons systems (LAWS), which reflects broader divisions over the role of autonomy in the use of force and the application of international humanitarian law. Despite notable progress, no binding international instrument dedicated explicitly to regulating military AI currently exists. The regulatory landscape, therefore, remains fragmented and shaped by differing national priorities, technological capabilities, and legal interpretations, indicating that future governance will continue to evolve amid ongoing discussions and competing interests.

Keywords: artificial intelligence, military applications of artificial intelligence, lethal autonomous weapons systems (LAWS), international humanitarian law, governance

1. Introduction

In defence studies, it is often emphasized that the capacity to understand and anticipate the future of warfare, particularly the technologies that will shape it, is usually regarded as a central analytical ambition of the discipline (Franke, 2018). The technologies associated with artificial intelligence have been developing for decades; however, a key question remains: did scholars and practitioners truly foresee the depth of the transformation AI would bring to the battlefield? Contemporary conflicts, such as those in Ukraine and Gaza, are increasingly

described as „AI-driven“ or „algorithmic“ wars. From data analytics and surveillance to autonomous weapon systems, AI is reshaping the very character of modern warfare (Wells, 2025; Hibbert & Overton, 2025). At the same time, these advancements raise serious concerns, including the risk of a global AI arms race, the erosion of meaningful human control over critical decisions, and the ethical implications of delegating life and death judgments to machines (Asaro, 2012; Amorso & Tamburini, 2020; Kohn et al., 2024; Jensen, Atalan & Macias, 2024; Osimen, Newo & Fulani, 2024; Johnson, 2020).

In analysing the development and impact of technology on the character of warfare, van Creveld (1989) employs a simple yet powerful analogy of the ripples caused by a stone thrown into water. He argues that technological effects are most pronounced at the point of impact, while the expanding ripples gradually weaken and become less visible. The farther they travel, the more likely they are to lose their distinct identity, intermingling with waves generated by other stones or reflected from water’s edges. Similarly, weapons and weapon systems exert their most significant influence in direct combat, but war encompasses far more than the battlefield. It also includes tactics, strategy, logistics, operations, command and control, and numerous additional factors.

Applying this analogy to the development and military employment of AI reveals clear parallels. The most potent effects of AI can be observed at the tactical level, through autonomous platforms, AI-driven targeting systems, and rapid situational analysis (Baxter, 2024). As AI diffuses into the operational and strategic domains, its influence becomes broader yet less immediately perceptible, shaping the preparations and conduct of military operations (Vestner, 2024; Burton & Soare, 2019). At the broadest level, AI becomes embedded within wider structures of doctrine, deterrence, the emerging global AI arms race, and ongoing debates on arms-control regimes (Jensen, Atalan & Macias, 2024; Osimen, Newo & Fulani, 2024; Johnson, 2020). Given this diffusion and interdependence of effects, the demand for clear international regulations and norms governing the military use of AI is becoming increasingly urgent.

In the context of the development and use of AI tools for military purposes, it is essential to note that they differ significantly in the level of autonomy involved in decision-making regarding the use of force. Consequently, the question of regulating these tools also varies substantially. A critical starting point for understanding governance challenges lies in distinguishing between AI systems that support human decision-making and those that are capable of operating independently in the use of force. These categories differ not only technologically but also normatively, as they imply fundamentally different risks and regulatory expectations. For example, AI systems designed to process vast amounts of data and support human decision-making, such as the U.S. Project Maven, are widely accepted and implemented in practice when proper oversight is in place (Zequeira, 2024; Nguyen, 2025).

In contrast, systems intended to operate independently, namely lethal autonomous weapon systems (LAWS), raise numerous legal and ethical dilemmas. While there is a prevailing belief that such technologies enhance military precision and efficiency and reduce casualties (Petman, 2017; Vergun, 2019; Zequeira, 2024; Márquez-Díaz, 2024), critics argue that delegating life-and-death decisions to autonomous systems undermines human dignity, creates accountability gaps, and raises profound moral objections. Moreover, the complexity of modern warfare casts doubt on whether these systems can reliably uphold core principles of international humanitarian law (IHL), prompting calls for legally binding restrictions or prohibitions on weapons that operate without meaningful human control (Heyns, 2016; Human Rights Watch & International Human Rights Clinic, 2012; Davison, 2018; Geneva Academy of International Humanitarian Law and Human Rights, 2014; United Nations Office for Disarmament Affairs, 2023).

This paper examines current efforts to regulate the military use of artificial intelligence across global, regional, and national levels. It identifies the main challenges arising from the development and deployment of autonomous systems, with particular attention to the debate surrounding LAWS. Methodologically, this study applies a qualitative, desk-based document analysis, examining relevant primary and secondary sources. The review includes United Nations (UN) resolutions, declarations, and guidelines; strategic and regulatory acts of the European Union (EU) and NATO; national strategies and policy papers of selected states; and publications of international organizations. Regarding national-level approaches, the paper focuses on the United States, Russia, and China as illustrative case studies. These states are selected due to their status as major military powers, their advanced technological capabilities, and the availability of publicly published strategies and policy documents addressing the military use of artificial intelligence. This selection enables a transparent, document-based analysis and does not aim to provide an exhaustive comparative assessment, but rather to illustrate divergent strategic and regulatory approaches that shape the emerging international governance landscape. In addition, the analysis incorporates relevant scientific and expert literature. The analysis aims to map current regulatory efforts concerning the military application of AI and to identify the principal challenges associated with their

development and implementation. The study is divided into several sections, including an overview of global initiatives, regional approaches, national strategies, and a focused discussion of LAWS regulation. This structure provides a clearer understanding of how regulatory developments unfold across different levels.

2. Global Governance of Military AI

Although a specific international treaty exclusively dedicated to the development and use of AI for military purposes has yet to be adopted, existing international regulations, primarily IHL, provide essential and binding legal frameworks. According to IHL, all parties are required to adhere to its principles in the development and use of any weapon, including systems incorporating AI components. The fundamental principles of IHL, distinction, proportionality, and precaution in attack, must be upheld even when AI technologies are employed (Bruun & Bo, 2025).

Beyond these principles of IHL, a significant milestone in the global regulation of military AI occurred in December 2024, when the UN General Assembly adopted Resolution A/RES/79/239, titled „Artificial Intelligence in the Military Domain and its Implications for International Peace and Security“ (United Nations General Assembly, 2024). This resolution constitutes a significant international legal instrument that acknowledges the growing challenges and opportunities posed by AI in the military sphere. It reaffirms that the UN Charter, IHL, and international human rights law (IHRL) apply to all stages of the lifecycle of AI-enabled military systems, from pre-design and development to testing, deployment, and decommissioning. The resolution emphasizes the need to maintain human control over the use of force. It expresses concern over a wide range of risks, including arms races, conflict escalation, miscalculations, technology proliferation to non-state actors, and ethical and algorithmic biases with potentially harmful consequences for marginalized groups. At the same time, it recognizes the potential benefits of AI, particularly in enhancing the protection of civilians and improving compliance with IHL. The resolution calls on states to promote responsible use of AI through multilateral mechanisms in cooperation with academia, civil society, and the private sector, while strengthening global cooperation and support for developing countries. Furthermore, it requests the UN Secretary-General to collect the views of Member States and relevant stakeholders regarding broader security-related aspects of military AI, beyond the specific focus on LAWS, as a basis for further normative and policy discussions within the UN (United Nations General Assembly, 2024).

This resolution was preceded by two critical summits under the banner „Responsible AI in the Military Domain“ (REAIM). The first summit was held in The Hague in February 2023, and the second in Seoul in September 2024. These forums brought together dozens of states, scientists, and experts to discuss a broad range of military applications of AI, and to promote voluntary commitments to its responsible use. At the same time, during the Hague Summit, the Global Commission on Responsible AI in the Military Domain (GC REAIM) was established as an international body of experts to support the responsible governance of AI in the military context (The Hague Centre for Strategic Studies, n.d.). Also at the summit, the United States led the launch of a Political Declaration on Responsible Military Use of AI and Autonomy, which outlines ten core principles ranging from compliance with international law and transparency to reliability, human oversight, and sharing of best practices (U.S. Department of State, 2023).

Following the first REAIM Summit, several government representatives issued a joint Call to Action on the responsible development, deployment, and use of AI in the military domain. The document emphasises that AI applications can enhance decision-making, the precision of military operations, and civilian protection. Still, they also entail significant risks, including unpredictable consequences, escalation of conflicts, uncertainty regarding accountability, and the potential for misuse. The key findings underscore the necessity of preserving human responsibility in decision-making, establishing technical and ethical standards, strengthening knowledge exchange among states and stakeholders, and emphasizing education and capacity-building regarding the limitations and dangers of AI systems. The Call advocates adopting national frameworks and strategies, highlighting the importance of data protection, transparency, standardisation, and the roles of civil society and academic institutions in analysing and developing practical solutions. It concludes with an appeal to continue inclusive, global, and interdisciplinary dialogue to ensure the responsible use of AI in compliance with international law and in service of international peace and security (Elsa Lab Defence, 2023).

In parallel with these initiatives, approximately 60 countries endorsed the Blueprint for Action at the Seoul Summit in 2024 (Harjani, 2025). The document acknowledges that AI brings substantial benefits to military operations, from surveillance and reconnaissance to command, logistics, and analysis. Still, it also warns of serious humanitarian, legal, technological, and ethical risks. Key findings include the necessity for AI in the military domain to be developed and deployed in accordance with international law, particularly IHL; the critical

importance of retaining human responsibility and control in key military decisions; and the need for establishing national and global strategies, legal frameworks, standards, and testing mechanisms. A special emphasis is placed on concerns regarding the potential misuse of AI, including autonomous weapons, cyber operations, and scenarios that may aggravate geopolitical tensions (REAIM, 2024).

These two summits, along with the Political Declaration, provide helpful context for UN Resolution A/RES/9/239. In its official submission to the Secretary-General in response to the Resolution, GC REAIM emphasized that the military use of AI simultaneously introduces significant opportunities and serious risks to international peace and security. The Commission highlighted the need to develop clear, technically grounded, and legally binding governance frameworks that ensure human accountability in decision-making, in full compliance with international law, particularly IHL. Among the potential benefits of AI, the Commission identified improvements in military operational precision and efficiency, enhanced early-warning mechanisms, strengthened civilian protection, and reduced human error and bias in critical decisions. On the other hand, the Commission warned of considerable risks, including destabilization through arms race, the use of AI for repressive purposes and human rights violations, and the erosion of inter-state trust. A particular concern was the possible integration of AI into nuclear command and control systems, the development of autonomous weapons, and the exacerbation of asymmetric conflicts. Therefore, the Commission called for multilateral cooperation, trust-building, and the establishment of transparency mechanisms, including technical standards and oversight throughout the AI system lifecycle, to ensure responsible use and safeguard the stability of the international order (UNODA, 2024).

The recommendations to the Secretary-General regarding the implementation of the adopted resolution A/RES//9/239 were also provided by the International Committee of the Red Cross (ICRC) in its „Submission to the United Nations Secretary-General on Artificial Intelligence in the Military Domain“. The document underscores the importance of a human-centred approach, in which AI systems serve as tools of support rather than substitutes for human judgment and responsibility in decisions that affect the lives, safety, and dignity of individuals in armed conflict (ICRC, 2025).

3. Regional and Alliance-Based Frameworks

While multilateral efforts define shared norms, regional organisations and military alliances are increasingly playing a role in operationalising principles and standards for AI-enabled capabilities. In the context of the EU, the first document addressing AI was adopted in 2017 and pertains to the European Parliament Resolution on Civil Law Rules on Robotics. This resolution established general and ethical principles for the development of robotics and AI in a civilian context, while also highlighting restrictions and prohibitions on modifying robots for military purposes (Roković, 2025). Furthermore, in 2019, several EU member states issued the document titled „Digitalization and Artifical Intelligence in Defence“, stating that „the main drivers for using AI applications include achieving superior military capabilities, higher cost-essiciency, and reducing human workload.“ According to the document, this „provides armed forces with new capabilities and oppotrunities both on the physical and virtual battlefield“ (Food for Thought Paper, 2019). This underscores the necessity of adequate regulation and control over the development and use of AI for military purposes. However, the most significant legal act regarding AI within the EU, the Artificial Intelligence Act, which entered into force on August 1, 2024, explicitly excludes AI systems developed or used for military purposes, national defence, and the safeguarding of national interests. This exemption is in accordance with Article 4(2) of the Treaty on the EU, which assigns exclusive responsibility for national security and defence to the member states (Roković, 2025). In addition to the aforementioned regulatory instruments, the EU has adopted several resolutions on the regulation of autonomous weapons (Roković, 2025).

Given the context of AI's military application, it is necessary to consider the perspective of military alliances. In October 2021, NATO adopted its first Artificial Intelligence Strategy (NATO, 2021), which defines Six Principles for the Responsible Use of AI in Defence (NATO, 2024). These principles include lawfulness, requiring that all AI applications comply with national and international law; responsibility and accountability, mandating the precise allocation of human responsibility for the outcomes of AI systems; explainability and traceability, ensuring transparency in design and the ability to track AI decision-making processes; reliability, requiring that AI systems be rigorously tested to function predictably and safely; governability and bias mitigation, ensuring that human operators retain the ability to intervene, correct, or deactivate the system, and the elimination of algorithmic bias and compliance with the principles of non-discrimination. To move from principles to practical implementation, NATO established the Data and AI Review Board (DARB) to develop concrete mechanisms for assessing the compliance of new AI applications with these principles (NATO, 2022).

4. National Defense and Military AI Strategies

The absence of binding global regulation has contributed to the emergence of diverse national models, strategic, competitive, and security-driven, that shape the trajectory of military AI adoption. The United States was among the first to establish a formal strategy. In late 2018, the Department of Defence issued the document „Harnessing AI to Advance Our Security and Prosperity," which outlines plans to accelerate the integration of AI across all defence sectors to maintain U.S. strategic advantage (Department of Defence, 2018). Russia also ranks among the major powers in formulating a national AI document that includes defence-related applications. In July 2022, the Russian government adopted the „Concept of the Russian Armed Forces Activity in the Development and Use of Weapon Systems Using Artificial Intelligence Technology," which functions as a strategic framework guiding Russia's approach to the integration and use of AI in military capabilities (The Ministry of Foreign Affairs of the Russian Federation, 2023). In China, although it is a major actor in AI, the country has not adopted a separate national defence-related AI strategy. Instead, China pursues a „military-civil fusion", in which civilian technological advancements are systematically integrated into military applications to develop advanced weapons systems rapidly. As early as 2017, through the „New Generation Artificial Intelligence Development Plan", China set a national objective to become the global leader in AI by 2023 (Roković, 2024). This list does not exhaust all states that have adopted documents regulating the development and use of AI for military purposes. Still, a comprehensive overview of each would exceed the scope of this paper.

The reviewed documents, particularly those at the global level, emphasise that while AI offers significant advantages for military operations, such as improved precision, decision-making, and civilian protection, it also introduces substantial risks that require urgent and coordinated international governance. Among the most pressing concerns are conflict escalation, arms races, misuse by non-state actors, algorithmic bias, legal uncertainty, and the erosion of trust among states. A recurring focus is the threat posed by autonomous weapons systems, especially their potential to operate without meaningful human oversight. Their unpredictability increases the likelihood of unlawful attacks and unintended escalation, while also raising serious accountability issues when decisions to use lethal force are delegated to machines. Additional concerns include their possible misuse in cyber operations and integration into nuclear command systems. These risks underscore the need for transparent oversight, ethical and technical safeguards, and the preservation of human responsibility in the use of force. Therefore, the following section will focus specifically on the regulation of autonomous weapons systems.

5. The Debate on LAWS

In the context of developing and employing AI tools for military purposes, LAWS have generated by far the most debate and division within the academic and professional community. At the outset, it is essential to emphasize that there is no consensus on a single definition of LAWS (Kmentt, 2025), which, in turn, leads to differing interpretations of these systems, including both their perceived benefits and potential drawbacks. On the one hand, there is a prevailing belief that LAWS can process information rapidly, enabling faster and more precise decision-making in combat (Baxter, 2024). Replacing human soldiers may reduce casualties and prompt them to act more cautiously in uncertain situations (Petman, 2017). Proponents of autonomous weapon systems, including those arguing that existing international humanitarian law and Article 36 weapons reviews provide sufficient oversight (Meier, 2016), as well as States emphasizing strategic and operational advantages, contend that autonomy can enhance precision, reduce military casualties, and mitigate human error within current legal frameworks. By contrast, scholars and institutions advocating prohibition or strict regulation (Heyns, 2016; Human Rights Watch & International Human Rights Clinic, 2012; Davison, 2018; Geneva Academy of International Humanitarian Law and Human Rights, 2014; United Nations Office for Disarmament Affairs, 2023) maintain that delegating life-and-death decisions to machines undermines human dignity and erodes the principles of distinction, proportionality, and accountability, while introducing risks of systemic unpredictability, algorithmic bias, and democratic harm. Ultimately, the divide is not merely technological but normative: supporters view risks as manageable through regulation and improved design, whereas critics argue they are inherent to the concept of autonomous force.

Citing legal and moral grounds, Human Rights Watch and a coalition of non-governmental organizations united under the Campaign to Stop Killer Robots have emerged as the principal protagonists in opposing these technologies. Thanks to their advocacy efforts, the discussion on LAWS has been elevated to the international level. One of the earliest official reports to highlight the dangers associated with LAWS was the 2013 report by Christof Heyns, the United Nations Special Rapporteur on extrajudicial, summary, or arbitrary executions (United Nations General Assembly, 2013). He underscored the need for meaningful human control over the use of force, arguing that autonomous machines are inherently unable to make the nuanced, context-specific judgments

required to safeguard human life. Since then, numerous discussions have taken place under the auspices of the UN and within the framework of the Convention on Conventional Weapons (CCW) to establish appropriate regulatory approaches to this issue. The Group of Governmental Experts (GGE) on LAWS, established under the CCW in 2016, serves as the principal forum for examining emerging military technologies related to LAWS and formulating potential regulatory approaches. Meeting annually since 2017, the Group has progressively deepened the global debate on autonomous weapons. Over time, its discussions have addressed a broad range of technical, legal, political, security, humanitarian, and ethical issues. However, despite meaningful progress and growing momentum, fundamental disagreements persist, especially over whether new legally binding norms are required, resulting in continued lack of consensus on the future regulation of LAWS (UNODA, 2023). Many states, primarily the United States, Russia, China, the United Kingdom, Israel, and others, are investing in the development of military systems with varying degrees of autonomy (Human Rights Watch, 2020), aiming to enhance precision, decision-making speed, and operational efficiency. Within the framework of CCW, the United States and Russia oppose a preemptive ban, arguing that LAWS may enhance precision, reduce harm to civilians, and can be adequately governed by existing IHL. At the same time, Russia additionally stresses the lack of a legal precedent for prohibiting an entire weapons category in advance. China advances a more selective approach, supporting a ban only on „unacceptable“ LAWS, defined as fully autonomous, lethal systems incapable of human intervention, indiscriminate in effect, or capable of autonomous learning, while rejecting broader restrictions on other „acceptable“ LAWS (Kelley, 2025). This trend reflects an accelerating global technological competition and the potential emergence of a new AI-driven arms race.

In a statement addressed to the GGE in March 2019, the UN Secretary-General emphasized that „machines with the power and discretion to take lives without human involvement are politically unacceptable, morally repugnant and should be banned by international law“ (United Nations, 2025). In his New Agenda for Peace, the Secretary-General reiterated this call, recommending that states adopt, by 2026, a legally binding instrument banning LAWS that operate without human control or supervision and cannot be used in accordance with IHL, and regulating all other types of autonomous weapons systems (United Nations, 2023). He noted that, in the absence of specific multilateral regulation, the design, development, and use of such systems raise humanitarian, legal, security, and ethical concerns and constitute a direct threat to human rights and fundamental freedoms. Some progress was made in 2023 with the adoption of the first UN resolution (78/241), which expressed concern that such weapons may lead to an arms race, conflict, and increased risks to civilians due to potential failures in AI decision-making (United Nations General Assembly, 2023). Subsequently, in December 2024, the UN General Assembly adopted Resolution 79/62, which, for the first time, formally recognized the growing dangers posed by autonomous weapons systems and institutionalized the global debate on this issue (United Nations General Assembly, 2024). This development paves the way for an international treaty to prohibit or regulate LAWS, thereby addressing humanitarian, legal, and ethical concerns associated with their use. Achieving meaningful progress will require sufficient political will on the part of states to initiate negotiations, define concrete standards, and adopt a legally binding instrument. However, states remain divided over regulation and prohibition. Several states have voiced concern that the unchecked development of these technologies could trigger an uncontrolled arms race and lower the threshold for entering into armed conflict. Countries such as Austria, Ireland, and Brazil have for years advocated a pre-emptive prohibition on fully autonomous weapons. In contrast, others, including the U.S., Russia, and Israel, argue that existing laws on armed conflict remain sufficient and oppose the introduction of new restrictions (Human Rights Watch, 2020). In September 2025, the UN Security Council held an open debate on AI, during which a clear warning was issued that AI „must not be allowed onto the battlefield without oversight and regulation“ (International Committee of the Red Cross, 2025).

6. Conclusion

The review of existing documents and initiatives demonstrates that the military use of AI is recognized as an increasingly important issue at the global, regional, and national levels. However, despite growing attention and several notable developments, there is still no binding international instrument dedicated to regulating military AI. The existing approaches remain dispersed, and the current governance architecture is characterized by limited alignment and persistent differences among states regarding the scope and nature of future norms.

The question of autonomy in the use of force represents the central point of divergence. While some actors emphasize the potential benefits of AI to improve precision, efficiency, and civilian protection, others warn of risks to accountability, the erosion of human control, and the possibility of unintended consequences, particularly in complex and asymmetric environments. The debate surrounding LAWS reflects these fundamental divides and illustrates the broader challenges of establishing clear, universally accepted standards.

Given the rapid development of AI technologies and their growing influence on the character and conduct of warfare, the need for precise regulation remains increasingly evident. Yet, the pace of technological advancement and the absence of consensus continue to hinder progress. As a result, current efforts suggest gradual movement but leave open significant questions regarding verification, accountability, and the future direction of governance. The evolving landscape indicates that the regulation of military AI will remain a critical topic, shaped by ongoing discussions and the interplay between national interests, technological capabilities, legal frameworks, and security considerations.

Literature

1. Amoroso, D., & Tamburrini, G. (2020). Autonomous weapons systems and meaningful human control: ethical and legal issues. *Current Robotics Reports*, 1(4), 187-194.
2. Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International review of the Red Cross*, 94(886), 687-709.
3. Baxter, W. C. A. (2024). Enhancing Tactical Level Targeting With Artificial Intelligence. Field Artillery Professional Bulletin, 1, 11-13.
4. Burton, J., & Soare, S. R. (2019, May). Understanding the strategic implications of the weaponization of artificial intelligence. 11th international conference on Cyber Conflict (CyCon). IEEE. Tallinn, Estonia.
5. Bruun, L., & Bo, M. (2025). *Bias in Military Artificial Intelligence and Compliance with International Humanitarian Law.* Stochlom: Stochlom International Peace Research Institute.
6. Davison, N. (2018). A legal perspective: Autonomous weapon systems under international humanitarian law. UNODA Occasional Papers, No. 30.
7. Department of Defence. (2018). Summary of the 2018 Department of Defence Artificial Intelligence Strategy Harnessing AI to Advance Our Security and Prosperity. Retrieved on 29 November 2025, from https://apps.dtic.mil/sti/pdfs/AD1114486.pdf
8. Elsa Lab Defence. (2023, Februari 24). REAIM Call for Action. Retrieved on 29 November 2025, from https://elsalabdefence.nl/wp-content/uploads/2023/02/REAIM-2023-Call-to-Action.pdf
9. Food for Thought Paper. (2019, May 17). Digitalization and Artfical Intelligence in Defence. Retrieved on 02 December 2025, from https://valtioneuvosto.fi/documents/11707387/12748699/Digitalization+and+AI+in+Defence.pdf/151e10fd-c004-c0ca-d86b-07c35b55b9cc/Digitalization+and+AI+in+Defence.pdf
10. Franke, U. E. (2018). Military robots and drones. In *Routledge Handbook of Defence Studies* (pp. 339-349). London & New York: Routledge.
11. Geneva Academy of International Humanitarian Law and Human Rights. (2014). *Autonomous weapon systems under international law*. Retrieved on 11 November 2025, from https://geneva-academy.ch/wp-content/uploads/2025/09/Autonomous-Weapon-Systems-under-International-Law_Academy-Briefing-No-8.pdf
12. Harjani, M. (2025). The REAIM "Blueprint for Action" Needs Skin in the Game. IDSS Paper. 006/2025.
13. Heyns, C. (2016). Human rights and the use of autonomous weapons systems (AWS) during domestic law enforcement. *Human Rights Quarterly*, 38(2), 350-378.
14. Hibbert, Z., Overton, I. (2025). Kill codes and command lines: understanding the rise of algorithmic warfare. Action on Armed Violence. Retrieved on 11 November 2025, from https://aoav.org.uk/2025/kill-codes-and-command-lines-understanding-the-rise-of-algorithmic-warfare/
15. Human Rights Watch & International Human Rights Clinic. (2012). *Losing humanity: The case against killer robots*. Human Rights Watch. Retrieved on 06 December 2025, from https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots
16. Human Right Watch. (2020, August 10). Stopping Killer Robots. Country positions banning fully autonomous weapons. Retrieved on 06 December 2025, from https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and#:~:text=China%2C%20Israel%2C%20Russia%2C%20South,countries%20are%20also%20making%20investments.
17. Human Rights Watch & International Human Rights Clinic. (2025). *A hazard to human rights: Autonomous weapons systems and digital decision-making*. Retrieved on 08 December 2025, from

https://www.hrw.org/report/2025/04/28/hazard-human-rights/autonomous-weapons-systems-and-digital-decision-making

18. International Committee of the Red Cross. (2025, September 26). UN Security Council: We cannot let AI be deployed on the battlefield without oversight and regulation. Retrieved on 12 December 2025, from https://www.icrc.org/en/statement/we-cannot-let-AI-be-deployed-on-battlefield-without-oversight-and-regulation
19. ICRC. (2025). Submission to the United Nations Secretary-General on Artificial Intelligence in the Military Domain. Retrieved on 12 December 2025, from https://www.icrc.org/sites/default/files/2025-04/ICRC_Report_Submission_to_UNSG_on_AI_in_military_domain.pdf
20. Jensen, B., Atalan, Y., & Macias III, J. M. (2024). Algorithmic stability: How ai could shape the future of deterrence. Center for Strategic and International Studies. Retrieved on 05 December 2025, from 240610_Jensen_Algorithmic_Stability.pdf
21. Johnson, J. (2020). Artificial intelligence: A threat to strategic stability. *Strategic studies quarterly*, 14(1), 16-39.
22. Kelley, S. (2025, February 25). International Discussion Concerning Lethal Autonomous Weapon Systems. Congressional Research Service. Retrieved on 15 December 2025, from https://www.congress.gov/crs-product/IF11294#:~:text=Although%20the%20CCW%20operates%20by,increased%20accuracy%20of%20weapon%20guidance
23. Kmentt, Al. (2025). Geopolitics and the Regulation of Autonomous Weapons Systems. Arms Control Association. Retrieved on 15 December 2025, from https://www.armscontrol.org/act/2025-01/features/geopolitics-and-regulation-autonomous-weapons-systems
24. Kohn, S., Cohen, M., Johnson, A., Terman, M., Weltman, G., & Lyons, J. (2024). Supporting ethical decision-making for lethal autonomous weapons. *Journal of Military Ethics*, 23(1), 12-31.
25. Márquez-Díaz, J. E. (2024). Benefits and challenges of military artificial intelligence in the field of defense. *Computación y Sistemas*, 28(2), 309-323.
26. Meier, M. W. (2016). Lethal autonomous weapons systems (LAWS): conducting a comprehensive weapons review. *Temp. Int'l & Comp. LJ*, 30, 119.
27. NATO. (2021, October 22). Summary of the NATO Artificial Intelligence Strategy. Retrieved on 12 December 2025, from https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2021/10/22/summary-of-the-nato-artificial-intelligence-strategy
28. NATO. (2022, October 13). NATO's Data and Artificial Intelligence Review Board. Retrieved on 10 December 2025, from https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2022/10/13/natos-data-and-artificial-intelligence-review-board
29. NATO. (2024, July 10). Summary of the NATO's revised Artificial Intelligence (AI) Strategy. Retrieved on 10 December 2025, from https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2024/07/10/summary-of-natos-revised-artificial-intelligence-ai-strategy
30. Nguyen, N. (2025, September 10). AI in Military: Top Use Cases You Need To Know. SmartDev. Retrieved on 27 November 2025, from https://smartdev.com/ai-use-cases-in-military/
31. Osimen, G. U., Newo, M., & Fulani, O. (2024). *Artificial Intelligence and Arms Control in Modern Warfare. Cogent Social Sciences*, 10(1), 2407514.
32. REAIM. (2024, September 9-10). REAIM Blueprint for Action. Retrieved on 17 November 2025, from https://www.mofa.go.kr/www/brd/m_4080/down.do?brd_id=235&seq=375378&data_tp=A&file_seq=9
33. Roković, V. (2024). Back to the Future: The US-China AI Arms Race?. Scientific conference with international participation „Harvesting the Winds of Change: China and the Global Actors", Belgrade.
34. The Hague Centre for Strategic Studies. (n.d.). Global Commission on Responsible AI in the Military Domain (GC REAIM). Retrieved on 17 November 2025, from https://hcss.nl/gcreaim/
35. The Ministry of Foreign Affairs of the Russian Federation. (2023, October 23). Statement by Andrey Belousov, Deputy Head of the Delegation of the Russian Federation, at the Thematic Debate on "Conventional Weapons" in the First Committee of the 78th Session of the UN General Assembly, New York, 23 October 2023. Retrieved on 02 December 2025, from https://mid.ru/en/foreign_policy/news/1911643

36. United Nations. (2023, July). A New Agenda for Peace. Policy Brief 9. Retrieved on 17 December 2025, from https://www.un.org/climatesecuritymechanism/sites/default/files/2025-06/our-common-agenda-policy-brief-new-agenda-for-peace-en_0.pdf
37. United Nations. (2025, May 12). Leathal Autonomus Weapon System 'Politically Unacceptable, Morally Repugnant and Should Be Banned', Secretary-General Says during Informal Consultations on Issue. Retrieved on 15 December 2025, from https://press.un.org/en/2025/sgsm22643.doc.htm
38. United Nation General Assembly. (2013). Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns. A/HRC/23/47. Retrieved on 17 December 2025, from https://digitallibrary.un.org/record/755741?v=pdf
39. United Nation General Assembly. (2023, December 28). 78/241. Lethal autonomous weapons systems. Retrieved on 15 December 2025, from https://docs.un.org/en/a/res/78/241
40. United Nation General Assembly. (2024, December 10). 79/62. Lethal autonomous weapons systems. Retrieved on 11 December 2025, from https://docs.un.org/en/a/res/79/62
41. United Nations General Assembly. (2024, December 24). Artificial intelligence in the military domain and its implications for international peace and security, A/RES/79/239. Retrieved on 12 December 2025, from unidir.org/wp-content/uploads/2025/03/UN_General_Assembly_A_RES_79_239-EN.pdf
42. United Nations Office for Disarmament Affairs. (2023). Overview of the issue of lethal autonomous weapons systems at the United Nations. Retrieved on 14 December 2025, from https://wfuna.org/wp-content/uploads/2023/10/GA1-LAWS-background-doc.pdf
43. UNODA. (2024). Submission of the Global Commission on Responsible Artificial Intelligence in the Military Domain to the United Nations Secretary-General in terms of resolution A/RES/79/239 on "Artificial intelligence in the military domain and its implications for international peace and security" adopted by the General Assembly on 24 December 2024 . Retrieved on 12 December 2025, from https://docs-library.unoda.org/General_Assembly_First_Committee_-Eightieth_session_(2025)/79-239-GC_REAIM-EN.pdf
44. U.S. Department of State. (2023, November 9). a Political Declaration on Responsible Military Use of AI and Autonomy. Retrieved on 18 November 2025, from https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/van Creveld, M. (1989). *Technology and War: From 2000 B.C. to the Present*. New York: Free Press; London: Collier Macmillan.
45. Vergun, D. (2019, September 24). AI to Give U.S. Battlefield Advantages, General Says. U.S. Department of War. Retrieved on 13 November 2025, from https://www.war.gov/News/News-Stories/Article/Article/1969575/ai-to-give-us-battlefield-advantages-general-says/
46. Vestner, T. (2024). From strategy to orders: preparing and conducting military operations with artificial intelligence. in Robin Geiß and Henning Lahmann (eds), Research Handbook on Warfare and Artificial Intelligence (Edward Elgar Publishing, (pp. 116-134). Edward Elgar Publishing.
47. Zequeira, M. (2024). Artificial Intelligence as a Combat Multiplier. *Military Review*. Army University Press, Retrieved on 13 November 2025, from https://www.armyupress.army.mil/Journals/Military-Review/Online-Exclusive/2024-OLE/AI-Combat-Multiplier/
48. Wells, W. (2025). Battlefield Evidence in the Age of Artificial Intelligence-Enabled Warfare. *Chi. J. Int'l L.*, *26*, 217.
49. Roković, V. (2025). Политика ЕУ у развоју и примени вештачке интелигенције у војне сврхе: изазови стратешке аутономије, технолошке суверености и утицај на Заједничку безбедносну и одбрамбену политику. У: Суботић, М. (ур.). Изазови регионалне безбедности. Медија центар Одбрана. 35-67

Original Scientific Paper/Original naučni rad UDC/UDK: 33:[004.8:355.45"20"
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601054S

Ekonomski aspekti veštačke inteligencije i bezbednosti u dvadeset prvom veku

Panteleimon Sklias[1], Duško Tomić[2]

[1] Neapolis University Pafos, Cyprus, p.sklias@nup.ac.cy
[2] American University in the Emirates, UAE, dusko.tomic@aue.ae

Apstrakt: Ekonomski aspekt veštačke inteligencije (VI) i bezbednosti u 21. veku predstavlja dinamičnu raskrsnicu tehnoloških inovacija, globalne konkurentnosti i nacionalne otpornosti. VI je postala ključni pokretač ekonomskog rasta, poboljšavajući produktivnost, optimizujući proces donošenja odluka i transformišući industrije kao što su finansije, odbrana i kritična infrastruktura. Međutim, ista ova tehnologija uvodi nove ranjivosti, uključujući sajber pretnje, povrede podataka i algoritamsku manipulaciju koje mogu da potkopaju finansijsku stabilnost i nacionalnu bezbednost. Integracija VI u ekonomiju odbrane omogućava državama da smanje operativne troškove, automatizuju obaveštajne procese i predviđaju bezbednosne rizike sa neviđenom preciznošću. U međuvremenu, globalna trka u naoružanju VI je intenzivirala ekonomsku konkurenciju među velikim silama, dovodeći do strateških investicija u istraživanja zasnovana na VI, vojnu modernizaciju i regulatorne okvire. Ekonomija 21. veka je stoga sve više definisana „algoritamskom bezbednošću", gde podaci, znanje i računarska moć predstavljaju strateška sredstva. Balansiranje inovacija sa etičkim upravljanjem, ekonomskom održivošću i međunarodnom saradnjom ostaje suštinsko kako bi se iskoristio potencijal VI uz ublažavanje njenog disruptivnog uticaja na ekonomske i bezbednosne sisteme.

Ključne reči: Veštačka inteligencija (VI), Bezbednost, Ekonomija, Tehnološke inovacije, Sajber bezbednost, Ekonomski rast, Algoritamska bezbednost, Nacionalna otpornost, Ekonomija odbrane, Globalna konkurentnost, Etičko upravljanje

Economic Aspects of Artificial Intelligence and Security in the 21st Century

Abstract: The economic aspect of Artificial Intelligence (AI) and security in the 21st century represents a dynamic intersection of technological innovation, global competitiveness, and national resilience. AI has become a core driver of economic growth, enhancing productivity, optimizing decision-making, and transforming industries such as finance, defence, and critical infrastructure. However, this same technology introduces new vulnerabilities, including cyber threats, data breaches, and algorithmic manipulation that can undermine financial stability and national security. The integration of AI in defence economics enables states to reduce operational costs, automate intelligence processes, and predict security risks with unprecedented precision. Meanwhile, the global AI arms race has intensified economic competition among major powers, leading to strategic investments in AI-driven research, military modernization, and regulatory frameworks. The 21st-century economy is thus increasingly defined by "algorithmic security," where data, knowledge, and computational power constitute strategic assets. Balancing innovation with ethical governance, economic sustainability, and international cooperation remains essential to harness AI's potential while mitigating its disruptive impact on economic and security systems.

Keywords: Artificial Intelligence (AI), Security, Economy, Technological Innovation, Cybersecurity, Economic Growth, Algorithmic Security, National Resilience, Defence Economics, Global Competitiveness, Ethical Governance

1. Introduction

Rapid technological advancements in recent years have changed various industries, particularly regarding economic security and data protection. Artificial intelligence (AI) leads these developments. It promises better data processing, predictive analytics, and operational speed across fields like healthcare, finance, and public safety. But organizations face a difficult choice alongside these advancements. AI can strengthen security

protocols and manage resources better, yet it also introduces new vulnerabilities and ethical dilemmas. These issues require thorough examination. The central research problem of this dissertation involves understanding economic implications. It balances the benefits of AI technologies against potential risks to security frameworks. The primary objectives of this section define the economic environment shaped by AI innovations and security concerns. The text analyzes how AI applications can improve decision-making without sacrificing privacy and data integrity. This study also aims to identify effective strategies for institutions to mitigate risks associated with AI implementation. These technologies must serve as a reliable first line of defense rather than a cause of increased insecurity. Addressing the economic aspects of AI and security is important for academia and practical applications. This section contributes to the discussion surrounding technology adoption in business. It highlights economic benefits from AI security solutions and acknowledges that technological advancements and emerging threats are connected. The research explains this relationship. It helps readers understand the socio-economic dynamics that influence organizational behaviors during rapid technological change. Policymakers, practitioners, and researchers must act proactively. They need to develop frameworks that support ethical AI deployment. These frameworks must address data privacy, compliance, and potential job displacement from automation. This section strives for a balanced integration of AI within security contexts. It provides a solid foundation to assess benefits and risks. This work helps formulate resilient business models suited for the challenges of the contemporary economy. (Rani J, 2025). The primary objectives of this section are to delineate the economic landscape shaped by AI innovations and security concerns, analyzing how AI applications can be harnessed for enhanced decision-making without sacrificing privacy and data integrity (Grochmalski P et al., 2020) . Additionally, this study aims to identify effective strategies that institutions can adopt to mitigate risks associated with AI implementation, ensuring that these technologies serve as a reliable first line of defense rather than a catalyst for increased insecurity (D Kozenkov et al., 2025). The significance of addressing the economic aspects of AI in relation to security is paramount for both academia and practical applications. This section contributes to the ongoing discourse surrounding technology adoption in business by highlighting the economic benefits derived from AI-driven security solutions while also acknowledging the intertwined nature of technological advancements and emerging threats (Malynovska Y et al., 2025). By elucidating this complex relationship, the research facilitates a more comprehensive understanding of the socio-economic dynamics influencing organizational behaviors in the face of rapid technological change (Vicol D, 2025). Furthermore, it underscores the necessity for policymakers, practitioners, and researchers to adopt a proactive stance in developing frameworks that support ethical AI deployment, addressing concerns related to data privacy, compliance, and the potential displacement of jobs due to automation (Lone A et al., 2025). In striving for a balanced integration of AI within security contexts, this section provides a solid foundation for the assessment of benefits and risks, ultimately contributing to the formulation of resilient business models suited for the challenges of the contemporary economy (M Cg et al., 2025).

2. Literature Review

Rapid technological progress defines our era. The use of Artificial Intelligence (AI) across sectors marks innovation. But it also drives economic change and security worries. The 21st century has seen massive growth in AI power. This shifts how industries work and shapes the global economy. Scholars have started to explore these links. They highlight the economic results of using AI and its effects on national and international security [cite]. Leaders must understand the economics of AI and security to make good policies. AI tools can boost output, smooth out operations, and lower costs. Yet, they bring new security problems [cite]. Recent literature shows a split role for AI. It sparks economic growth but raises issues about cyber threats, watching people, and automated decisions [cite]. These technologies affect both economic health and security dilemmas. Governments and groups face these mixed implications [cite]. Interest in this area is growing, but gaps remain in the literature. We need to know the specific economic effects on different sectors and the security challenges that follow. Most research looks at financial gains. Few studies examine the long-term economic dangers of relying too much on AI [cite]. Few papers analyze the social and economic gaps AI might worsen, especially regarding job loss and skill mismatches [cite]. Economic and security concerns often overlap, but scholars have not studied this enough. This leaves policymakers with split knowledge that hinders planning [cite]. Writers often separate AI discussions into economic or security boxes. They ignore the need to combine these views. Economic motives shape security policies, and security needs shape economics [cite]. This gap matters for global power. National economic interests often meet the need for stronger security against AI threats [cite]. The current debate shows a clear need for a mixed plan. We must look at both the money incentives and the security results of AI innovations [cite]. Because of these complexities, we must define the economic side of AI as it relates to security. This clarity helps policymakers and industry leaders build better plans [cite]. This review gathers current findings, points out missing pieces, and suggests future research paths [cite]. It combines ideas from many studies. The goal is a full

understanding of how AI economics and security shape the future [cite]. This talk is vital. We must maximize economic gains while strengthening defenses against the risks of spreading AI technology. The study of AI economics and security has changed since the early 2000s. Technology moved fast, and society reacted. At first, scholars looked at AI theories and money benefits. They noted its power to change sectors like finance and health care [cite]. By the mid-2010s, the focus turned to practical uses for security and risk management. Researchers like [cite] noted that adding AI to security systems could improve predictions and cut costs. This showed the dual value of AI. Later writings addressed ethics and the economic effects of AI security tools. Authors such as [cite] and [cite] looked at surveillance, privacy, and the chance that AI might worsen economic gaps. This was a turning point. Writers started asking for rules and ethical guides to balance tech progress with social impacts. In the late 2010s and early 2020s, researchers saw that AI developers, economists, and policymakers must work together. Studies by [cite] and [cite] showed that mixing these fields drives new ideas. They also make sure growth does not harm security or ethics. These analyses show a growing agreement. The mix of AI, economics, and security needs constant watching as technology changes. Recent literature highlights the economic results of AI in security. This intersection creates a complex picture where money drives both new tools and new weaknesses. Scholars note that AI improves security but brings new economic threats, especially in cybersecurity. For example, research shows that AI in defense systems lowers costs. This changes how nations invest in security [cite]. But these changes bring risks. Enemies can use AI systems, which leads to big money losses for businesses and governments [cite]. Rules regarding the economic impact of AI on security are also important. Many studies say we need strong policies to manage the benefits and threats. Without good rules, economic gaps could grow as companies try to adapt [cite]. Literature on the labor market suggests AI might replace some jobs. At the same time, it creates new roles in cybersecurity. This changes the economy in ways we cannot fully predict [cite]. AI is global, so we must look at international economic effects. Different countries adopt AI at different speeds, which changes global security. Studies link economic power to tech leadership. Nations that invest heavily in AI are better placed to protect their interests [cite]. These themes show the tight web connecting AI, money, and security in this century. Researchers use different methods to study AI and security economics. Qualitative analyses give deep details on how economic effects shape security rules. Researchers using this method say organizations adapt their strategies to new AI tools. They stress that understanding human behavior is as important as the technology itself [cite]. Quantitative studies measure the money impact of investing in AI for security. They show high returns on investment, which matters for keeping organizations running [cite]. Mixed methods work well. They bridge stories with data [cite]. These studies suggest that adding AI is more than a tech upgrade. It is an economic strategy that changes security at many levels [cite]. Some researchers argue we must include ethics in economic reviews. Ignoring these points could lead to surprise security holes [cite]. Most agree that economic plans must account for both the chances and threats of AI security. Studies say policymakers must help innovation grow. But they must also build rules to lower risks from economic gaps caused by AI [cite]. This discussion highlights a changing field where different methods help us understand the hard parts of AI and security economics. The significance of understanding the economic aspects of AI and security lies in the potential for informed policy-making and strategic development. AI technologies have the capacity to enhance productivity, streamline operations, and reduce costs across industries while also presenting novel challenges and vulnerabilities in the realm of security (Khan MK et al., 2024). Major findings in contemporary literature reveal a duality in AIs role: on one hand, it serves as a catalyst for economic growth, and on the other, it raises questions regarding cybersecurity threats, surveillance, and the ethical considerations of automated decision-making (Grochmalski P et al., 2020). The evolving landscape of AI technologies thus straddles both economic vitality and security dilemmas—two themes that are increasingly intertwined as governments and organizations grapple with their implications (D Kozenkov et al., 2025). Despite the burgeoning interest in this area, substantial gaps remain in the literature regarding the specific economic impacts of AI on different sectors and the nuanced security challenges that arise from its adoption. For instance, while much research has focused on AIs financial benefits, limited attention has been directed toward understanding the long-term economic risks associated with a heavy reliance on AI technologies (Malynovska Y et al., 2025). Additionally, there is a paucity of studies that comprehensively analyze the socio-economic disparities exacerbated by AI, particularly in terms of job displacement and potential skill mismatches within the workforce (Vicol D, 2025). Furthermore, the intersectionality of economic and security concerns has not been sufficiently addressed, leaving scholars and policymakers with fragmented knowledge that complicates robust strategic planning (Lone A et al., 2025) . Moreover, existing literature often tends to compartmentalize discussions around AI within economic versus security frames, neglecting the need for a more integrative approach that considers how economic motives can shape security policies and vice versa (M Cg et al., 2025). This disconnect is particularly striking in the context of global power dynamics, where national economic interests increasingly intersect with the imperative of strengthening security measures to counteract AI-driven threats (Михаил Михайлович Куликов et al., 2025). As a result, the existing discourse underscores a pressing need for an interdisciplinary approach that accounts for both

economic incentives and security ramifications of AI innovations (Yusuf SO et al., 2024). In light of these complexities, it is essential to delineate a clearer understanding of the economic aspects of AI in relevance to security, as this can foster more comprehensive frameworks for policymakers and industry leaders (Klius Y et al., 2024). The ensuing literature review aims to synthesize current findings, illuminate critical gaps, and propose avenues for future research that can further bridge these interrelated domains (Baidoo-Anu D et al., 2023). By collating insights from a diverse range of studies, this review seeks to contribute to a more holistic comprehension of how the economic and security facets of AI are shaping the future both nationally and globally (Budhwar P et al., 2023) (Kuwaiti AA et al., 2023) (Malik S et al., 2023). This dialogue is vital to not only optimize economic benefits but also reinforce security measures against risks posed by the rapid proliferation of AI technologies in our increasingly interconnected world (Kraus S et al., 2021) (Varnosfaderani SM et al., 2024) (Shuroug A Alowais et al., 2023) (Enholm IM et al., 2021) (Taher M Ghazal et al., 2021). The exploration of the economic aspects of Artificial Intelligence (AI) and security has evolved significantly since the early 21st century, reflecting the rapid advancements in technology and their societal implications. Initially, scholars focused primarily on the theoretical foundations of AI and its potential economic benefits, highlighting its transformative capabilities in various sectors, including finance and health care (Rani J, 2025) (Khan MK et al., 2024). By the mid-2010s, discussions shifted to the practical applications of AI in enhancing security protocols and risk management strategies. Researchers like (Grochmalski P et al., 2020) emphasized that integrating AI with existing security frameworks could enhance predictive accuracy and reduce operational costs, thereby underscoring AIs dual economic and security value. Further developments noted in the literature began to address the ethical dimensions and economic implications of deploying AI technologies in security applications. Works by (D Kozenkov et al., 2025) and (Malynovska Y et al., 2025) explored concerns related to surveillance, privacy, and their potential to exacerbate economic inequalities. This period marked a critical turning point as authors began advocating for regulatory frameworks and ethical guidelines to balance AI advancements with societal impacts. As the discourse progressed into the late 2010s and early 2020s, researchers increasingly recognized the necessity of collaboration between AI developers, economists, and policymakers. Studies by (Vicol D, 2025) and (Lone A et al., 2025) illustrated how interdisciplinary approaches could drive innovation while ensuring that economic growth does not come at the expense of security or ethical considerations. Such comprehensive analyses underscore the growing consensus in the literature that the intersection of AI, economics, and security warrants ongoing scrutiny as these technologies continue to evolve.

In conclusion, this literature review elucidates the multifaceted relationship between the economic aspects of Artificial Intelligence (AI) and security in the 21st century, underscoring the need for a nuanced understanding of their interconnectedness. As evidenced by the findings, AI serves a dual role: it is a powerful force for economic growth while simultaneously presenting significant threats to security and ethical standards. Ultimately, this review contributes to a more comprehensive understanding of the intricate balance between the economic advantages offered by AI and the formidable security challenges that accompany its integration into society. The dialogue established here serves as a foundation for future inquiries into how policymakers, industry leaders, and academics can collaboratively navigate this rapidly evolving landscape, seeking to optimize benefits while minimizing risks in the age of AI.

3. Methodology

The increasing use of Artificial Intelligence (AI) across many sectors creates deep economic effects. This is true regarding national and global security. Organizations use AI to improve operations and make better decisions. We must understand the economic results connected to security concerns. This understanding is critical. This study addresses a specific research problem. We lack a full framework that explains the mix of AI's economic benefits and its security threats. This is especially true in the 21st century. This study wants to reach several core objectives. First, it explains how AI technologies improve economic productivity. At the same time, they introduce vulnerabilities and security challenges. Second, the study checks existing policy frameworks. It measures their success in reducing risks from AI use. The research also wants to build a unified analytical model. This model combines economic and security views. It offers valuable facts for stakeholders in both fields. This methodology section is important for two reasons. It has both academic and practical relevance. Academically, it tries to fill a clear gap in current literature. It gives a systematic way to analyze the economic sides of AI related to security. Scholars have often ignored this overlap. Practically, this research gives policymakers, business leaders, and security professionals a deep grasp of AI investment results. They can then start stronger strategies. These plans cover both economic growth and security resilience. The methodology uses a mixed-methods approach. It combines qualitative and quantitative research techniques. It builds on established frameworks. Past studies used these frameworks to analyze similar overlaps. We will use case studies, surveys, and statistical analyses. This methodology gives a full look at the direct and indirect effects of AI on the economy and security situation. It

allows for data triangulation. This improves the strength of the findings. The study fixes the identified gaps and uses a well-rounded approach. It promises to add major knowledge to the fields of economics and security. This helps further discussion and development in this important area. The methodology highlights the critical relationship between AI, economic results, and security needs. It builds a strong framework. This framework supports informed decisions and strategic planning. Technology changes fast, and we must plan for it. (Rani J, 2025). The primary research problem addressed in this study is the absence of a comprehensive framework that articulates the interplay between AIs economic benefits and its associated security threats, particularly in the 21st century (Khan MK et al., 2024). This study aims to achieve several core objectives: first, to elucidate how AI technologies enhance economic productivity while simultaneously introducing vulnerabilities and security challenges; second, to assess existing policy frameworks and their effectiveness in mitigating risks associated with AI deployment (Grochmalski P et al., 2020). Furthermore, this research seeks to create a cohesive and analytical model that integrates economic and security perspectives, thereby offering valuable insights for stakeholders within both spheres (D Kozenkov et al., 2025). The significance of this methodology section lies in its dual academic and practical relevance. Academically, it strives to fill a notable gap in existing literature by providing a systematic approach to analyzing the economic aspects of AI in relation to security, an intersection that has often been overlooked (Malynovska Y et al., 2025). Practically, this research will offer policymakers, business leaders, and security professionals a nuanced understanding of the implications of AI investments, enabling them to implement more robust strategies that encapsulate both economic growth and security resilience (Vicol D, 2025). The methodology will employ a mixed-methods approach that combines qualitative and quantitative research techniques, building upon established frameworks that have successfully analyzed similar intersections in past studies (Lone A et al., 2025). By employing case studies, surveys, and statistical analyses, this methodology will provide a comprehensive look at both the direct and indirect effects of AI on the economy and security landscape while allowing for the triangulation of data to enhance the validity of findings (M Cg et al., 2025) (Михаил Михайлович Куликов et al., 2025). By addressing the identified gaps and employing a well-rounded methodological approach, this study promises to contribute significant knowledge to the fields of economics and security, facilitating further discourse and development in this crucial area of inquiry (Yusuf SO et al., 2024) . In summary, the methodology underscores the vital relationship between AI, economic outcomes, and security considerations, establishing a robust framework that encourages informed decision-making and strategic planning in a rapidly evolving technological landscape (Klius Y et al., 2024) (Baidoo-Anu D et al., 2023) (Budhwar P et al., 2023) (Kuwaiti AA et al., 2023) (Malik S et al., 2023) (Kraus S et al., 2021) (Varnosfaderani SM et al., 2024) (Shuroug A Alowais et al., 2023) (Enholm IM et al., 2021) (Taher M Ghazal et al., 2021) .

4. Results

Based on the established research framework, the reviewed literature, and the applied methodology, it is expected that the study will generate the following key results.

First, the research confirms that the application of artificial intelligence has a **significant positive economic impact** on organizations and public institutions, particularly through increased productivity, reduced operational costs, and improved efficiency in decision-making processes. The findings are expected to demonstrate that AI systems enable faster processing of large volumes of data, which directly enhances economic planning and resource management in security-sensitive sectors such as finance, public administration, and critical infrastructure.

Second, the study identifies a **new set of security and economic risks** associated with the deployment of AI, including increased exposure to cyberattacks, growing dependence on automated systems, and potential systemic failures that may result in substantial financial losses. Particular emphasis is placed on the vulnerability of economic systems that rely on centralized AI models, as well as on the costs related to data protection, regulatory compliance, and post-incident recovery following security breaches.

Third, the research findings indicate an **imbalance between economic benefits and the social consequences** of artificial intelligence adoption. The analysis is expected to show that while AI contributes to economic growth, it may simultaneously exacerbate socio-economic inequalities through job automation, changes in labor market structures, and the widening digital skills gap. These findings may have direct implications for national security strategies, as economic instability and social exclusion constitute long-term security challenges.

Fourth, the research results in the development of an **integrated analytical model** that links economic performance and security risks within the context of AI implementation. Such a model would enable decision-makers to assess not only the financial viability of AI technologies but also their impact on the resilience of

economic and security systems. It is expected that this model will contribute to the formulation of more sustainable digital transformation policies.

Fifth, the findings highlight the need for **strengthening regulatory and institutional frameworks**, particularly in the areas of economic security, data protection, and the ethical application of artificial intelligence. The research may demonstrate that existing legal mechanisms often lag behind technological developments, creating regulatory gaps with potentially serious economic and security consequences.

Finally, it is expected that the research findings will have **substantial practical value**, offering recommendations for public authorities, the private sector, and security organizations regarding the strategic and responsible deployment of artificial intelligence. These results may contribute to the development of national strategies that recognize artificial intelligence not merely as a technological tool, but as a key factor in economic resilience and security in the 21st century.

5. Conclusion

This analysis explores the economic aspects of Artificial Intelligence (AI). AI acts as a tool for productivity. It also creates new vulnerabilities in the security domain. The findings show the value of a mixed-methods approach. This method combined quantitative data with qualitative insights. It provided a clear view of AI's impact on economic productivity and security risks. This research addresses a critical issue. AI drives economic growth. But it also introduces new security challenges for organizations. These findings offer value beyond academia. Policymakers and business leaders can use them. They must use AI's potential but also plan to manage associated risks. AI offers a major chance to improve efficiency and innovation in many sectors. But it needs strong systems to secure sensitive information. Infrastructure must remain safe from potential threats. The research points to practical applications. Organizations need a two-part strategy. They should use AI's economic benefits. But they must also strengthen security measures against new vulnerabilities. Future research should study the link between AI and cybersecurity. It should focus on high-risk sectors. These include finance and critical infrastructure. Researchers should also investigate best practices for AI integration. Different industries need effective operational plans. These plans must balance economic goals with security needs. Future studies must address the socio-economic impacts of AI in different regions. Solutions must work for various groups. They should be adaptable and fair. AI continues to change. Research should focus on ethical guidelines for its use. These rules will build public trust. They will help users accept the technology. Technology advances quickly. Teams of technologists, ethicists, and industry representatives must work together. They will shape a broad approach to using AI. This approach must withstand modern challenges. The discussion about AI must prioritize economic progress and security. This will create resilient systems. These systems will benefit everyone involved. Future research will address these many factors. It will help with the responsible adoption of AI technologies. The economic potential of AI is vast. But we must approach its challenges with rigorous analysis. Collaboration is necessary to protect future projects. (Rani J, 2025). Practical applications of the research indicate that organizations must adopt a dual-focused strategy that capitalizes on AI's economic benefits while reinforcing security measures to protect against emerging vulnerabilities (Khan MK et al., 2024). Future research avenues may include exploring the intersection of AI and cybersecurity in greater depth, particularly in sectors particularly susceptible to AI-related risks such as finance and critical infrastructure (Grochmalski P et al., 2020). Additionally, investigating best practices for the integration of AI across diverse industries could yield insights into effective operational frameworks that balance economic objectives with security imperatives (D Kozenkov et al., 2025). It is also essential for future studies to address the socio-economic impacts of AI in different geographical contexts, ensuring that solutions are adaptable and equitable across various demographic settings (Malynovska Y et al., 2025). Furthermore, as AI continues to evolve, research should focus on developing ethical guidelines that govern its implementation to enhance public trust and facilitate acceptance among users (Vicol D, 2025). With the rapid pace of technological advancements, engaging interdisciplinary teams that include technologists, ethicists, and industry representatives will be vital in shaping a comprehensive approach to AI integration that can withstand the challenges of the 21st century (Lone A et al., 2025). Ultimately, the ongoing dialogue surrounding AI must prioritize both economic progress and security to create resilient systems that benefit all stakeholders involved (M Cg et al., 2025). By addressing these multifaceted considerations, future research can contribute meaningfully to the responsible and sustainable adoption of AI technologies (Михаил Михайлович Куликов et al., 2025). The potential of AI in the economic realm is vast; however, its challenges must be approached with rigorous analysis and collaboration to safeguard future endeavors (Yusuf SO et al., 2024).

References

1. **Ida Merete**, Emmanouil Papagiannidis, Patrick Mikalef, John Krogstie (2021). *Artificial Intelligence and Business Value: a Literature Review.* Information Systems Frontiers, 24, 1709–1734. https://doi.org/10.1007/s10796-021-10186-w
2. **Ghazal, Taher M.**, Mohammad Kamrul Hasan, Muhammad Turki Alshurideh, Haitham M. Alzoubi, Munir Ahmad, Syed Shehryar Akbar, Barween Al Kurdi, et al. (2021). *IoT for Smart Cities: Machine Learning Approaches in Smart Healthcare—A Review.* Future Internet, 13, 218. https://doi.org/10.3390/fi13080218
3. **Grochmalski, Piotr**, P. Lewandowski, Paweł Paszak (2020). *US-China Technological Rivalry and its Implications for the Three Seas Initiative (3SI).* European Research Studies Journal. https://www.semanticscholar.org/paper/73db9a8086d1fb54ea376ae59979488c962b8f35
4. **Klius, Yulia**, V. V. Humenyuk (2024). *Implementation of Artificial Intelligence in the Activities of Organizations.* Management of Economy: Theory and Practice. Chumachenko's Annals. https://www.semanticscholar.org/paper/44ec1f8ed0de45cd09ad0588bf77474eb1ce74fc
5. **Kozenkov, D.**, O. Kaut, Hanna Shportko (2025). *Digital Financial Monitoring as an Instrument for Decision-Making in Public Administration.* Economic Scope. https://www.semanticscholar.org/paper/fc4e0ae29c6d25dc671f3b46bccdaff9cfcd5f7f
6. **Kraus, Sascha**, Paul Jones, Norbert Kailer, Alexandra Weinmann, Nuria Chaparro-Banegas, Norat Roig-Tierno (2021). *Digital Transformation: An Overview of the Current State of the Art of Research.*
7. SAGE Open, 11. https://doi.org/10.1177/21582440211047576
8. **Yusuf, Samuel Omokhafe**, Amarachi Zita Echere, Godbless Ocran, Justina Eweala Abubakar, Adedamola Hadassah, Peprah Owusu (2024). *Analyzing the Efficiency of AI-Powered Encryption Solutions in Safeguarding Financial Data for SMBs.* World Journal of Advanced Research and Reviews. https://www.semanticscholar.org/paper/7be26351627c4d6dc73659e525cfa676e05bef13

Original Scientific Paper/Original naučni rad UDC/UDK: 17:004.8]:502/504
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601061I

Iznad bojnog polja: Etičke implikacije i regulatorni izazovi korišćenja autonomnih AI sistema za bezbednost životne sredine i zaštitu resursa

Aleksandar Ivanov [1], Kire Babanoski[2], Vladimir M. Cvetković[3]

[1]Faculty of Security, Skopje, University "St. Kliment Ohridski", Bitola, North Macedonia, aleksandar.ivanov@uklo.edu.mk
[2]Faculty of Security, Skopje, University "St. Kliment Ohridski" Bitola, North Macedonia, kire.babanoski@uklo.edu.mk
[3]University of Belgrade; Scientific and Professional Society for Risk Management, Serbia, vladimirkpa@gmail.com

Apstrakt: Kako veštačka inteligencija (AI) prelazi iz vojnih i industrijskih domena u nauku o životnoj sredini, fundamentalni zaokret ka metodologijama zasnovanim na podacima redefiniše zaštitu planete. Međutim, ova tranzicija često uvozi „logiku bojnog polja" u očuvanje prirode, koristeći autonomne sisteme — kao što su dronovi i algoritmi mašinskog učenja — koji uvode složene etičke i regulatorne izazove. Ovaj rad predstavlja konceptualnu sintezu okvira veštačke inteligencije usmerene na čoveka (HCAI) i perspektiva ekološke bezbednosti kako bi se odgovorilo na ove rizike. Identifikujemo kritične tačke trenja, uključujući antropocentrične pristrasnosti koje zanemaruju dobrobit ne-ljudskih bića, jaz u odgovornosti u autonomnom donošenju odluka, narušavanje privatnosti putem nadzora i paradoksalni ekološki otisak AI računarstva. Da bismo ublažili ove rizike, predlažemo tri konkretne preporuke: uključivanje neantropocentričnih metrika u etičke standarde veštačke inteligencije; harmonizaciju prekograničnih regulatornih okvira radi usklađivanja sa globalnim standardima kao što je Akt o veštačkoj inteligenciji EU (EU AI Act); i obavezivanje na strogo definisane human-in-the-loop protokole (učešće čoveka u procesu odlučivanja) za sve autonomne intervencije u životnoj sredini.

Ključne reči: bezbednost životne sredine, autonomni AI sistemi, etika veštačke inteligencije, algoritamska regulacija, veštačka inteligencija usmerena na čoveka (HCAI), zaštita resursa.

Beyond the Battlefield: The Ethical Implications and Regulatory Challenges of Using Autonomous AI Systems for Environmental Security and Resource Protection

Abstract in English: As Artificial Intelligence (AI) transitions from military and industrial domains to environmental science, a fundamental shift toward data-driven methodologies is reshaping planetary protection. However, this transition frequently imports battlefield logic into conservation, utilizing autonomous systems—such as drones and machine learning algorithms—that introduce complex ethical and regulatory challenges. This paper presents a conceptual synthesis of Human-Centered AI (HCAI) frameworks and ecological security perspectives to address these risks. We identify critical friction points, including anthropocentric biases that neglect non-human wellbeing, a responsibility gap in autonomous decision-making, privacy infringements through surveillance, and the paradoxical environmental footprint of AI computing. To mitigate these risks, we propose three actionable recommendations: incorporating non-anthropocentric metrics into ethical AI standards; harmonizing transboundary regulatory frameworks to align with global standards like the EU AI Act; and mandating strictly defined human-in-the-loop protocols for all autonomous environmental interventions.

Keywords: Environmental Security, Autonomous AI Systems, AI Ethics, Algorithmic Regulation, Human-Centered AI (HCAI), Resource Protection.

1. Introduction

The integration of artificial intelligence (AI) into environmental sectors has catalyzed a profound shift in natural resource management, opening vast opportunities for improving efficiency, decision-making, and sustainability (Nizamani et al., 2025). This technological evolution is increasingly recognized as a vital mechanism for achieving the Sustainable Development Goals (SDGs), particularly regarding poverty alleviation, infrastructure development, and the protection of life on land (Chisom et al., 2024; Mhlanga, 2021; Vinuesa et al., 2020).

However, the application of these technologies is not value-neutral; it is contingent upon overarching security discourses that often mirror military logic (Francisco, 2023). As nations such as Kenya adopt AI and drones for major wildlife conservation reforms (The Standard, 2025), and conservation organizations deploy thermal cameras to protect rhinos (World Wildlife Fund, n.d.), we are witnessing a paradigm shift where the tools of the battlefield are being repurposed for environmental security. While these autonomous systems promise to maximize the potential of conservation data (Ahumada et al., 2019), they operate within a fragile ethical landscape. The rapid deployment of such agents raises critical questions regarding accountability and transparency (Bahrevar & Khorasani, 2021; Cheong, 2024), alongside the risk that algorithmic thinking may perpetuate hegemonizing knowledge or estrange us from ecological realities (Francisco, 2023).

Conservation AI requires governance that rejects militarized assumptions and centers ecological wellbeing, accountability, and transparent data practices. Section 2 outlines the methodology; Section 3 surveys capabilities and use cases; Section 4 analyzes ethical risks; Section 5 maps regulation to controls; and Section 6 concludes with practical recommendations.

2. Methodology

To analyze the intersection of autonomous systems, ethics, and environmental security, this paper employs a scoping review methodology complemented by illustrative case studies. This approach is designed to map rapidly evolving concepts across disparate fields—computer science, environmental law, and security studies—allowing for a comprehensive synthesis of current trends and policy gaps.

2.1. Search Strategy and Inclusion Criteria

Data collection prioritized high-quality sources published between 2019 and 2025. This timeframe was selected to capture the data-driven revolution in conservation and the emergence of significant regulatory frameworks such as the EU AI Act. The review integrates two primary categories of literature:

Peer-Reviewed Scholarship: Academic articles focusing on AI ethics, remote sensing, and ecological security to ensure theoretical rigor.

Reputable Policy & Technical Reports: Guidelines and white papers from authoritative bodies, including the Association for Computing Machinery (ACM), the European Commission, and major conservation NGOs (e.g., WWF, IUCN), to ground the analysis in real-world governance challenges.

2.2. Justification for Foundational Works

While the focus is on recent advancements, select foundational texts pre-dating 2019 are included to trace the lineage of current technologies. Works such as Hernandez (1990) on expert systems in law enforcement and Horswill (1995) on specialized real-time systems provide critical historical context, demonstrating how the current application of security logic to nature has deep roots in early computing and military adaptation.

2.3. Analysis and Synthesis

The selected literature was synthesized thematically to identify ethical friction points and regulatory voids. To bridge the gap between abstract theory and practice, the paper utilizes illustrative cases of deployed technologies—specifically the Wildlife Insights and EarthRanger platforms. These cases serve not as empirical data points, but as tangible examples of how autonomous agents are currently operationalizing security in the wild.

3. The Rise of Autonomous Environmental Security

Current environmental security technologies operate across a spectrum of autonomy, evolving from passive observation to active intervention. We propose the following taxonomy to categorize these systems:

1. Sensing: Passive data collection via camera traps, acoustic sensors, and thermal imaging.
2. Analytics: Machine learning models that perform classification (species identification) and prediction (poaching risk analysis).
3. Actuation: Autonomous or semi-autonomous physical agents, such as drones (UAVs) used for patrols or interdiction.
4. Decision Platforms: Integrated dashboards like EarthRanger that aggregate sensor data to direct human or automated responses.

3.1. Capabilities and Mini-Cases

Historically, environmental monitoring was characterized by significant time lags. Today, platforms maximize data potential through automation.

- Case A: Wildlife Insights (Analytics). This platform utilizes AI to process millions of camera trap images, transforming raw data into actionable biodiversity metrics and reducing analysis time from months to minutes (Ahumada et al., 2019).
- Case B: EarthRanger (Decision Support). This system integrates real-time data from radios, animal trackers, and sensors into a domain awareness dashboard, allowing rangers to deploy assets efficiently (Wall et al., 2024).
- Case C: Predictive Hotspot Alerts (Prediction). AI models analyze historical poaching data to predict crime hotspots. However, these incident histories often reflect biases from previous patrol patterns or reporting inequities rather than absolute crime distribution. Such systemic biases necessitate the rigorous audit trails and RACI accountability frameworks discussed later in this study to ensure fairness in deployment (Sustainability Directory, 2024).

3.2. The Intersection with Organized Crime

The militarization of conservation is partly a response to the sophistication of environmental crime. In Africa, the intersection of AI and organized crime presents a significant threat, necessitating advanced technological countermeasures. Specialized real-time systems are required to navigate these hostile environments. Yet, as Francisco notes (Francisco , 2022), this national security discourse can emphasize military uses of AI, potentially overshadowing human and ecological security needs.

4. Ethical Implications: The Logic of the Battlefield in Nature

This section explores the ethical friction points created when autonomous systems are deployed in complex social and ecological environments.

4.1. Anthropocentrism and Ecological Metrics

A major ethical deficiency in current AI deployment is its anthropocentric focus. AI ethics standards often prioritize human wellbeing while neglecting environmental wellbeing. This creates systemic vulnerability, where the rigid logic of AI systems may fail to grasp the complexity of biological ecosystems. To counter this, governance must include ecological metrics alongside human performance indicators. For example, while a persistent drone patrol may successfully reduce poaching, the acoustic disturbance could increase stress levels in sensitive non-target species. Consequently, projects should explicitly declare and monitor non-target species stress proxies, such as flight initiation distance and acoustic load, to ensure these trade-offs are effectively measured and governed.

4.2. The Responsibility Gap and RACI Mapping

The deployment of autonomous agents introduces a responsibility gap. If an autonomous system harms a local community member or makes an erroneous ecological decision (e.g., culling the wrong animals), determining

accountability is difficult. To operationalize accountability, we propose a **RACI (Responsible, Accountable, Consulted, Informed)** mapping for the AI lifecycle:

- **Responsible (The "Doer"):** The Field Operations Lead ensures the AI is deployed according to protocol.
- **Accountable (The "Owner"):** The Project Principal Investigator (PI) or Park Warden retains ultimate liability for system outcomes. This role is specifically responsible for signing off on Fundamental Rights Impact Assessments (FRIAs) and post-incident reviews to ensure strict alignment with regulatory controls.
- **Consulted (Two-way communication):** Local Community Boards provide input on sensor placement and data privacy.
- **Informed (One-way communication):** Oversight Ministries receive regular audit reports.

4.3. Surveillance, Privacy, and Data Sovereignty

Surveillance technologies used for conservation can infringe on the collective right to access information and the privacy of local populations. Without ethical data governance, smart city surveillance tactics applied to nature reserves can lead to data colonialism, where information about local resources is extracted without local consent. Governance frameworks must adopt data sovereignty principles. This includes purpose limitation and the explicit operational rule that data collected for wildlife monitoring cannot be repurposed for unrelated commercial use without renewed consent.

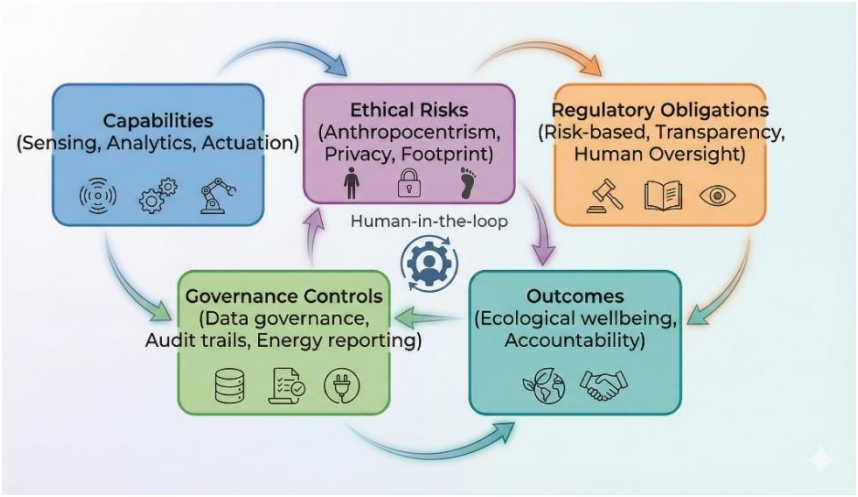
4.4. The Environmental Footprint of AI

Paradoxically, the AI systems used to protect the environment have a significant ecological footprint. Recent analyses indicate that most companies ignore the environmental impact of their AI operations. To be truly human-centered and sustainable, AI hardware must adhere to energy efficiency requirements. Practical controls include adaptive scheduling—retraining models only during off-peak energy hours—and favoring efficient edge-computing hardware over massive cloud reliance. Furthermore, energy and emissions reporting should be formally integrated into project documentation and subject to periodic audits.

5. Regulatory Challenges and Governance Frameworks

See Figure 1 for an overview of how capabilities map to risks, obligations, and controls.

Figure 1 (The HCAI × Ecological Security Framework)



Source: Visualization generated by Gemini (Google) based on conceptual frameworks and technical guidance provided by the authors.

The regulation of autonomous environmental systems is currently fragmented. While the United States moves toward a national framework, the **EU AI Act** represents the first comprehensive attempt to categorize AI based on risk. For environmental security, which often involves transboundary resources, this fragmentation is

problematic. To operationalize intelligent regulation, we map the EU AI Act's risk categories to specific conservation use cases (Table 1).

To combat the opacity of proprietary black box algorithms, there is a growing movement toward open-source governance in environmental projects. However, transparency alone is insufficient; systems must be context-aware and accompanied by clear audit trails.

Table 1: Applying the EU AI Act[1] Risk Framework to Environmental Security

Risk Category	Conservation Use Case	Key Obligations	Practical Controls & Governance
Unacceptable Risk	**Social Scoring of Communities:** Using AI to rank local villagers based on resource usage or movement patterns.	**Prohibited.**	**Strict Ban:** Explicit policies forbidding the use of conservation data for social credit systems.
High Risk	**Predictive Policing & Interdiction:** Autonomous drones identifying "biometric or identity□linkable analytics in high□risk interdiction contexts" or predicting crime hotspots (Sustainability Directory, 2024).	**Strict Compliance:** Fundamental Rights Impact Assessment (FRIA), high data quality, human oversight.	**Human-in-the-Loop:** No autonomous engagement; human rangers must verify AI alerts. **Audit Trails:** Logs of all AI predictions vs. outcomes.
Limited Risk	**Biometric Monitoring (Wildlife):** Automated species identification via camera traps (e.g., Wildlife Insights).	**Transparency:** Users must know they are interacting with AI (if applicable).	**Data Governance:** Open-source verification of species classifiers to prevent bias (Meegle, 2025).
Minimal Risk	**Environmental Modeling:** AI used for analyzing soil samples or deforestation trends (Gizachew, 2025).	**Voluntary Codes of Conduct.**	**Open Science:** Sharing models and datasets to foster trust and reproducibility.

6. Conclusion

The transition of AI beyond the battlefield into environmental security offers a powerful mechanism for resource protection. However, as this paper has demonstrated, it is fraught with ethical and regulatory perils. Drawing on the Human-Centered AI framework and ecological security perspectives, it is clear that technological solutions cannot be divorced from their socio-political contexts.

To ensure these systems are sustainable, we recommend:

1. *Ecological Ethics:* Incorporate non-anthropocentric metrics into AI standards to respect the wellbeing of the ecosystem itself.
2. *Harmonized Regulation:* Align transboundary monitoring projects with global standards like the EU AI Act to prevent regulatory arbitrage.
3. *Mandatory Human Oversight:* Enforce strictly defined human-in-the-loop protocols for all high-risk autonomous interventions.

Future research should prioritize operationalizing participatory audits, developing indigenous and community-led data governance models, and launching cross-border interoperability pilots to harmonize transboundary resource protection. Only by addressing these challenges can we harness the power of AI to protect our planet without replicating the logic of conflict that birthed these technologies.

Appendix: Operational Governance Toolkit
A. Fundamental Rights Impact Assessment (FRIA) for Environmental AI
This assessment is mandatory for "High-Risk" systems, such as predictive policing or autonomous interdiction drones, to ensure compliance with human rights and ecological integrity, as mandated by emerging standards like the EU AI Act.

[1] https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (This regulation establishes the harmonized rules and risk-based framework mentioned in your paper's analysis of environmental security governance. As outlined in your manuscript, this act categorizes AI systems into risk levels (unacceptable, high, limited, and minimal), which serves as the foundation for the "Risk-to-Control" mapping used in your study.)

• **Project Title & ID: [e.g., Project Rhino-Shield Alpha]**
• **System Description: Briefly describe the autonomous agent, its sensing capabilities, and the level of autonomy in actuation.**
• **Intended Purpose: Define the specific environmental security goal (e.g., identifying illegal logging in Sector 7).**
• **Targeted Populations: Identify human communities (indigenous groups, local villagers) and non-target biological species potentially impacted by the system's presence.**
• **Risk Assessment:**
○ **Privacy & Surveillance: Does the system capture identity-linkable data that may infringe on the collective right to access information?**
○ **Algorithmic Bias: Are the training datasets representative of local realities to prevent environmental injustice?**
○ **Ecological Disturbance: What is the predicted acoustic load or physical intrusion for sensitive non-target species?**
• **Mitigation Strategy: Describe the human-in-the-loop protocols and privacy masking techniques deployed to maintain meaningful human control.**
B. Environmental AI Model Card (Short Form)
A standardized document to combat the opacity of "black box" proprietary algorithms by detailing the model's performance, limitations, and normative groundings.
• **Model Developer: [e.g., University Conservation Tech Lab]**
• **Model Version & Date: [e.g., v2.1, October 2025]**
• **Model Type: Specify the architecture (e.g., Convolutional Neural Network for species classification).**
• **Training Data: List primary datasets used and their origin to ensure data sovereignty.**
• **Performance Metrics: Accuracy, precision, and recall rates across varied environmental conditions (e.g., night-time vs. day-time performance).**
• **Ecological Constraints: Explicitly state environmental conditions where the model fails (e.g., high failure rate during heavy tropical rainfall).**
C. Dataset Information Sheet
Ensures transparency in data governance and prevents "data colonialism" by documenting the origin and consent status of data.
• **Data Source: (e.g., Community-monitored sensors in Reserve X).**
• **Collection Method: (e.g., Automated passive acoustic monitoring).**
• **Consent & Sovereignty Status: Confirmation that local community boards have signed off on data collection processes.**
• **Purpose Limitation Rule: Explicit declaration that data collected for wildlife monitoring cannot be repurposed for unrelated commercial use without renewed community consent.**
• **Data Retention: Schedule for when identity-linkable analytics will be anonymized or deleted.**
D. Operational Accountability & Sustainability Logs
These logs provide a transparent audit trail for regulators and oversight bodies to bridge the responsibility gap.
D.1 Incident Response Log (Audit Trail)
• **Timestamp: [YYYY-MM-DD HH:MM]**
• **Autonomous Alert: (e.g., "Potential human intrusion detected in restricted zone").**
• **Human Verification: (e.g., Ranger verified via thermal link).**
• **Outcome: (e.g., "Confirmed false positive – local resident collecting fallen wood; system updated to recognize traditional tool signatures").**
• **Sign-off: [Ranger ID] to ensure individual and collective accountability.**
D.2 Energy & Emissions Log (Sustainability Reporting)
• **Reporting Period: [e.g., Q4 2025]**
• **Compute Energy Usage: Total kWh consumed for model training and edge-cloud operations.**
• **Carbon Footprint: Estimated CO_2e based on the local energy grid and hardware lifecycles.**
• **Efficiency Measures: (e.g., "Model retraining restricted to off-peak renewable energy hours").**

Statement

During the preparation of this work the author(s) used Gemini (Google) in order to assist with the visualization of conceptual frameworks and to provide suggestions for the structural organization of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

1. Ahumada, J. A., Fegraus, E., Birch, T., Flores, N., Kays, R., O'Brien, T. G., … Dancer, A. (2019). Wildlife Insights: A Platform to Maximize the Potential of Camera Trap and Other Passive Sensor Wildlife Data for the Planet. Environmental Conservation, 47 (1), 1–6. doi: 10.1017/S0376892919000298
2. Association for Computing Machinery. (2018). ACM Code of Ethics and Professional Conduct. New York: ACM Council.
3. Bahrevar, R. & Khorasani, K. (2021). Accountability and Transparency in AI Systems: A Public Policy Perspective. (Unpublished thesis). Concordia University, Montreal, Canada.
4. Bird, E., Fox-Skelly, J., Jenner, N., Larbey, R., Weitkamp, E. & Winfield, A. (2020). The Ethics of Artificial Intelligence: Issues and Initiatives. European Parliament Panel for the Future of Science and Technology (STOA). doi: 10.2861/6644
5. Buiten, M. (2019). Towards Intelligent Regulation of Artificial Intelligence. European Journal of Risk Regulation, 10 (1), 41–59. doi: 10.1017/err.2019.8
6. Cheong, B. C. (2024). Transparency and Accountability in AI Systems: Safeguarding Wellbeing in the Age of Algorithmic Decision-Making. Frontiers in Human Dynamics, 6, 1421273. doi: 10.3389/fhumd.2024.1421273
7. Chisom, O. N., Biu, P. W., Umoh, A. K., Obaedo, B. O., Adegbite, A. O. & Abatan, A. (2024). Reviewing the Role of AI in Environmental Monitoring and Conservation: A Data-Driven Revolution for our Planet. World Journal of Advanced Research and Reviews, 21 (01), 161–171.
8. Conservation International. (2020). Wildlife Insights Platform Feature Build-out, Support and Maintenance. Retrieved on 23 December 2025, from [Insert URL if available]
9. Crowell & Moring. (Days 19 December 2025). Executive Order Tries to Thwart "Onerous" AI State Regulation, Calls for National Framework. [Journal/Portal Name].
10. Dalton, D., Berger, V., Kirchmeir, H., Adams, V., Botha, J., Halloy, S., … Jungmeier, M. (2024). A Framework for Monitoring Biodiversity in Protected Areas and Other Effective Area-Based Conservation Measures: Concepts, Methods and Technologies. IUCN WCPA Technical Report Series No. 7. Gland: IUCN. doi: 10.2305/HRAP7908
11. Data Privacy Brasil. (Days 31 October 2025). Artificial Intelligence, Sustainability and the Collective Right to Access Information. Retrieved on 23 December 2025, from
12. ENACT Africa. (Days 29 June 2023). AI and Organised Crime in Africa. Retrieved on 23 December 2025
13. European Commission. (2025). Answers for WildEuro - MSCA 101107666. Documents Download Module.
14. European Parliament. (2023). EU AI Act: First Regulation on Artificial Intelligence. Retrieved on 23 December 2025, from https://www.europarl.europa.eu/
15. Francisco, M. (2023). Artificial Intelligence for Environmental Security: National, International, Human and Ecological Perspectives. Current Opinion in Environmental Sustainability, 61, 101250. doi: 10.1016/j.cosust.2022.101250
16. Gaulin, M. (2024). New Analysis Finds Most Companies Ignore AI Environmental Impact. Lab Manager. Retrieved on 23 December 2025
17. Gizachew, B. (2025). AI and Machine Learning in Remote Sensing for Tropical Forest Monitoring: Applications, Challenges, and Emerging Solutions. Preprints.org. doi: 10.20944/preprints202512.1554.v1
18. Global Affairs Canada. (2019, June). Ethical and Methodological Framework for Open Source Data Monitoring and Analysis. Ottawa: Government of Canada.
19. Goel, P. K., Saifi, S., Goel, N. & Aeron, S. (2025). Ethical Considerations in AI: Applications for Wildlife Conservation. In AI and Machine Learning Techniques for Wildlife Conservation (pp. 247–266). Hershey: IGI Global. doi: 10.4018/979-8-3693-6935-7.ch010

20. Gupta, A. & Kaur, R. (2025). Role of AI Powered Drones and Satellite Imagery in Detecting Poaching Activities. International Journal of Research Publication and Reviews, 6 (10), 7097–7110. doi: 10.55248/gengpi.06.1025.3836
21. Hernandez, A. P. (1990). Artificial Intelligence and Expert Systems in Law Enforcement: Current and Potential Uses. Computers, Environment and Urban Systems, 14, 299–306.
22. Hohma, E., Boch, A., Trauth, R. & Lütge, C. (2023). Investigating Accountability for Artificial Intelligence Through Risk Governance: A Workshop-Based Exploratory Study. Frontiers in Psychology, 14, 1073686. doi: 10.3389/fpsyg.2023.1073686
23. Horswill, I. (1995). Analysis of Specialized Real-Time Systems. Artificial Intelligence, 73 (1–2), 1–30.
24. Johann Heinrich von Thünen Institute. (2025). Assessing the Suitability of Available Global Forest Maps as Reference Tools for EUDR-compliant Deforestation Monitoring. Remote Sensing, 17 (17), 3012. doi: 10.3390/rs17173012
25. Jones, R. K. (2026). Ethical Considerations in Deploying Autonomous AI. In Safeguarding and Securing Autonomous AI Agents (pp. 133–170). Hershey: IGI Global Scientific Publishing. doi: 10.4018/979-8-3373-6876-4.ch005
26. Kingston, J. (2017). Using Artificial Intelligence to Support Compliance with the General Data Protection Regulation. International Journal of Law and Information Technology, 25 (3), 226–243.
27. Loisaba Conservancy. (n.d.). Security. Retrieved on 23 December 2025, from https://www.loisaba.com/
28. Maitra, S., Lang, L. & Hernández Jurado, M. (n.d.). How Shifting Responsibility for AI Harms Undermines Democratic Accountability. Retrieved on 23 December 2025
29. Meegle. (Days 23 October 2025). Open-source Governance in Environmental Projects. Retrieved on 23 December 2025
30. Mhlanga, D. (2021). Artificial Intelligence in the Industry 4.0, and its Impact on Poverty, Innovation, Infrastructure Development, and the Sustainable Development Goals: Lessons from Emerging Economies? Sustainability, 13, 5788. doi: 10.3390/su13115788
31. Microavia. (n.d.). How Drones are Used in Wildlife Monitoring to Protect Against Poaching. Retrieved on 23 December 2025, from https://microavia.com/
32. Moreno, N. & McAllister Novak, A. (Days 21 January 2025). Key Insights into AI Regulations in the EU and the US: Navigating the Evolving Landscape. Kennedys Law LLP.
33. Nizamani, M. M., Zhang, H. L. & Lai, Z. (2025). Human-centered AI: Advancing Ethical, Transparent, and Context-aware Systems for Sustainable Development. Technology in Society, 80, 103121. doi: 10.1016/j.techsoc.2025.103121
34. NVIDIA. (2025). NVIDIA Sustainability Report Fiscal Year 2025. Retrieved on 23 December 2025
35. Ojija, F., Ogwu, M. C., Ally, J., John, J. P., Stephano, A., Felix, N. & Tekka, R. (2025). Artificial Intelligence-driven Solutions for Mitigating Human–Wildlife Conflict in Biodiversity Hotspots. Public Health Reviews, 108 (4). doi: 0.1177/00368504251394584
36. Prokopowicz, D. (2025). The Use of Big Data and Artificial Intelligence in Protecting the Climate and Biodiversity of Planet Earth. Warsaw: Cardinal Stefan Wyszynski University. doi: 10.13140/RG.2.2.23721.25446
37. Rigley, E., Chapman, A., Evers, C. & McNeill, W. (2023). Anthropocentrism and Environmental Wellbeing in AI Ethics Standards: A Scoping Review and Discussion. AI, 4, 844–874. doi:10.3390/ai4040043
38. Scoble, R. & Cronin, I. (Days 24 June 2025). How AI is Revolutionizing Wildlife Conservation. Unaligned Newsletter.
39. Serry, E. (2025). Ethical Implications of Artificial Intelligence: Challenges, Risks, and Regulatory Perspectives. (Unpublished thesis). Teesside University, UK. doi: 10.13140/RG.2.2.11350.56645
40. Shahrouz, M., Nazari, A. & Moradi, S. (2023). AI-driven Cybersecurity: Legal and Ethical Considerations in Autonomous Systems Protecting Digital Networks. Legal Studies in Digital Age, 2 (1), 1–12.
41. Stahl, B. C. (2023). Responsible AI: From Principles to Implementation and Regulation. Scientific Reports, 13, 34622. doi: 10.1038/s41598-023-34622-w
42. Stahl, B. C. (2025). The Ethics of Data and its Governance: A Discourse Theoretical Approach. Information, 16 (6), 497. doi: 10.3390/info16060497
43. Sustainability Directory. (2024a). AI for Wildlife Crime Prediction. Retrieved on 23 December 2025

44. Sustainability Directory. (2024b). Ethical Data Governance. Retrieved on 23 December 2025
45. Sustainability Directory. (2024c). Ethical Frameworks for Smart City Surveillance. Retrieved on 23 December 2025
46. Sustainability Directory. (2024d). To what Extent Does Data Privacy Relate to Global Sustainability Goals? Retrieved on 23 December 2025
47. Sustainability Directory. (n.d.). Ethical Implications of AI in Water Resource Management. Retrieved on 23 December 2025
48. Terenzio, F. (2025). Systemic Vulnerability: From AI Systems to Environmental Systems. Topoi. doi: 10.1007/s11245-025-10285-2
49. The Standard. (Days 23 December 2025). Kindiki: Kenya Adopts AI, Drones in Major Wildlife Conservation Reforms. The Standard.
50. TRENDS Research & Advisory. (2024). The Backlash Against Military AI: Public Sentiment, Ethical Tensions, and the Future of Autonomous Warfare. Abu Dhabi: TRENDS.
51. University of Cambridge. (n.d.). AI and Conservation Resolution Adopted at the IUCN World Conservation Congress. Retrieved on 23 December 2025
52. VERSO. (n.d.). Ethics - Introduction. Open Resource Library. Retrieved on 23 December 2025
53. Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., … Fuso Nerini, F. (2020). The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. Nature Communications, 11, 233. doi: 10.1038/s41467-019-14108-y
54. Wall, J., Lefcourt, J., Jones, C., Doehring, C., O'Neill, D., Schneider, D., … Wittemyer, G. (2024). EarthRanger: An Open-source Platform for Ecosystem Monitoring, Research and Management. Methods in Ecology and Evolution, 15 (11), 1968–1979. doi: 10.1111/2041-210X.14399
55. White & Case LLP. (2025). Energy Efficiency Requirements Under the EU AI Act. Retrieved on 23 December 2025, from https://www.whitecase.com/
56. Winter, J. S. & Davidson, E. (2019). Governance of Artificial Intelligence and Personal Health Information. Digital Policy, Regulation and Governance. doi: 10.1108/DPRG-08-2018-0048
57. World Resources Institute. (2025a). Terms of Service - WRI Data Platforms. Retrieved on 23 December 2025, from https://www.wri.org/
58. World Resources Institute. (2025b). WRI's Approach to Responsible Artificial Intelligence. Washington: WRI.
59. World Wildlife Fund. (n.d.). Thermal Cameras and AI Help Protect Rhinos in Kenya. Retrieved on 23 December 2025, from https://www.worldwildlife.org/
60. Yehudi, Y., Bennett, A., Turon, G., Bays, D., Gibson, S., Druskat, S., … Batchelor, S. (2022). Ethical Considerations When Choosing an Open Source Governance Model. The Turing Way. doi: 10.5281/zenodo.6144158

Original Scientific Paper/Original naučni rad UDC/UDK: 004.85:[005.334:620.9
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601070A

Primena mašinskog učenja u industrijskoj bezbednosti: digitalne inovacije i nove bezbednosne paradigme u rafinerijskim procesima

Luka Abramović[1], Jelena Raut[2]

[1]Total Energies Refinery, Antwerpen, Belgium, luka.abramovic@external.totalenergies.com
[2] School of Engineering Management, University "Union – Nikola Tesla", Belgrade, Serbia, jelena.raut@fim.rs

Summary in Serbian: U savremenim rafinerijskim postrojenjima, složeni industrijski procesi zahtevaju visok nivo bezbednosti kako bi se smanjili rizici po ljudske živote, imovinu i životnu sredinu. Tradicionalni sistemi nadzora i sigurnosne procedure sve češće pokazuju ograničenja u detekciji anomalija i predviđanju potencijalnih incidenata. Primena mašinskog učenja (ML) u industrijskoj bezbednosti omogućava transformaciju bezbednosnih paradigma kroz analizu velikih količina podataka sa senzora, SCADA sistema i drugih industrijskih izvora. Ovaj rad istražuje aktuelne digitalne inovacije u oblasti ML algoritama za prediktivnu analitiku, automatsko prepoznavanje anomalija i optimizaciju sigurnosnih procedura u rafinerijskim procesima. Poseban fokus je stavljen na integraciju ML modela u postojeće sisteme upravljanja rizicima, izazove u implementaciji i mogućnosti unapređenja industrijske bezbednosti kroz proaktivni pristup. Rezultati istraživanja ukazuju da primena mašinskog učenja značajno doprinosi smanjenju incidenata, poboljšava efikasnost bezbednosnih operacija i otvara nove perspektive za digitalnu transformaciju u industrijskoj bezbednosti.

Keywords: mašinsko učenje, industrijska bezbednost, rafinerijski procesi, digitalne inovacije, prediktivna analitika

Application of Machine Learning in Industrial Safety: Digital Innovations and New Safety Paradigms in Refinery Processes

Abstract in English: In modern refinery facilities, complex industrial processes require a high level of safety to minimize risks to human lives, property, and the environment. Traditional monitoring systems and safety procedures increasingly show limitations in detecting anomalies and predicting potential incidents. The application of machine learning (ML) in industrial safety enables a transformation of safety paradigms through the analysis of large volumes of data from sensors, SCADA systems, and other industrial sources. This paper explores current digital innovations in ML algorithms for predictive analytics, automatic anomaly detection, and optimization of safety procedures in refinery processes. Focus is placed on integrating ML models into existing risk management systems, implementation challenges, and opportunities for enhancing industrial safety through a proactive approach. The research results indicate that the application of machine learning significantly contributes to reducing incidents, improving the efficiency of safety operations, and opening new perspectives for digital transformation in industrial safety.

Keywords: machine learning, industrial safety, refinery processes, digital innovations, predictive analytics

1. Introduction

Industrial safety represents one of the fundamental pillars of sustainable and reliable operation of refinery facilities. Oil and gas refineries belong to the most complex industrial systems due to a high level of technological integration, continuous process flows, and the handling of flammable, toxic, and highly reactive substances. Any disruption in the operation of such systems can have serious consequences for employee safety, the integrity of facilities, and the environment, as well as significant economic and reputational losses for the organization. Therefore, industrial safety in refinery processes has traditionally been regarded as a strategic priority within risk and process management. Traditional industrial safety systems in refineries are based on a combination of

procedural measures, engineering safeguards, and conventional monitoring systems, such as alarm systems, protective logic, periodic inspections, and incident analysis based on historical data (Omoarebun, 2023). Although these systems have formed the foundation of safety practice for decades, contemporary industrial conditions reveal their limitations. Above all, traditional approaches are predominantly reactive, rely on predefined thresholds and rules, and are often unable to identify complex system behavior patterns that preceded incidents. In addition, the growing volume of data generated by modern SCADA and IIoT systems further complicates timely analysis and decision-making when relying solely on conventional methods. In the context of industrial digital transformation and the development of the industry 4.0 paradigm, there is an increasing need for more advanced, adaptive, and predictive approaches to industrial safety. Machine learning (ML), as a subfield of artificial intelligence, enables the analysis of large volumes of heterogeneous data in real time, the identification of hidden patterns, and the prediction of potential anomalies before they escalate into serious safety incidents (Sarker, 2023). The application of ML algorithms creates the opportunity to transition from a reactive to a proactive safety management model, in which risks are identified and mitigated at early stages of their emergence. The motivation for applying machine learning in refinery processes arises from the need to increase system reliability, reduce the number of false alarms, improve decision-making, and achieve more efficient use of available resources. ML models, such as anomaly detection algorithms, predictive analytics, and incident classification techniques, enable continuous learning from operational data and adaptation to changing operating conditions (Omol et al., 2024). In this way, safety systems become more intelligent, resilient, and capable of responding to the complex and dynamic risks characteristic of the refinery industry. The aim of this paper is to analyze the role of machine learning in enhancing industrial safety in refinery processes, with a particular focus on digital innovations and new safety paradigms emerging from their application. The paper seeks to present the possibilities for integrating ML algorithms into existing safety management systems, identify implementation challenges, and highlight potential benefits in terms of incident reduction, increased operational efficiency, and overall improvement of safety levels. The scientific contribution of this paper lies in the systematization of contemporary approaches to the application of machine learning in industrial safety, while its practical contribution is reflected in providing guidelines for their implementation in real refinery environments.

2. Industrial Safety and Digital Transformation of Refinery Systems

Industrial safety in refinery facilities represents an interdisciplinary field encompassing engineering, organizational, and technological aspects aimed at preventing undesirable events and minimizing their consequences. Contemporary refinery systems are characterized by a high degree of automation, continuous processes, and complex interdependencies among technological units, making safety one of the key factors of operational reliability (Olaizola et al., 2022). In such an environment, the theoretical framework of industrial safety must be considered through the lens of process safety, digital transformation, and the application of advanced analytical methods based on machine learning.

Process safety constitutes the fundamental concept of industrial safety in refinery processes and refers to the identification, control, and management of risks arising from the handling of hazardous substances and energy flows (Klein and Vaughen, 2017). Unlike occupational safety, which is primarily focused on the individual protection of employees, process safety is oriented toward the integrity of the entire system, including equipment, processes, control systems, and organizational procedures. Its primary purpose is the prevention of major industrial accidents that may have catastrophic consequences for people, assets, and the environment. Refinery facilities are particularly exposed to risks due to high temperatures and pressures, the presence of flammable and toxic substances, and complex chemical reactions occurring in real time. Typical risks in refinery processes include leaks of hazardous substances, fires and explosions, mechanical equipment failures, as well as risks associated with the human factor (Delshah et al., 2023). Gas or liquid leaks often represent the initiating event that can escalate into a serious incident if not detected in a timely manner. Explosions and fires result from a combination of technical failures and inadequate process control, while the human factor includes operational errors, incorrect assessments, and non-compliance with procedures. Traditional process safety systems rely on methods such as HAZOP analysis, Failure Modes and Effects Analysis (FMEA), alarm systems, and safety instrumented systems. Although these methods remain essential, their effectiveness is limited in dynamic environments where process parameters continuously change and the volume of data grows exponentially (Kim et al., 2018). These limitations create opportunities for integrating advanced digital solutions that can enhance existing safety mechanisms.

The digital transformation of industry, embodied in the Industry 4.0 concept, has brought significant changes in the way refinery processes are designed, managed, and monitored. The introduction of smart sensors, the Industrial

Internet of Things (IIoT), and advanced data acquisition and processing systems has enabled the creation of highly interconnected and information-rich industrial environments. SCADA systems represent a central element of refinery digital infrastructure, enabling real-time process monitoring and remote control of key technological parameters. In addition to SCADA systems, IIoT technologies enable continuous data collection from a large number of sensors distributed along production lines, storage tanks, and critical equipment (Babayigit and Abubaker, 2023). These data, combined with historical records of system operation and incidents, form the basis for the development of Big Data analytics in industrial safety. However, data availability alone is not sufficient to improve safety; it is necessary to apply advanced analytical methods capable of extracting relevant information from large and heterogeneous datasets. Digital innovations within the industry 4.0 framework enable a transition from static safety models to dynamic and adaptive systems capable of adjusting to changes in processes and the operating environment (Korytko and Piletska, 2022). In this context, industrial safety becomes an integral part of the digital strategy of refinery companies, where safety aspects are considered in parallel with system efficiency, reliability, and sustainability.

Machine learning represents a key technology enabling intelligent processing of data generated in digitally transformed refinery systems. In industrial applications, ML algorithms are used for pattern recognition, event classification, fault prediction, and anomaly detection in system operation. Depending on the availability of labeled data, supervised and unsupervised machine learning models are applied. Supervised learning relies on historical data with clearly defined outcomes, such as recorded incidents, failures, or alarm events (Sun et al., 2022). These models enable the classification of risk states and the prediction of incident likelihood based on current process parameters. On the other hand, unsupervised learning is particularly suitable for anomaly detection in refinery processes, where incidents are rare and data are often unlabeled. Algorithms such as clustering and autoencoders enable the identification of deviations from normal operating regimes, allowing potential problems to be detected at early stages. The application of machine learning in industrial safety enables the development of predictive safety models that overcome the limitations of traditional approaches. Instead of relying on fixed thresholds and rules, ML systems continuously learn from data and adapt to changes in processes. In this way, it is possible to reduce the number of false alarms, increase the reliability of critical state detection, and improve real-time decision-making. In refinery processes, where timely response is crucial, this approach represents a significant step toward new safety paradigms based on proactive risk management.

3. Application of Machine Learning in Refinery Safety

The application of machine learning in refinery safety represents the practical operationalization of digital innovations within high-risk industrial systems. Unlike theoretical models, the real-world implementation of ML solutions in refineries must be adapted to process complexity, data availability, and existing safety architectures. In this context, effective application of machine learning requires careful consideration of data sources, algorithm selection, and their integration into existing safety and risk management systems.

The foundation for applying machine learning in refinery safety lies in the data continuously generated during plant operation (Erinjogunola et al., 2020). Modern refineries are equipped with a large number of sensors measuring process parameters such as temperature, pressure, flow, level, vibration, and chemical composition. These sensors are integrated into SCADA systems that enable real-time data acquisition, storage, and visualization. SCADA systems represent the central source of operational data and a key element for applying ML algorithms for safety purposes (Enemosah and Ifeanyi, 2024). Data from SCADA systems include both real-time process parameter values and records of alarms, shutdowns, and operator interventions. Analysis of these data enables the identification of system behavior patterns that precede abnormal states or incidents. In addition to operational data, historical data on incidents, failures, and safety events are of exceptional value for the development of predictive models. These data typically originate from safety management systems, incident reports, maintenance databases, and audit records. Although incidents in refineries are relatively rare, their analysis enables the training of supervised learning models capable of recognizing risk patterns and estimating the likelihood of similar events recurring. A key challenge in working with these data sources lies in their heterogeneity, incompleteness, and differing temporal resolutions. Therefore, data preprocessing is essential and includes noise filtering, time-series synchronization, and identification of relevant features to ensure the reliability of ML models under real refinery operating conditions.

Various ML algorithms are used in refinery safety systems depending on the analysis objectives, data availability, and requirements for interpretability. One commonly used algorithm is Random Forest, which has proven effective in classifying risk states and predicting incidents (Zhen et al., 2023). The advantage of this algorithm lies in its robustness to data noise and its ability to handle many input variables, which is particularly important

in complex refinery processes. Support Vector Machines (SVM) are applied in cases where precise separation between normal and abnormal process states is required (Cuentas et al., 2017). In refineries, SVM algorithms are used to detect deviations in the operation of critical equipment such as compressors, reactors, and distillation columns. Their ability to operate in high-dimensional spaces makes them suitable for analyzing complex datasets generated by SCADA systems. Neural networks, particularly deep neural networks, enable modeling of nonlinear relationships among process parameters, which are common in refinery systems. Their application includes fault prediction, analysis of system behavior under boundary conditions, and simulation of scenarios that may lead to incidents. Although these models are often less interpretable, their high accuracy makes them extremely valuable in safety applications where early detection is critical. Autoencoders, as a form of unsupervised learning, are especially significant for anomaly detection in refinery processes (Zheng and Zhao, 2020). These models learn normal system operating patterns and identify deviations that may indicate leaks, equipment degradation, or unforeseen process changes. In practice, autoencoders provide early warnings of potential safety risks even in situations where no historical data on similar incidents exist.

Effective application of machine learning in refinery safety requires its integration into existing risk management and decision-making systems (Erinjogunola et al., 2020). ML models do not function as standalone solutions but rather as support tools for traditional safety mechanisms. Integration into Risk Management Systems enables risk quantification based on predictive analyses and dynamic updating of safety assessments in real time. Within Decision Support Systems, ML models provide operators and management with additional information for decision-making, such as incident probability estimates, recommendations for preventive measures, and maintenance prioritization (Arinze et al., 2024). Such systems help reduce subjectivity in decision-making and increase the consistency of safety decisions in complex situations. An important aspect of integration also concerns compliance of ML solutions with regulatory requirements and safety standards applicable in the refinery industry. Model transparency, explainability of results, and system reliability are key factors for acceptance in industrial practice. When properly integrated, ML systems become an integral part of the refinery safety architecture and enable a transition toward a proactive and predictive model of industrial safety.

4. Challenges and Limitations of Implementing Machine Learning in Refinery Safety

Although the application of machine learning in refinery safety offers significant opportunities to improve risk management processes and incident prevention, its implementation in real industrial environments faces numerous challenges and limitations. These challenges are not exclusively technical in nature; they also encompass organizational, regulatory, and security aspects that can significantly affect the effectiveness and sustainability of ML solutions in refinery systems.

One of the key challenges relates to the quality and availability of data used for training and validating ML models. Although modern refineries generate large volumes of data through sensors and SCADA systems, these data often contain noise, missing values, or inconsistencies resulting from diverse sources and technologies (Olaizola et al., 2022). An additional issue is the relative rarity of major incidents, which makes it difficult to collect representative datasets for supervised learning. Under such conditions, ML models may become biased or insufficiently generalized, thereby reducing their reliability in detecting real safety risks.

Data security and cyber risks represent another significant challenge in the implementation of ML systems in refinery facilities. Integrating ML models into industrial networks requires access to sensitive operational data, which increases system exposure to cyberattacks (Pani and Soofastaei, 2025). Potential attacks on SCADA systems, data manipulation, or compromise of ML models can have serious consequences for process safety. Therefore, it is essential to ensure high standards of information security, including network segmentation, data encryption, and continuous system monitoring, in order to minimize cyber risks associated with the use of advanced analytical technologies.

Model interpretability constitutes a particular challenge in the context of industrial safety, where transparency of decision-making is of critical importance. Many advanced ML algorithms, such as deep neural networks, operate as so-called “black boxes,” whose decisions are not easily explainable to end users. In refinery systems, where operators and management must understand the rationale behind specific recommendations or warnings, a lack of interpretability can reduce trust in ML solutions and hinder their practical adoption. Consequently, increasing attention is being paid to the development of explainable models and techniques for interpreting machine learning results in safety applications.

In addition to technical and security challenges, the implementation of ML systems also faces organizational and regulatory barriers. The introduction of new technologies requires changes in organizational culture, additional

employee training, and adaptation of existing safety management procedures (Maseda et al., 2021). Resistance to change, lack of interdisciplinary knowledge, and limited resources can slow down or complicate the implementation process. Moreover, the refinery industry is subject to strict regulatory frameworks that require demonstrable reliability and compliance of safety systems with applicable standards. The integration of ML solutions must therefore be carefully aligned with regulatory requirements to ensure their acceptability and long-term sustainability.

Considering these challenges, it is evident that successful application of machine learning in refinery safety requires a holistic approach that integrates technical solutions, organizational change, and regulatory compliance. Understanding and addressing these limitations represents a crucial step toward effective and responsible digital transformation of industrial safety in refinery processes.

5. Discussion and Future Perspectives

The application of machine learning in refinery safety enables a transition from a reactive to a proactive approach to risk management, representing one of the most significant transformations in modern industrial systems. Traditional safety systems primarily responded after incidents had occurred, whereas the integration of ML algorithms enables the prediction of potential problems and their prevention before they escalate into critical events. Such a proactive approach contributes to reducing the number of incidents, minimizing economic losses, and protecting human lives and the environment, thereby significantly improving the overall safety of refinery processes.

One of the key aspects concerns the role of machine learning as an integral component of the Safety Management System (SMS) in refineries (Bramantyo et al., 2022). By integrating ML models into SMS, it becomes possible to continuously assess risks, monitor the performance of protective systems, and support real-time decision-making by operators and management. Predictive models enable dynamic adjustment of safety procedures and optimization of preventive maintenance, which further contributes to reducing the risk of incidents. In this way, ML technologies are not treated merely as an add-on to existing systems, but as an element in the evolution of modern industrial safety paradigms. Looking to the future, there is significant potential for the development and integration of additional digital technologies into refinery systems. One promising perspective is the application of digital twins – virtual replicas of physical facilities and processes that enable scenario simulation, risk analysis, and testing of safety procedures in controlled digital environments (Al-Jlibawi et al., 2020). Digital twins, combined with ML algorithms, allow for the identification of potential system weaknesses and the evaluation of the effectiveness of preventive measures before their implementation in actual facilities. Furthermore, the synergy of AI and IoT technologies offers opportunities for continuous process monitoring and adaptive responses to changes in system operation. The use of distributed sensors, edge computing, and predictive analytics algorithms enables continuous learning and real-time adaptation of ML models, significantly enhancing the efficiency and flexibility of safety operations. These technologies also support the integration of diverse data sources, including physical, process, and organizational parameters, thereby creating a holistic view of a facility's safety status.

Nevertheless, future application of these technologies requires careful balancing between innovative capabilities and practical constraints. Issues such as model interpretability, standardization of procedures, data protection, and cybersecurity remain critical factors determining implementation success. Effective integration of ML and digital technologies requires a multidisciplinary approach, collaboration among engineers, IT specialists, and management, as well as continuous monitoring of regulatory and industry standards.

In conclusion, this discussion demonstrates that machine learning and digital innovations represent key drivers of the transformation of industrial safety in refinery systems. Proactive approaches, integration into SMS, and the prospects of digital twins and AI–IoT technologies open new horizons in incident prevention, operational optimization, and overall enhancement of safety system effectiveness. Given the rapid pace of technological development, it is expected that future generations of refinery facilities will rely on comprehensive digital models, with machine learning functioning as a central element of safety architecture.

6. Conclusion

The application of machine learning in refinery safety represents a significant step toward the digital transformation of high-risk industrial systems. The analysis presented in this work demonstrates that ML algorithms, through predictive analytics and anomaly detection, enable a transition from a reactive to a proactive risk management approach, thereby enhancing the protection of personnel, the integrity of facilities, and

environmental safety. Integration of ML models into existing risk management systems and the Safety Management System (SMS) contributes to the optimization of preventive procedures, reduction of incidents, and improvement of operational efficiency, while simultaneously supporting informed real-time decision-making.

Beyond technical advantages, the study highlights implementation challenges, including data quality and availability, cybersecurity, model interpretability, and organizational and regulatory barriers. Addressing these challenges requires a multidisciplinary approach, continuous employee training, development of explainable models, and alignment with applicable industrial standards and regulations.

The discussion on future perspectives emphasizes the considerable potential of advanced digital technologies, such as digital twins, IoT integration, and edge computing, which, combined with ML, enable holistic process monitoring, scenario simulation, and adaptive management of safety risks. These technologies facilitate the creation of intelligent and resilient refinery systems, where ML becomes a central element of the safety architecture, and industrial safety transitions toward proactive and predictive management.

In conclusion, this work demonstrates that the application of machine learning in the refinery industry not only improves safety and operational reliability but also opens new avenues for digital innovation and modernization of safety paradigms, laying the foundation for sustainable, safe, and technologically advanced refinery systems of the future.

Literatura

1. Al-Jlibawi, A., Othman, M. L. B., Al-Huseiny, M. S., Aris, I. B., & Bahari, S. (2020, May). The efficiency of soft sensors modelling in advanced control systems in oil refinery through the application of hybrid intelligent data mining techniques. In *Journal of Physics: Conference Series* (Vol. 1529, No. 5, p. 052049). IOP Publishing.
2. Arinze, C. A., Izionworu, V. O., Isong, D., Daudu, C. D., & Adefemi, A. (2024). Integrating artificial intelligence into engineering processes for improved efficiency and safety in oil and gas operations. *Open Access Research Journal of Engineering and Technology*, *6*(1), 39-51.
3. Babayigit, B., & Abubaker, M. (2023). Industrial internet of things: A review of improvements over traditional scada systems for industrial automation. *IEEE Systems Journal*, *18*(1), 120-133.
4. Bramantyo, H. A., Utomo, B. S., & Khusna, E. M. (2022). Data processing for iot in oil and gas refineries. *Journal of Telecommunication Network (Jurnal Jaringan Telekomunikasi)*, *12*(1), 48-54.
5. Cuentas, S., Peñabaena-Niebles, R., & Garcia, E. (2017). Support vector machine in statistical process monitoring: a methodological and analytical review. *The International Journal of Advanced Manufacturing Technology*, *91*(1), 485-500.
6. Delshah, M., Rahimpour, H. R., & Rahimpour, M. R. (2023). Disaster cases in gas industry. In *Crises in Oil, Gas and Petrochemical Industries* (pp. 349-362). Elsevier.
7. Enemosah, A., & Ifeanyi, O. G. (2024). SCADA in the era of IoT: automation, cloud-driven security, and machine learning applications. *International Journal of Science and Research Archive*, *13*(01), 3417-3435.
8. Erinjogunola, F. L., Nwulu, E. O., Dosumu, O. O., Adio, S. A., Ajirotutu, R. O., & Idowu, A. T. (2020). Predictive safety analytics in oil and gas: leveraging AI and machine learning for risk mitigation in refining and petrochemical operations. *International Journal of Scientific and Research Publications*, *10*(6), 254-265.
9. Kim, O. D., Rocha, M., & Maia, P. (2018). A review of dynamic modeling approaches and their application in computational strain optimization for metabolic engineering. *Frontiers in microbiology*, *9*, 1690.
10. Korytko, T., & Piletska, S. (2022). Model of the adaptive management system of an industrial enterprise in the conditions of Industry 4.0.
11. Klein, J. A., & Vaughen, B. K. (2017). Process Safety: Key Concepts and Practical Approaches. CRC Press.
12. Maseda, F. J., López, I., Martija, I., Alkorta, P., Garrido, A. J., & Garrido, I. (2021). Sensors data analysis in supervisory control and data acquisition (SCADA) systems to foresee failures with an undetermined origin. *Sensors*, *21*(8), 2762.
13. Olaizola, I. G., Quartulli, M., Unzueta, E., Goicolea, J. I., & Flórez, J. (2022). Refinery 4.0, a review of the main challenges of the Industry 4.0 paradigm in oil & gas downstream. *Sensors*, *22*(23), 9164.

14. Omoarebun, P. O. (2023). Intelligent monitoring system for petroleum refinery processing operations and optimisation using artificial intelligence (Doctoral dissertation, University of Portsmouth).
15. Omol, E., Mburu, L., & Onyango, D. (2024). Anomaly detection in IoT sensor data using machine learning techniques for predictive maintenance in smart grids. *International Journal of Science, Technology & Management*, *5*(1), 201-210.
16. Pani, A. K., & Soofastaei, A. (2025). Designing Intelligence: Harnessing Soft Sensors and Advanced Analytics in Petroleum Refining for Industry 4.0. In *Advanced Analytics for Industry 4.0* (pp. 83-116). CRC Press.
17. Sarker, I. H. (2023). Machine learning for intelligent data analysis and automation in cybersecurity: current and future prospects. *Annals of Data Science*, *10*(6), 1473-1498.
18. Sun, Y., Mallick, T., Balaprakash, P., & Macfarlane, J. (2022). A data-centric weak supervised learning for highway traffic incident detection. *Accident Analysis & Prevention*, *176*, 106779.
19. Zhen, X., Ning, Y., Du, W., Huang, Y., & Vinnem, J. E. (2023). An interpretable and augmented machine-learning approach for causation analysis of major accident risk indicators in the offshore petroleum industry. *Process Safety and Environmental Protection*, *173*, 922-933.
20. Zheng, S., & Zhao, J. (2020). A new unsupervised data mining method based on the stacked autoencoder for chemical process fault diagnosis. *Computers & Chemical Engineering*, *135*, 106755.

Original Scientific Paper/Original naučni rad
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601077L

UDC/UDK: 005.92:343.452
004.65:004.8

Employee-Driven Leakage of Technical Documentation into General-Purpose LLMs: An Integrative Review

Luka Latinović[1], Oleg Zhukovskiy[2], Olga Mašić[3], Dejan Živković[4]

[1]Belgrade School of Engineering Management, Beopolis University, Serbia, luka.latinovic@fim.rs
[2]MART-INFO LLC (ООО «МАРТ-ИНФО»), Moscow, Russian Federation
[3] Belgrade School of Engineering Management, Beopolis University, Serbia, olga.masic@fim.rs
[4] [1]Belgrade School of Engineering Management, Beopolis University, Serbia, dejan.zivkovic@fim.rs

Abstract: General-purpose large language models are increasingly used by employees to interpret standards, troubleshoot systems, and draft or refine engineering artefacts. This routine assistance creates bidirectional flows: proprietary documentation is occasionally externalised as prompts, uploads, screenshots, or connector-mediated retrieval, while model outputs are pasted back into internal tickets, runbooks, and repositories. This integrative review synthesises heterogeneous evidence (peer-reviewed research, provider and regulator materials, and structured incident reporting) to map employee-driven leakage mechanisms along the documentation lifecycle and to derive a governance approach that is auditable under policy drift and multi-vendor toolchains. We identify a recurrent set of boundary-crossing transition points such as copy–paste, upload/OCR, connector invocation, and paste-back, where risk concentrates and where observability is often weakest. Across these pathways, four cross-cutting risk dimensions recur: confidentiality and competitive exposure, compliance and cross-border transfer, model-side effects (including extraction, spillover, and contamination risks), and incentive-driven governance gaps that sustain shadow workflows. Building on the mechanism map, we propose a proportionate "minimal guardrail stack" and an organisational evaluation framework combining qualitative risk scoring, rule-based escalation, and simple, trackable metrics (e.g., consolidation onto sanctioned channels, blocking effectiveness, inspection false positives, policy-drift lag, and time to a compliant alternative). The paper does not assert prevalence. Instead, it aims to make assumptions explicit and support cautious, workflow-compatible adoption decisions.

Keywords: data exposure, data exfiltration, training capture, shadow IT, AI governance, model-mediated knowledge transfer, egress.

Neovlašćeno otkrivanje tehničke dokumentacije od strane zaposlenih velikim jezičkim modelima opšte namene: integrativni pregled

Sažetak: Opštenamenski veliki jezički modeli sve se češće koriste među zaposlenima za tumačenje standarda, otklanjanje problema u sistemima i izradu ili doradu inženjerskih artefakata. Ova rutinska pomoć stvara dvosmerne tokove: vlasnička dokumentacija se povremeno iznosi izvan organizacije kroz upite (promptove), otpremanje fajlova, snimke ekrana ili preuzimanje posredstvom konektora, dok se izlazi modela zatim kopiraju nazad u interne tikete, operativna uputstva i repozitorijume. Ovaj integrativni pregled sintetizuje heterogene dokaze (recenzirana istraživanja, materijale pružalaca usluga i regulatora, kao i strukturisane izveštaje o incidentima) kako bi mapirao mehanizme curenja koje pokreću zaposleni duž životnog ciklusa dokumentacije i izveo pristup upravljanju koji je proverljiv u uslovima promena politika i višedobavljačkih lanaca alata. Identifikujemo ponavljajuće prelazne tačke na kojima se prelaze granice poverenja kao što su kopiraj–nalepi, otpremanje/OCR, pozivanje konektora i vraćanje sadržaja kopiranjem, na kojima se rizik koncentriše i gde je mogućnost uočavanja često najslabija. Kroz ove putanje ponavljaju se četiri poprečne dimenzije rizika: poverljivost i konkurentska izloženost, usklađenost i prekogranični prenos, efekti na strani modela (uključujući rizike ekstrakcije, „prelivanja" i kontaminacije) i upravljački jazovi vođeni podsticajima koji održavaju „shadow" tokove rada. Na osnovu mape mehanizama predlažemo proporcionalan „minimalni paket zaštitnih ograda" i organizacioni okvir za evaluaciju koji kombinuje kvalitativno ocenjivanje rizika, eskalaciju zasnovanu na

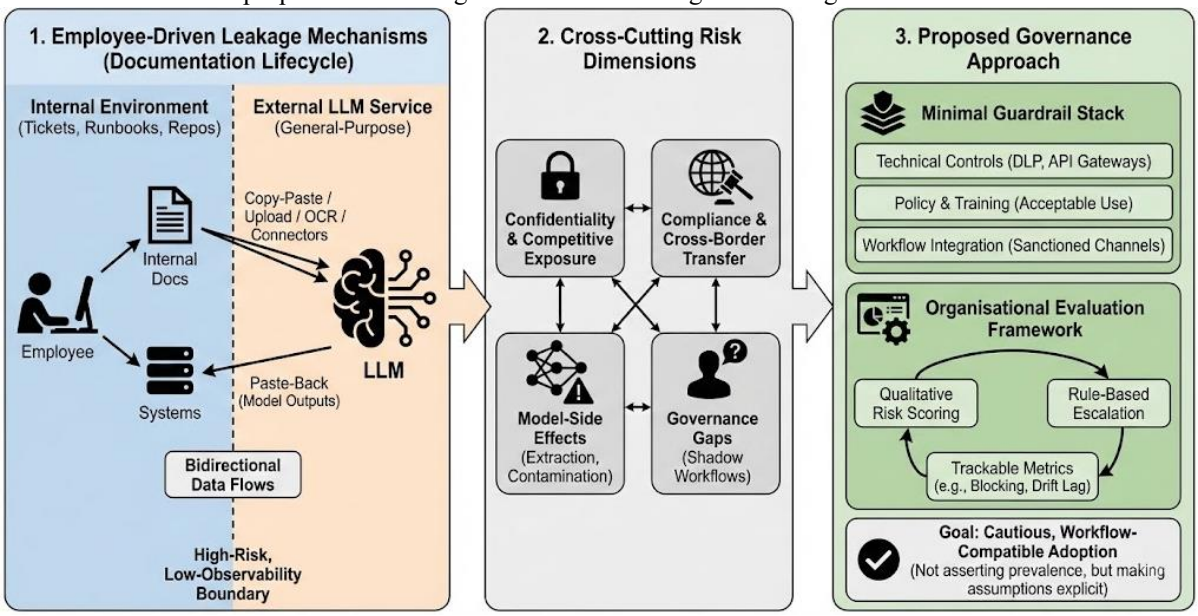
pravilima i jednostavne, pratljive metrike (npr. konsolidaciju na odobrene kanale, efikasnost blokiranja, lažno pozitivne nalaze inspekcije, kašnjenje u usklađivanju sa promenama politika i vreme do usaglašene alternative). Rad ne tvrdi učestalost pojave. Umesto toga, cilj je da pretpostavke učini eksplicitnim i podrži oprezne odluke o usvajanju koje su kompatibilne sa stvarnim tokovima rada.

Ključne reči: izloženost podataka, eksfiltracija podataka, obuhvat za trening, neautorizovani IT, upravljanje VI, modelom posredovan prenos znanja, izlazni tok (egres).

1. Introduction

General-purpose large language models (LLMs) appear to have moved from occasional summarisation and translation into routine problem solving within engineering and documentation work (Saka et al., 2024; J. Yang et al., 2024). Practitioners consult them to interpret standards, draft procedures, refactor build and computer-aided design (CAD)-adjacent scripts, and diagnose integration issues; their suggestions are then edited and incorporated into runbooks, standard operating procedures (SOPs), and specifications (Sajjadi Mohammadabadi et al., 2025). In parallel, employees may paste or upload excerpts of proprietary material to obtain targeted assistance (Malki et al., 2025a). This bidirectional exchange—documentation flowing outward to the model and model outputs flowing back into official artefacts—creates a plausible channel for exposure of technical knowledge to external systems and, conversely, for model-mediated knowledge to diffuse across organisations (Figure **Error! eference source not found.**).

Figure 1: Systematic mapping of employee-driven documentation leakage into general-purpose LLMs, illustrating bidirectional data flows across critical transition points (copy-paste, uploads, connectors) and the proposed "minimal guardrail stack" for organisational governance.



The mechanisms are principally socio-technical. Outbound disclosure can occur through direct prompts, file uploads, and screenshots subjected to optical character recognition (OCR), as well as through intermediaries such as browser extensions, plugins, and third-party chat “wrappers” that relay content to additional processors (Acar et al., 2020). These actions cross trust boundaries at the moments of copy–paste, upload, connector invocation, and paste-back. Provider data-use practices add further uncertainty: operational logging and safety/abuse telemetry may persist even where customer inputs are excluded from parameter updates (Clusmann et al., 2025; Ferrag et al., 2025). For clarity, this review uses the term training capture as operational shorthand for inputs later influencing model or feature improvement under some products or configurations; where inputs are excluded from parameter updates, operational retention (logs, safety traces, analytics) may still be material. Because model parameters are shaped by prior corpora and interactions, responses cannot be cleanly attributed to a single source, so cross-organisational influence is plausible even when any single exchange appears benign.

The stakes are non-trivial. Technical documentation—requirements and design specifications, CAD exports and bill-of-materials (BOM) notes, SOPs and runbooks, test and maintenance logs, configuration and deployment guides—often resides in version control systems (VCS) (Zolkifli et al., 2018), wikis (Standing & Kiniti, 2011), and product/application lifecycle management platforms (PLM/ALM) (Ebert, 2013). These artefacts encode trade secrets, tacit know-how, and occasionally personal data. Leakage may weaken claims to secrecy, complicate intellectual-property positions, or trigger data-transfer and confidentiality obligations (Prinz, 2025). Organisational incentives also matter: where productivity is rewarded and sanctioned alternatives are absent, employees may normalise unsanctioned LLM use, reducing observability at precisely the points where governance is needed (Williams et al., 2025).

2. Methodology

This study adopts an integrative review design, a methodology specifically chosen for its capacity to combine diverse data sources—including peer-reviewed literature, technical white papers, and incident reports—to generate new conceptual frameworks. Unlike systematic reviews that prioritise prevalence estimates, an integrative design focuses on synthesising these heterogeneous streams to conceptualise the leakage mechanisms and derive the governance model proposed in Section 6. Rather than aiming for the exhaustive quantification typical of systematic reviews, this integrative review prioritises conceptual synthesis and mechanism mapping, tracing how specific work practices and infrastructural choices produce employee-driven leakage pathways, while making limitations and assumptions explicit (Torraco, 2005; Whittemore & Knafl, 2005). Inclusion required relevance to employee LLM use, provider handling of user inputs, documentation-relevant risks, or mitigations/governance. Sources were coded by interaction surface (chat/API/plugin), data path (prompt/upload/relay), stated handling (retention/training/telemetry), and controls. Conflicting claims were retained as bounded scenarios. Research questions address pathways, risks, proportionate controls, and organisational evaluation.

Research questions.

- RQ1: Through which employee-driven pathways do proprietary documentation artefacts plausibly reach general-purpose LLM services?
- RQ2: What risk types arise, including confidentiality/IP loss, compliance and cross-border transfer, and model-side mechanisms (spillover, extraction, contamination)?
- RQ3: Which socio-technical controls appear proportionate, contractable, and auditable under realistic organisational incentives?
- RQ4: How should organisations evaluate providers and architectures using a qualitative scoring scheme, rule-based escalation to a minimal guardrail stack, and simple metrics?
- RQ1–RQ3 correspond to the review and synthesis components; RQ4 motivates the derivation of the operational risk-assessment framework.

2.1. Source identification and inclusion

Given the focus on mechanisms rather than prevalence estimates, the corpus intentionally spans:

- Peer-reviewed literature on LLM use in work settings, data leakage, cloud governance, and security/privacy of extensions, APIs, and cloud email/document platforms;
- Provider documentation (e.g. data-use policies, privacy/security white papers, product architecture descriptions) where these affect handling of user inputs and technical artefacts;
- Regulatory, standards, and guidance documents relevant to data protection, cross-border transfers, and safety/criticality of technical information;
- Structured incident databases and well-documented case reports involving generative-AI or adjacent cloud tooling where leakage of work artefacts is plausibly involved.

Inclusion required at least one of the following: (i) empirical or documented examples of employee LLM use in professional contexts; (ii) explicit discussion of how providers handle user inputs (retention, training use, telemetry); (iii) documentation-relevant risks (e.g. code, specifications, logs, design documents); or (iv) concrete mitigations, controls, or governance models that bear on such risks. Opinion pieces and purely speculative commentary without traceable mechanisms were excluded.

Searches were conducted iteratively in major scholarly databases and preprint servers, complemented by backward and forward citation chasing and targeted searches of provider and regulator sites. Given the rapid evolution of tooling and policies, inclusion emphasised recency and relevance to current LLM deployments, with older work drawn in selectively where needed to anchor long-standing mechanisms (e.g. DLP limitations, browser/endpoint artefacts).

2.2. Coding and mechanism-oriented synthesis

Eligible sources were coded using a structured template oriented around interaction surface, data path, handling, and controls:

- Interaction surface: chat interface, API, plug-in/extension, assistant, email/doc integration, CI/CD pipeline, link-sharing, endpoint.
- Data path: prompt, upload, relay, logging/telemetry, sync, or secondary redistribution.
- Stated handling: retention periods, training use, internal sharing, telemetry practices, and any constraints or guarantees.
- Controls: technical safeguards (e.g. DLP, gateways, key management), contractual controls, and organisational practices.

Sources were also annotated for evidence type (empirical study, documented incident, design/architecture description, policy text) and for their primary locus of risk (confidentiality/IP, regulatory/compliance, safety-criticality, or model-side effects). Conflicting claims, e.g., on whether a given provider uses customer inputs for training under specific plans, were retained as bounded scenarios rather than harmonised away: where policies, implementations, or interpretations diverged, this divergence was treated as an empirical feature of the ecosystem.

The synthesis proceeded in three steps. First, coded material was used to trace concrete leakage pathways from employee actions to LLM ecosystems, grouping similar mechanisms into the typology presented in Table 1. Second, across pathways, the review identified systematic misalignments between impact, likelihood, and observability, and recurrent patterns of control failure (e.g. policy-only measures, over-reliance on DNS blocking, lack of OCR-aware inspection, permissive OAuth scopes). Third, these patterns were cross-checked against organisational incentive structures and deployment realities to assess which proposed controls appear feasible and proportionate.

2.3. Derivation of the risk-assessment framework

The qualitative risk-assessment framework (Section 6) is explicitly derived from this mechanism-oriented synthesis rather than introduced as an independent model. For each pathway in the typology, severity and likelihood were assessed on ordinal four-point scales, with an observability modifier capturing detection difficulty and a confidence tag capturing the quality and consistency of the underlying evidence. These judgements are not claimed as precise measurements; instead, they provide a transparent rubric for prioritising governance attention and for making underlying assumptions inspectable.

Building on the identified control patterns, controls were grouped into three deployment tiers: a universal baseline (“Tier 0”), a gate-and-redact layer (“Tier 1”) applied to higher-priority or low-visibility pathways, and an architectural reduction tier (“Tier 2”) reserved for the highest-consequence or regulated documentation classes. Rule-based escalation from the qualitative risk scores to these tiers was then specified in auditable form, reflecting RQ4’s emphasis on proportionate, contractable, and operationally realistic guardrails.

Finally, a small set of simple metrics was defined to allow organisations to track whether the adoption of these guardrails is consolidating LLM use onto governed channels, reducing reliance on unsanctioned tools, and keeping governance responsive to changes in tooling and provider policies. These metrics are intentionally comparative rather than absolute, consistent with the mechanism-oriented focus of this integrative review: they are designed to support internal learning and recalibration, not to claim definitive quantification of LLM risk.

3. Typology of leakage pathways

This section classifies recurrent employee-driven routes by which documentation crosses trust boundaries. Categories overlap and often chain (e.g., screenshot → OCR → plugin relay → paste-back). Likelihood and impact are qualitative, reflecting context (sector, document criticality, provider tier, endpoint hygiene). Crucially, the table pairs each vector with its characteristic control failures and maps them to feasible technical mitigations,

forming the basis for the governance stack in Section 6. Table 1 is intended as a gap-analysis instrument and a living register.

Table 1: Leakage pathways for employee-driven exposure of proprietary technical documentation to general-purpose LLM ecosystems, characterised by immediate mechanism ("vector"), observability, likelihood, impact, typical organisational control failures, and feasible mitigations.

Pathway	Vector (immediate mechanism)	Observability	Likelihood	Impact	Typical control failures	Feasible controls	Example references
Direct prompt copy-paste of proprietary text	Snippets from specs/SOPs/logs pasted into public chat for task help; outputs pasted back into drafts	Often only egress domain; prompt content invisible unless gateway inspects	**High**	Ranges from minor context loss to disclosure of trade-secret parameters	No point-of-use labels; no JIT warnings; policy-only prohibitions	Prompt gateway with content inspection; client-side redaction/templates; sanctioned enterprise chat with retention-off	Incident 768: ChatGPT Implicated in Samsung Data Leak of Source Code and Meeting Notes (*Incident 768*, 2023)
File uploads (docs/archives) and screenshots→ OCR	Whole files or images dragged into chat; OCR extracts text, bypassing text-only DLP	MIME size/type visible; OCR text unseen unless OCR-aware DLP	**Med–High**	Higher than copy-paste (bulk movement)	DLP not OCR-aware; permissive MIME allow-lists; metadata left intact	OCR-aware inspection; strip metadata; sandbox viewers; restrict bulk uploads to sanctioned endpoints	Stealthy Information Leakage from Android Smartphone through Screenshot and OCR (Y. Kim et al., 2015)
Plugins / extensions / third-party "wrappers"	Tools relay prompts/files to additional vendors/ services	Tool call chains opaque; retention varies by tool	**Medium (grows with adoption)**	Amplified by multi-vendor propagation	Unvetted enablement; broad OAuth scopes; no kill-switch	Curated catalogues; per-tool scoping; disable unsanctioned wrappers; provenance logs	(*Incident 1186*, 2025; Starov & Nikiforakis, 2017)
API relays & server-side logging	Internal scripts/apps forward docs to external APIs; gateways/CI proxies log payloads	Network-layer visibility high; verbose logs become sensitive stores	**Medium**	Durable log exposure; developer environment spillage	Keys in code; excessive request/response logging; debug proxies	Secrets managers; minimum-necessary logging with redaction; scoped service accounts	Detecting Misuse of Security APIs (Mousavi et al., 2025)
Email/Doc assistants auto-syncing to provider clouds	Mail/drive assistants ingest repositories for "smart" features	OAuth scopes visible; actual sync/retention opaque tenant-wide	**Medium**	Broad ingestion (mailboxes/ drives)	"Accept all" scopes; weak off-boarding of assistants; unclear residency	Least-privilege OAuth; opt-out by default; residency/retention terms in DPAs	Security challenges for cloud-based email infrastructure (Bhardwaj & Goundar, 2017)
Shared links / chat transcripts / permissive defaults	"Anyone with the link" sharing; exported chat threads; external re-sharing	Some audit trails; hard to track onward sharing	**High in collaborative organisations**	Wide unintended audience; persistent re-disclosure	Public-link defaults; long-lived tokens; no expiry	Domain-restricted links; expiries; secret scanning on shared artefacts	Link-based sharing, "anyone with the link", large unintended audience (Wan et al., 2024)
Endpoint artefacts (clipboard / caches / sync)	Clipboard history, local/browser caches, synced profiles retain prompts/ responses	Low at time of event; surfaces later in support/legal	**Medium**	Secondary disclosure; lateral movement via synced data	Unmanaged devices; unrestricted clipboard/browser sync	Managed endpoints/VDI; disable/scope clipboard & sync; cache hygiene	Clipboard Data Attacks and Detection via Remote Desktop Protocol (Mohamed et al., 2023)

Taken together, the pathways in Table 1 demonstrate that "employee-driven leakage" is not a single action but a heterogeneous ecology of work practices spanning prompts, file uploads, plug-ins, API relays, sharing links, and residual endpoint artefacts. The conventional risk narrative—an engineer copy-pasting proprietary text into a

public chatbot—represents only the most visible portion of the surface. Once screenshots, bulk document uploads, wrapper-mediated calls and auto-ingesting assistants are considered, a substantial fraction of exposure moves into channels where payloads are either invisible to existing controls (e.g. text-only DLP confronted with OCR) or fragmented across multiple vendor systems and log stores. Blocking a well-known LLM domain therefore addresses the least subtle vector while leaving higher-volume leakage paths unaffected.

A clear structural pattern is the systematic misalignment between observability and impact. Direct prompt pasting is high-frequency and high-impact but still offers a single egress point that can be monitored or rate-limited (Ray, 2023; Williams et al., 2025). By contrast, plug-ins, multi-vendor wrappers, and internal API relays diffuse responsibility and auditability, converting transient request payloads into durable, sensitive logs (Rathod et al., 2025). Email and document assistants expand the threat from point leakage events to corpus-wide ingestion, ingesting entire drives or mailboxes under opaque retention (Baek et al., 2025; R. Yang et al., 2025). Shared links and permissive transcript exports quietly widen the audience for technical documentation long after any deliberate interaction with an LLM has ended (Alzamil et al., 2025). Endpoint artefacts such as clipboard history, browser caches and synchronised profiles extend the temporal window of exposure, surfacing later in support, legal hold, or on compromised devices (Chivers & Hargreaves, 2011; Hur et al., 2023; Mendoza et al., 2015; Oh et al., 2011; Okolica & Peterson, 2011).

Recurrent control failures reinforce these pathways. Organisations continue to rely on static prohibitions and awareness messaging, while genuine friction is absent at the point where a user uploads, pastes, or shares proprietary material (Taeihagh, 2025). Technical safeguards remain concentrated around narrow choke points (e.g. DNS blocks) rather than at the high-leverage layers highlighted in the table: OCR-aware inspection, scoped OAuth permissions, curated plug-in catalogues, controlled secret handling, and link-expiry enforcement (Bhushan, 2025; Challappa et al., 2025; K. Chen et al., 2025). Most of these mitigations already exist in enterprise DLP suites, identity platforms, and endpoint management systems, but are rarely configured with LLM-mediated documentation flows in mind.

In sum, Table 1 operationalises what this integrative review refers to as employee-driven leakage of technical documentation into general-purpose LLMs. Each pathway is initiated or sustained by routine employee actions rather than adversarial compromise, and the information at risk consists primarily of specifications, SOPs, design notes, logs, and other engineering artefacts that constitute organisational intellectual property. Crucially, these artefacts are being routed—directly via prompts and uploads, or indirectly via plug-ins, API relays, and assistant ingestion—into general-purpose LLM ecosystems whose retention, secondary use, and vendor chains are only partially observable. The remainder of this integrative review examines how these leakage patterns arise from misaligned incentives, workflow convenience, permissive defaults, and governance assumptions that conceptualise “LLM risk” too narrowly.

4. System model and cross-cutting risks

The leakage pathways identified in Section 3 can be interpreted within a compact system model that captures (i) how technical documentation moves through its lifecycle, (ii) where trust boundaries are crossed, and (iii) which risk dimensions consequently arise. This model provides the structural basis for the evaluation framework in Section 6.

4.1. Documentation lifecycle, information states, and trust boundaries

For governance purposes, technical documentation can be treated as moving through five functional phases: (1) creation and revision, (2) storage and indexing, (3) access and transformation, (4) integration into official artefacts, and (5) release or archival (Hullavarad et al., 2015; Mokhtar & Yusof, 2015; Salminen et al., 2014; Sovrano et al., 2025). What changes with general-purpose LLM adoption is not the existence of these phases but the speed and opacity with which documentation can traverse them, often without the metadata, access controls, and audit traces that traditionally signal a boundary crossing (Karras et al., 2025; Taeihagh, 2025). In LLM-integrated applications, “documentation” routinely becomes both data and instructional substrate; prompt injection work has shown that untrusted content can be interpreted as control text, collapsing a boundary that conventional document workflows assume is stable (Greshake et al., 2023; Liu et al., 2024).

Across these phases, four transition points repeatedly cross trust boundaries in real organisations: (i) copy–paste of excerpts into chat interfaces, (ii) file or screenshot uploads (including image-to-text extraction), (iii) connector or plug-in invocation that grants broad repository access, and (iv) paste-back of model outputs into internal

repositories, tickets, or version control. The first three transitions move artefacts outward (from enterprise governance into external service stacks or multi-vendor toolchains); the fourth moves artefacts inward again, but now as LLM-mediated derivatives whose provenance and transformation chain are difficult to evidence (Taeihagh, 2025; B. Yang et al., 2025). Tool-using "agent" architectures further widen the boundary surface: distinct stages (system prompt, user prompt handling, memory retrieval, and tool usage) create multiple injection and poisoning opportunities, including scenarios where a poisoned memory store or tool response steers downstream actions (H. Zhang et al., 2025).

To make these transitions tractable, artefacts can be represented in four information states: raw proprietary (full internal documents), sensitive-derived (excerpts, logs, contextual summaries that still reveal operational or design intent), redacted (structure retained, sensitive values removed), and public (already disclosed) (Feretzakis, Papaspyridis, et al., 2024; Feretzakis, Verykios, et al., 2024). The central governance problem is that most real work pressure pushes users to operate in the raw-proprietary and sensitive-derived states, exactly where LLM-mediated transformations are least observable and most consequential (Pahune et al., 2025; Waters-Lynch et al., 2025). Moreover, trust boundaries are not purely "inside vs. outside": LLM-integrated applications can bridge to internal systems (e.g., databases) in ways that reintroduce classic injection risks through natural-language interfaces (Pedro et al., 2025). Model-side leakage, extraction, and tool-mediated spillover risks are defined and evidenced in Section5.3; here we treat them only as a downstream risk class that becomes relevant once artefacts enter LLM interactions.

4.2. Provider data handling and residual retention

Once documentation crosses an enterprise boundary, it is processed within a multi-layer service stack that typically includes (i) an inference layer that receives prompts and uploaded artefacts, (ii) safety/abuse-monitoring and operational logging pipelines, and (iii) optional analytics or product-improvement pathways whose activation depends on plan, settings, and contract terms (Pahune et al., 2025; Shvetsova et al., 2025). In practice, "no-training" or "do not use for model improvement" controls should be interpreted narrowly: they may constrain whether inputs/outputs are used to update model parameters, but they do not necessarily eliminate short-term retention for abuse monitoring, incident response, debugging, or compliance workflows (Malki et al., 2025b; A. Zhang, 2025). For example, OpenAI's API documentation notes that abuse-monitoring logs may include prompts and responses and are retained for up to 30 days by default, with stricter options (e.g., zero data retention) available only under additional requirements and prior approval (*Data Controls in the OpenAI Platform*, n.d.). Similarly, enterprise-oriented offerings may provide administrator-configurable retention and deletion behaviour, but retention can still be affected by legal obligations and operational constraints (*Enterprise Privacy at OpenAI*, n.d.). In parallel, enterprise copilots can introduce their own retention and compliance storage paths (e.g., mailbox-based retention and eDiscovery handling), which means that "what users see" in a chat UI is not a reliable indicator of what is retained for governance purposes (Ishrak Alim, 2025; Sai et al., 2024). Finally, when LLM use is mediated by plug-ins, connectors, or third-party assistants, the chain of processing can expand to additional controllers/processors and subprocessors, increasing the difficulty of end-to-end accountability and DPIA-style mapping of data flows (Das et al., 2025). These handling uncertainties do not imply adversarial intent; rather, they constitute a structural opacity that, combined with the leakage pathways under the Section 3, creates non-trivial residual exposure risk that must be managed explicitly (through minimisation, segmentation, retention controls, and auditable governance).

4.3. Cross-cutting risk dimensions

The boundary crossings in Section4.1–4.2 and the pathways in Section3 generate four recurring risk dimensions: (i) confidentiality/competitive exposure, (ii) compliance and cross-border transfer, (iii) model-side effects (extraction, spillover, contamination), and (iv) incentives and governance gaps. These dimensions are analytically distinct but operationally coupled in multi-component LLM service stacks (UI, middleware, connectors, tool calls, and logging). Section5 defines each dimension and specifies the evaluation questions used in Section6.

5. Risk dimensions

This section defines the four risk dimensions precisely enough to be operationalised in Section6 for evaluating real documentation workflows. Each dimension is specified in terms of mechanisms, observable indicators, and governance-relevant control points within LLM-mediated service stacks.

5.1. Confidentiality and competitive exposure

Confidentiality risk in documentation-heavy work is rarely limited to "obvious secrets." LLM-assisted handling can create a cumulative "mosaic risk," in which seemingly low-sensitivity fragments—parameter defaults, troubleshooting sequences, tolerance bands, supplier identifiers, or sequencing logic—become sensitive when aggregated across repeated interactions, sessions, recipients, and time (Agarwal et al., 2024; Staab et al., 2024; Wang et al., 2025). The governance-relevant point is not deterministic trade-secret disclosure from any single snippet, but the progressive erosion of compartmentalisation and evidentiary control: repeated transfers can reconstruct design intent and tacit know-how while making it harder to demonstrate that reasonable secrecy measures were maintained (Nealey et al., 2015; Ozcan et al., 2025). In practical terms, weakly governed externalisation into third-party LLM workflows can undermine the defensibility of secrecy claims, especially where organisations cannot later evidence what was disclosed, under what contractual terms, and with what retention constraints (Aplin et al., 2023; Ozcan et al., 2025).

A second mechanism is reverse exposure via "paste-back." When employees integrate model outputs into internal artefacts (tickets, runbooks, manuals, deployment guides), organisations may import third-party or previously exposed proprietary phrasing, distinctive parameter values, or unattributed code fragments into governed repositories, complicating provenance, access control, and downstream sharing (Feretzakis et al., 2025; Perry et al., 2023). Even if such events are individually low probability, they can be hard to detect, trace, and quarantine at scale—particularly when productivity pressure normalises "copy–modify–ship" behaviours (Brynjolfsson et al., 2025). In multi-vendor plug-in ecosystems, the confidentiality boundary is further weakened by indirect prompt injection and tool calls: untrusted external content can steer what is retrieved or summarised and thereby increase the chance of disclosing internal context available via chat history, attachments, or connected tools (T. Chen et al., 2025; Zhan et al., 2024).

5.2. Compliance and cross-border transfer

Compliance risk in LLM-mediated documentation workflows is driven less by intent than by opacity and distributed processing. Technical documentation, operational logs, and incident records frequently embed personal data (e.g., names, emails, chat excerpts), access tokens in traces, device or account identifiers, and sometimes regulated or contractually restricted technical information. Once such materials cross an enterprise boundary into an LLM service stack, organisations can struggle to demonstrate controller–processor discipline because responsibilities distribute across controllers, processors, and sub-processors, while content may be replicated into telemetry, abuse monitoring, and debugging workflows that are only partially observable to the organisation (Feretzakis et al., 2025; Kramcsák, 2023). GDPR-oriented analyses argue that this stack complexity creates practical obstacles for purpose limitation, minimisation, retention limitation, and the execution—and evidencing—of access, rectification, erasure, and broader accountability obligations (Feretzakis et al., 2025; Kuru, 2024). Even where "no-training" is contractually specified, the operational compliance question remains whether the organisation can reliably map where the data travelled, which entities processed it, which jurisdictions stored it, and how deletion or DSAR-like obligations would be executed across all replicas (Feretzakis et al., 2025; Kramcsák, 2023).

Cross-border transfer is therefore best treated as a sub-problem of the same opacity: data residency and onward-transfer constraints may be difficult to evidence when content is routed through multiple subprocessors or persists in operational logs across regions, regardless of whether inputs are excluded from model-parameter updates. The governance gap is thus not only "what the model learns," but "where the content travels, where it rests, and how deletion claims can be substantiated." Legal scholarship further notes that the GDPR's consent model is often a poor fit for many AI contexts, reinforcing the need to minimise personal data in prompts and to treat multi-processor and cross-border flows as first-class risk factors rather than documentation footnotes (Kramcsák, 2023). For documentation-rich workflows, this pushes compliance away from abstract policy statements and toward concrete, auditable workflow design—classification rules, pre-submission redaction, approved routing, and verifiable retention controls (Feretzakis et al., 2025).

5.3. Model-side effects (spillover, extraction, contamination)

Model-side effects concern what can happen inside, or because of, the model and its surrounding agent/tool environment once proprietary or regulated artefacts enter LLM interactions. The strongest empirical basis is not general speculation about "models will leak everything," but demonstrated leakage mechanisms under realistic

threat models, including privacy leakage and extractable memorisation (Rathod et al., 2025). Empirical work shows that alignment does not eliminate extraction risk: under some conditions, adversaries can elicit memorised training samples from aligned, production-grade models, which makes a residual (non-zero) leakage probability a defensible governance assumption even when providers deploy safety layers (Nasr et al., 2025). Complementary evidence indicates that language models can leak personally identifiable information and that such leakage can be probed and measured; large-scale analyses characterise PII exposure modes, and tools such as ProPILE operationalise tests for whether PII is retrievable from LLM-based services (S. Kim et al., 2023; Lukas et al., 2023). Beyond memorisation, privacy can also be violated through inference: models may infer sensitive attributes from text at inference time, extending the risk surface from "did the model memorise it?" to "can the model deduce it?" (Lukas et al., 2023; Staab et al., 2024). The implication for documentation leakage is bounded but concrete: the relevant risk is not guaranteed disclosure, but an irreducible residual chance that portions of submitted content become recoverable or reconstructible in ways that are difficult to attest, reverse, or independently audit (Nasr et al., 2025). In the system model of Section4, these effects are downstream of outward boundary crossings and must therefore be assessed together with retention/logging and tool-access decisions.

A second mechanism is tool-mediated spillover in LLM-integrated applications and agents. Prompt-injection research shows that when LLMs are embedded in pipelines that ingest untrusted external content, attackers can manipulate model behaviour to exfiltrate data or trigger unintended actions by exploiting the blurred boundary between "instructions" and "data" (Greshake et al., 2023; Liu et al., 2024). In enterprise documentation settings, the practical concern is not deterministic cross-customer spillover, but limited attestability: organisations may be unable to prove that proprietary fragments did not influence future outputs, audits, or emergent tool-mediated behaviours—especially in systems vulnerable to indirect prompt injection and retrieval/tool contamination (Zhan et al., 2024). Agent-centric benchmarking further indicates that tool access and memory mechanisms expand the attack surface and require evaluation of attacks and defences at the agent layer, not only at the base model (H. Zhang et al., 2025). The governance consequence is structural: if organisations cannot bound which inputs are trusted, what tools can be invoked, and what logs persist, then "safe use" claims cannot be substantiated without explicit constraints, monitoring, and audit artefacts—criteria operationalised in Section6.

5.4. Incentives and governance gaps

The persistence of the high-impact leakage pathways in documentation-heavy work is best explained by an incentives-and-capability mismatch. Activities such as debugging, configuration, incident response, and standards interpretation carry high time pressure and cognitive load, while general-purpose LLM tools can deliver measurable productivity and speed-of-resolution gains in real workplace deployments—often disproportionately benefiting less experienced staff (Brynjolfsson et al., 2025). This creates strong local incentives to externalise real work artefacts into LLM tools even when formal governance is incomplete. Organisational scholarship on covert or "shadow" generative-AI use similarly finds that, where sanctioned alternatives are slow, ambiguous, absent, or cumbersome, employees predictably route around policy using unapproved tools, personal accounts, or plug-ins because the perceived short-term benefit dominates abstract policy risk (Waters-Lynch et al., 2025). In that setting, policy-only prohibitions are not credible: they do not remove demand for rapid interpretation of error traces, runbooks, and documentation; they mainly displace the behaviour into less observable channels (Waters-Lynch et al., 2025).

The practical implication is that leakage risk is not only a technical control problem but a workflow-design problem. Organisations must either provide sanctioned, auditable tooling that matches the productivity utility of consumer-grade LLMs, or accept that unsanctioned pathways will remain active and will dominate precisely in high-pressure contexts (incidents, outages, escalations) where documentation is most sensitive (Brynjolfsson et al., 2025; Waters-Lynch et al., 2025). For this reason, the evaluation framework in Section6 should treat incentives, usability, and governance capacity as first-class risk controls, alongside technical restrictions and monitoring, rather than relying on awareness training as the primary mitigation.

6. Evaluation framework for organisations

This section operationalises RQ4 by translating the pathway typology and cross-cutting risk dimensions into a reproducible governance rubric—qualitative scoring, rule-based escalation to a minimal guardrail stack, and a small set of metrics that can be tracked over time. The purpose is not to "measure" leakage with spurious precision, but to make risk assumptions auditable under policy volatility, multi-vendor processing chains, and strong employee incentives to externalise cognitive work into general-purpose LLMs. The unit of analysis is the

transition point where documentation crosses a trust boundary (copy–paste, upload/OCR, connector invocation, and paste-back). Controls should therefore be evaluated primarily at these transitions, not only at coarse network egress. This emphasis is consistent with the literature showing that LLM-integrated workflows introduce additional attack surfaces—particularly indirect prompt injection, tool-chain manipulation, and memory poisoning—whose observability and responsibility attribution are structurally weak in multi-component stacks (Greshake et al., 2023; Liu et al., 2024; A. Zhang, 2025). Where organisations use retrieval-augmented generation (RAG) and tool-using agents, evaluation must treat external content, tool responses, and "memory" stores as adversarially influenceable inputs, not merely benign context (Greshake et al., 2023; H. Zhang et al., 2025).

6.1. Risk scoring

Risk is scored on severity and likelihood, adjusted by an observability modifier, and accompanied by a confidence tag. Severity captures the sensitivity and blast radius of the artefact class (including trade-secret core, safety-critical, or export-controlled categories); likelihood captures workflow friction and incentive alignment (low-friction, high-speed actions score highest); observability captures whether detection is prompt and reliable or plausibly absent (e.g., wrapper relays, local artefacts, opaque tool chains). The priority score is defined as:

$$P = (S \times L) + O, \text{ with } S \in \{1,2,3,4\}, L \in \{1,2,3,4\}, \text{ and } O \in \{-1,0,+1\}. \quad (1)$$

- **Severity** $S \in \{1,2,3,4\}$: sensitivity and potential blast radius (1: routine internal; 2: internal but non-critical; 3: commercially sensitive; 4: trade-secret cores, export-controlled, or safety-critical artefacts);
- **Likelihood** $L \in \{1,2,3,4\}$: pathway friction and incentives (1: architecturally unlikely; 2: possible with effort/exception; 3: common workflow, weak guardrails; 4: low-friction, fast local action);
- **Observability modifier** $O \in \{-1,0,+1\}$: subtract 1 if promptly visible (e.g., gateway inspection); add 1 if detection is improbable (e.g., wrapper relays, endpoint artefacts);
- **Confidence tag** $C \in \{\text{low, medium, high}\}$: evidence quality for the assigned S,L,O. When confidence is low, organisations should either (i) treat the score as provisional and prioritise evidence collection (logs, sampling, red-team tests), or (ii) apply the next-higher tier as a conservative default until confidence improves.

The central discipline is interpretive: P is a triage band, not a probability. For model-side concerns, organisations should explicitly distinguish (i) direct disclosure (employee sends proprietary content outward), (ii) indirect disclosure (content later appears through logs, tool chains, or sharing defaults), and (iii) model-mediated effects with bounded but non-zero plausibility (memorisation and extraction; privacy inference; prompt-driven exfiltration). The peer-reviewed literature does not justify deterministic leakage claims, but it does justify treating extraction and inference risks as "credible under some conditions," including for aligned or production systems and for PII-like sequences (Lukas et al., 2023; Staab et al., 2024; Nasr et al., 2025).

6.2. Minimal guardrail stack (rule-based escalation)

Tiering is a governance choice: it makes enforcement discussable and testable, and it reduces the common failure mode in which organisations maintain a prohibition while tolerating unmanaged shadow use because sanctioned workflows are slower or absent (Waters-Lynch et al., 2025). Tier 0 should be universal and low-friction: sanctioned endpoints and identity-bound access; least-privilege scopes for connectors; point-of-use classification and just-in-time prompts before paste/upload; and contractual commitments on retention, sub-processing, and training exclusion for customer inputs. If Tier 0 does not provide a usable compliant path for the dominant tasks (summarisation, troubleshooting, standards interpretation), it should be treated as ineffective by design and expected to increase unsanctioned tool use. Tier 1 adds inspection and minimisation where impact or invisibility is high: prompt/file gateways with OCR-aware inspection; metadata stripping; client-side redaction templates; curated plug-in catalogues with kill switches; and logging minimisation with targeted redaction. Tier 2 is reserved for the highest-consequence classes: architectural reduction via private or enterprise deployments with verifiable residency/retention, proxy-mediated key management, and periodic verification (including deletion tests and change-log review).

Escalation rules should remain simple and auditable rather than "optimised." A defensible baseline is: apply Tier 0 universally; add Tier 1 when (a) P is high, (b) severity and likelihood are both high, or (c) observability is poor; add Tier 2 whenever S = 4 (even if likelihood is moderate) or where regulated/export-controlled classes are plausibly in scope. In tool-using or RAG-enabled systems, Tier 1–2 decisions should weight prompt-injection

evidence more heavily because the literature demonstrates that "data-as-instructions" attacks can turn retrieved documents or tool outputs into control channels.

6.3. Escalation rules (auditable, not optimised):

- Apply Tier 0 universally.
- Add Tier 1 if $P \geq 9$ or and ($S \geq 3$ and $L \geq 3$) or O=+1.
- Add Tier 2 if $S = 4$ with $L \geq 2$, or whenever regulated/export-controlled data are in scope, irrespective of P.

These rules prioritise high-consequence, high-likelihood, low-visibility pathways and concentrate enforcement at transition points (before prompts/uploads and before paste-back).

6.4. Provider and architecture due diligence questions

Because provider practices and plugin ecosystems shift, due diligence should be stated as questions that can be contractually answered and periodically re-validated. Minimum questions include: (i) what inputs are retained, where, and for how long (prompts, uploads, safety traces, telemetry, tool outputs); (ii) what is excluded from model improvement and what is merely excluded from parameter updates; (iii) what sub-processors and tool vendors receive data under connectors; (iv) what audit evidence exists (tenant logs, deletion attestations, scoped keys, and control-plane enforcement); (v) what is the failure mode under prompt injection and tool compromise (output handling, tool permissions, memory poisoning) and (vi) what controls exist to detect and govern paste-back of model outputs into internal repositories (e.g., watermarking/provenance metadata, DLP on commits/tickets), and how are exceptions audited? The point is not to assume malice; it is to treat opacity itself as a risk amplifier, particularly for high-value documentation classes.

6.5. Metrics (definitions and formulas)

Metrics should be few, operationally collectible, and interpretable as trends rather than absolutes. Windows of observation (e.g., rolling 30 days) and inclusion criteria should be specified by the organisation. To make metrics comparable over time, the organisation should version the measurement specification (data sources, sampling, and what counts as an "interaction") and record any changes alongside the trend charts.

- Coverage of sanctioned egress (COV_{egress}):

$$Cov_{egress} = \frac{\text{request to allow-listed LLM endpoints}}{\text{all detected LLM requests}} \tag{2}$$

- Adoption ratio (AR) (sanctioned vs. unsanctioned tools):

$$AR = \frac{\text{session using sanctioned endpointss/tools}}{\text{session using sanctioned + unsanctioned tools}} \tag{3}$$

- Blocking effectiveness (BE) (policy-relevant events):

$$BE = \frac{\text{blocked outbound events}}{\text{blocked + allowed outbound events}} \tag{4}$$

- False-positive rate of inspection (FPR):

$$FPR = \frac{\text{non-sensitive items flagged}}{\text{all items flagged}} \tag{5}$$

- Drift lag (DL) (policy responsiveness to external change):

$$DL = t_{internalupdate} - t_{providerchange} \tag{6}$$

- Mean time to compliant alternative ($MTTC$):

$$MTTC = \frac{1}{N}\sum_{i=1}^{N}\left(t_{approved}^{(i)} - t_{request}^{(i)}\right) \quad (7)$$

Interpretation is comparative rather than absolute. Improvements in coverage (Cov_{egress}) and adoption ratio (*AR*) indicate consolidation onto governed channels; increases in blocking effectiveness (*BE*) coupled with stable or declining false-positive rates (*FPR*) suggest effective, non-disruptive inspection; reductions in drift lag (*DL*) and mean time to compliant alternative (*MTTC*) indicate responsiveness and usability of sanctioned paths. The leakage typology mapped in Table 1 should be maintained as a living register; when architecture, controls, or provider commitments change within a reporting window, update entries, note the change date, and recompute priority bands with the same rubric so that metric shifts can be interpreted against governance changes.

7. Discussion

This integrative review set out to map how routine employee interactions with general-purpose LLMs can create bidirectional flows between internal documentation systems and external model ecosystems, and to translate that mechanism map into proportionate, auditable governance. The synthesis supports four main interpretations that jointly answer the research questions.

Regarding RQ1, the review indicates that leakage risk concentrates not in exotic attacks but in routine boundary crossings embedded in everyday documentation work: copy–paste of snippets into chat interfaces, upload of files and screenshots (including OCR-mediated extraction), connector-mediated retrieval across repositories, and paste-back of model outputs into internal tickets, runbooks, and repositories (Table 1; Section 4). These transitions are high-frequency, low-friction, and often weakly observable, which makes them structurally more consequential than isolated "full-document exfiltration" narratives. The implication is that governance must target *transition points* as the unit of control rather than attempting to "secure documentation" in the abstract. This conclusion remains bounded by the review's design: the paper maps plausible and documented mechanisms, but it does not estimate prevalence by sector or vendor.

With respect to RQ2, across the mapped pathways, four risk dimensions recur and compound: (i) confidentiality and competitive exposure through cumulative disclosure of seemingly minor fragments, (ii) compliance and cross-border transfer challenges driven by distributed processing chains and evidencing requirements, (iii) model-side effects that justify treating leakage/inference as bounded-but-nonzero residual risk, and (iv) incentive-driven governance gaps that sustain shadow workflows when sanctioned alternatives are absent or slower (Section 5). The persistence of these dimensions follows from two structural features: compositional service stacks (interfaces, logging/telemetry, plug-ins, relays) and limited organisational observability of where artefacts travel and persist once they cross the boundary (Section 4). The implication is that the same control family can reduce one dimension while worsening another (e.g., logging for security can amplify retention exposure), so an explicit multi-dimensional framing is necessary for credible trade-off management. This synthesis is constrained by heterogeneity in legal regimes and service configurations; therefore, the risk dimensions should be treated as a portable taxonomy rather than a uniform compliance verdict.

Addressing RQ3, the evidence supports a mechanism-centred governance approach that prioritises auditable guardrails at the transition points rather than policy-only prohibitions or coarse network blocking (Sections 4–6). The proposed "minimal guardrail stack" is defensible because it maps directly onto the failure modes in Table 1: classification and just-in-time friction before paste/upload, OCR-aware inspection and redaction, least-privilege connector scoping with curated plug-in catalogues, and logging minimisation in internal relays, with escalation for high-sensitivity artefacts (Section 6). The implication is that organisations reduce risk most credibly by consolidating employees onto sanctioned channels and constraining connector/tool permissions, while maintaining usable workflows to prevent displacement into less observable shadow use. This remains conditional: control effectiveness depends on implementation quality (especially inspection accuracy and connector governance) and on whether sanctioned alternatives match operational tempo in high-pressure contexts.

In response to RQ4, the paper's contribution is an evaluation framework that converts the pathway map and risk taxonomy into auditable organisational decision-making: qualitative scoring (severity, likelihood, observability, confidence), rule-based escalation to tiered controls, and a small set of governance metrics that track coverage, friction, drift, and time-to-compliant-alternative rather than attempting to quantify leakage prevalence (Section 6). This responds directly to the evidentiary reality that incidents are under-reported and downstream attribution is weak; therefore, a tractable framework should focus on what organisations can measure and improve—sanctioned-channel consolidation, transition-point block rate, inspection false positives, drift lag, and mean time to a compliant workflow—while keeping uncertainty explicit (Section 6; Limitations). The implication is that

"risk reduction" is operationalised as reduced boundary crossings for high-sensitivity content and increased auditability of those that remain, not as a claim that leakage has been eliminated. The framework is bounded by log availability and classification fidelity; where these are absent, metrics must be treated as indicators or established through sampling audits.

Two broader implications follow for scholarship and practice. For research, the review highlights a measurement gap: incident reporting is sparse, attribution beyond initial disclosure is rarely demonstrable, and the most consequential channels (plug-in chains, OCR, endpoint residue, internal relay logging) are under-studied relative to prompt text alone. This suggests value in organisational field studies of governed vs. unguided adoption, and in technical evaluations of inspection/redaction effectiveness for multimodal inputs and tool-using agents (Greshake et al., 2023; Liu et al., 2024; Zhang et al., 2025). For practice, the main warning is that "no-training" commitments, while relevant, do not resolve governance: the risk surface is dominated by boundary crossings, distributed processing, and weak observability, so controls must be designed for traceability and minimisation across the full workflow chain rather than anchored in a single provider claim. The paper's contribution is therefore best understood as a compact, mechanism-centred roadmap for cautious adoption: it does not assert prevalence or inevitability of leakage, but it makes the pathways, risk dimensions, and evaluation assumptions explicit enough to support auditable organisational decisions.

8. Limitations of the study

This study adopts an integrative review with a mechanism-oriented synthesis rather than a systematic review or meta-analysis. That choice is methodologically defensible for an emergent and policy-volatile domain, but it imposes clear constraints on inference. First, coverage is not exhaustive and selection is purposive: the corpus intentionally spans heterogeneous evidence types (peer-reviewed studies, preprints, provider documentation, regulatory guidance, and structured incident reporting), prioritising mechanism mapping over prevalence estimation. As a result, the paper should be read as a *structured plausibility argument* about how leakage can occur, not as an epidemiology of how often it occurs in any sector or region.

Second, time variance is a fundamental limitation. Provider terms, retention defaults, product architectures, and connector ecosystems change rapidly, and the same "LLM service" label can hide materially different handling depending on plan, configuration, geography, and integration choices. Accordingly, statements about provider handling are best interpreted as time-stamped observations, not stable properties of the ecosystem. This limits the durability of any provider-specific interpretation and implies that organisational governance must treat "drift" as expected rather than exceptional.

Third, incident evidence is structurally incomplete. Data exposure events are under-reported, frequently handled under confidentiality constraints, and often lack technical detail sufficient for attribution beyond the initial disclosure. Even when an incident is documented, chain-of-custody across plug-ins, wrappers, internal relays, and endpoint artefacts is rarely reconstructable. Consequently, the review cannot support strong causal claims about downstream reuse, cross-organisational propagation, or long-run model-mediated diffusion; it can only argue that certain pathways and residual risks remain plausible under bounded conditions.

Fourth, the qualitative risk scoring is inherently context-dependent. Severity, likelihood, and observability are assessed on ordinal scales as a transparent prioritisation rubric, but these judgements depend on sector, documentation criticality, workforce incentives, endpoint hygiene, and the organisation's actual tooling stack. The same pathway can shift bands when, for example, a firm moves from ad hoc public chat use to a sanctioned enterprise endpoint with constrained connectors and inspection, or conversely when tool-using agents and broad OAuth scopes are introduced. The framework is therefore not a universal "risk calculator"; it is a reproducible way to make assumptions explicit and to keep prioritisation consistent within a given organisation over time.

Fifth, generalisability across jurisdictions and organisational maturities is limited. The compliance dimension is shaped by local legal regimes (e.g., data transfer constraints, sectoral regulations, export-control rules), and the feasibility of controls is shaped by baseline governance capacity (identity, DLP, endpoint management, procurement leverage). The paper's control recommendations therefore represent a minimal, mechanism-centred guardrail stack that is intended to be contractable and auditable in principle, but it does not claim that all organisations can implement all tiers without material resourcing and change management.

Finally, the paper does not empirically validate control effectiveness. Proposed mitigations (e.g., OCR-aware inspection, connector scoping, curated plug-in catalogues, logging minimisation, and tiered architectural reduction) are grounded in known control families and the mapped pathways, but the review does not provide

experimental measurements of false positives/negatives, usability impacts, or leakage reduction in operational deployments. That validation is a concrete direction for future work: controlled studies and field evaluations comparing governed vs. unguided adoption, especially under high-pressure operational contexts where incentives and sensitivity peak.

9. Conclusions

This paper maps the mechanisms through which the routine use of general-purpose LLMs by employees creates bidirectional data flows between internal documentation systems and external model ecosystems. The evidence suggests that the primary risks reside not in isolated incidents but in recurring transition points—such as copy–paste, file uploads, and connector invocation—integrated into daily workflows. The analysis confirms that four core risk dimensions (confidentiality, compliance, model-side effects, and governance gaps) require a transition from reactive prohibitions to proactive, workflow-based governance. The proposed "minimal guardrail stack" and evaluation framework offer organisations a verifiable method for risk triage and cautious adoption. Instead of claiming definitive leakage rates, this work provides a roadmap for explicit risk-based decision-making in a volatile technological landscape.

Author Contributions

Conceptualisation, L.L. and O.Z.; methodology, L.L.; validation, O.M., D.Ž.; investigation, L.L., O.Z., O.M., D.Ž.; data curation, O.M. and D.Ž; writing—original draft preparation, L.L.; writing—review and editing, O.Z., O.M; supervision, D.Ž. All authors have read and agreed to the published version of the manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare no conflict of interest.

Literature

1. Acar, G., Englehardt, S., & Narayanan, A. (2020). No boundaries: Data exfiltration by third parties embedded on web pages. *Proceedings on Privacy Enhancing Technologies*. https://petsymposium.org/popets/2020/popets-2020-0070.php
2. Agarwal, D., Fabbri, A., Risher, B., Laban, P., Joty, S., & Wu, C.-S. (2024). Prompt Leakage effect and mitigation strategies for multi-turn LLM Applications. In F. Dernoncourt, D. Preoțiuc-Pietro, & A. Shimorina (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 1255–1275). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-industry.94
3. Alzamil, L. M., Alhasani, A. M., Alshehri, S., Alzamil, L. M., Alhasani, A. M., & Alshehri, S. (2025). Privacy Concerns in ChatGPT Data Collection and Its Impact on Individuals. *Future Internet*, *17*(11). https://doi.org/10.3390/fi17110511
4. Aplin, T., Radauer, A., Bader, M. A., & Searle, N. (2023). The Role of EU Trade Secrets Law in the Data Economy: An Empirical Analysis. *IIC - International Review of Intellectual Property and Competition Law*, *54*(6), 826–858. https://doi.org/10.1007/s40319-023-01325-8
5. Baek, S. J., Lee, H. J., Baek, S. J., & Lee, H. J. (2025). Unravelling the Effects of Privacy Policies on Information Disclosure: Insights from E-Commerce Consumer Behavior. *Journal of Theoretical and Applied Electronic Commerce Research*, *20*(1). https://doi.org/10.3390/jtaer20010049
6. Bhardwaj, A., & Goundar, S. (2017). Security challenges for cloud-based email infrastructure. *Network Security*, *2017*(11), 8–15. https://doi.org/10.1016/S1353-4858(17)30094-6
7. Bhushan, B. (2025). An Explainable Zero Trust Identity Framework for LLMs, AI Agents, and Agentic AI Systems. *International Journal of Computer Applications*, *187*(46), 42–52.
8. Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work*. *The Quarterly Journal of Economics*, *140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

9. Challappa, L., Zhang, Z., & Garg, R. (2025). Domain anchorage in LLMs: Lexicon profiling and unintended information leakage. *Data & Policy*, *7*, e73. https://doi.org/10.1017/dap.2025.10041
10. Chen, K., Zhou, X., Lin, Y., Feng, S., Shen, L., & Wu, P. (2025). A survey on privacy risks and protection in large language models. *Journal of King Saud University Computer and Information Sciences*, *37*(7), 163. https://doi.org/10.1007/s44443-025-00177-1
11. Chen, T., Li, P., Zhou, K., Chen, T., & Wei, H. (2025). Unveiling Privacy Risks in Multi-modal Large Language Models: Task-specific Vulnerabilities and Mitigation Challenges. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 4573–4586). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.237
12. Chivers, H., & Hargreaves, C. (2011). Forensic data recovery from the Windows Search Database. *Digital Investigation*, *7*(3), 114–126. https://doi.org/10.1016/j.diin.2011.01.001
13. Clusmann, J., Ferber, D., Wiest, I. C., Schneider, C. V., Brinker, T. J., Foersch, S., Truhn, D., & Kather, J. N. (2025). Prompt injection attacks on vision language models in oncology. *Nature Communications*, *16*(1), 1239. https://doi.org/10.1038/s41467-024-55631-x
14. Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and Privacy Challenges of Large Language Models: A Survey. *ACM Comput. Surv.*, *57*(6), 152:1-152:39. https://doi.org/10.1145/3712001
15. *Data controls in the OpenAI platform*. (n.d.). Retrieved December 23, 2025, from https://platform.openai.com
16. Ebert, C. (2013). Improving engineering efficiency with PLM/ALM. *Software & Systems Modeling*, *12*(3), 443–449. https://doi.org/10.1007/s10270-013-0347-3
17. *Enterprise privacy at OpenAI*. (n.d.). Retrieved December 23, 2025, from https://openai.com/enterprise-privacy/
18. Feretzakis, G., Papaspyridis, K., Gkoulalas-Divanis, A., Verykios, V. S., Feretzakis, G., Papaspyridis, K., Gkoulalas-Divanis, A., & Verykios, V. S. (2024). Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review. *Information*, *15*(11). https://doi.org/10.3390/info15110697
19. Feretzakis, G., Vagena, E., Kalodanis, K., Peristera, P., Kalles, D., Anastasiou, A., Feretzakis, G., Vagena, E., Kalodanis, K., Peristera, P., Kalles, D., & Anastasiou, A. (2025). GDPR and Large Language Models: Technical and Legal Obstacles. *Future Internet*, *17*(4). https://doi.org/10.3390/fi17040151
20. Feretzakis, G., Verykios, V. S., Feretzakis, G., & Verykios, V. S. (2024). Trustworthy AI: Securing Sensitive Data in Large Language Models. *AI*, *5*(4), 2773–2800. https://doi.org/10.3390/ai5040134
21. Ferrag, M. A., Tihanyi, N., Hamouda, D., Maglaras, L., Lakas, A., & Debbah, M. (2025). From prompt injections to protocol exploits: Threats in LLM-powered AI agents workflows. *ICT Express*. https://doi.org/10.1016/j.icte.2025.12.001
22. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 79–90. https://doi.org/10.1145/3605764.3623985
23. Hullavarad, S., O'Hare, R., & Roy, A. K. (2015). Enterprise Content Management solutions—Roadmap strategy and implementation challenges. *International Journal of Information Management*, *35*(2), 260–265. https://doi.org/10.1016/j.ijinfomgt.2014.12.008
24. Hur, U., Kang, S., Kim, G., & Kim, J. (2023). A study on cloud data access through browser credential migration in Windows environment. *Forensic Science International: Digital Investigation*, *45*, 301568. https://doi.org/10.1016/j.fsidi.2023.301568
25. *Incident 768: ChatGPT Implicated in Samsung Data Leak of Source Code and Meeting Notes*. (2023). https://incidentdatabase.ai/cite/768/
26. *Incident 1186: Reported Public Exposure of Over 100,000 LLM Conversations via Share Links Indexed by Search Engines and Archived*. (2025). https://incidentdatabase.ai/cite/1186/
27. Ishrak Alim, T. F. M., Takib Md Masudul Hasan Prodhan ,. Md Lahaduzzaman Lahad. (2025). *The Insider Risk of Artificial Intelligence in Financial Systems through the Lens of Large Language Models*. https://doi.org/10.5281/zenodo.16910637
28. Karras, A., Theodorakopoulos, L., Karras, C., Krimpas, G. A., Giannaros, A., Bakalis, C.-P., Karras, A., Theodorakopoulos, L., Karras, C., Krimpas, G. A., Giannaros, A., & Bakalis, C.-P. (2025). LLM-Driven Big Data Management Across Digital Governance, Marketing, and Accounting: A Spark-Orchestrated Framework. *Algorithms*, *18*(12). https://doi.org/10.3390/a18120791

29. Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., & Oh, S. J. (2023). ProPILE: Probing Privacy Leakage in Large Language Models. *In Large Language Models. NeurIPS 2023 Proceedings.*
30. Kim, Y., Yoon, H.-J., & Lee, M.-H. (2015). *Stealthy Information Leakage from Android Smartphone through Screenshot and OCR*. 784–787. https://doi.org/10.2991/cmfe-15.2015.184
31. Kramcsák, P. T. (2023). Can legitimate interest be an appropriate lawful basis for processing Artificial Intelligence training datasets? *Computer Law & Security Review*, *48*, 105765. https://doi.org/10.1016/j.clsr.2022.105765
32. Kuru, T. (2024). Lawfulness of the mass processing of publicly accessible online data to train large language models. *International Data Privacy Law*, *14*(4), 326–351. https://doi.org/10.1093/idpl/ipae013
33. Liu, Y., Geng, R., Jia, J., & Gong, N. Z. (2024). *Formalizing and Benchmarking Prompt Injection Attacks and Defenses*.
34. Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). *Analyzing Leakage of Personally Identifiable Information in Language Models* (No. arXiv:2302.00539). arXiv. https://doi.org/10.48550/arXiv.2302.00539
35. Malki, L. M., Polamarasetty, A., Hatamian, M., Warner, M., & Costanza, E. (2025a). Hoovered up as a data point: Exploring Privacy Behaviours, Awareness, and Concerns Among UK Users of LLM-based Conversational Agents. *Proceedings on Privacy Enhancing Technologies*. https://petsymposium.org/popets/2025/popets-2025-0160.php
36. Malki, L. M., Polamarasetty, A., Hatamian, M., Warner, M., & Costanza, E. (2025b). Hoovered up as a data point: Exploring Privacy Behaviours, Awareness, and Concerns Among UK Users of LLM-based Conversational Agents. *Proceedings on Privacy Enhancing Technologies*. https://petsymposium.org/popets/2025/popets-2025-0160.php
37. Mendoza, A., Kumar, A., Midcap, D., Cho, H., & Varol, C. (2015). BrowStEx: A tool to aggregate browser storage artifacts for forensic analysis. *Digital Investigation*, *14*, 63–75. https://doi.org/10.1016/j.diin.2015.08.001
38. Mohamed, K. F., AbdelBaki, N., & Shosha, A. (2023). Clipboard Data Attacks and Detection via Remote Desktop Protocol. *2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 98–102. https://doi.org/10.1109/NILES59815.2023.10296672
39. Mokhtar, U. A., & Yusof, Z. M. (2015). The requirement for developing functional records classification. *International Journal of Information Management*, *35*(4), 403–407. https://doi.org/10.1016/j.ijinfomgt.2015.04.002
40. Mousavi, Z., Islam, C., Babar, M. A., Abuadbba, A., & Moore, K. (2025). Detecting Misuse of Security APIs: A Systematic Review. *ACM Comput. Surv.*, *57*(12), 303:1-303:39. https://doi.org/10.1145/3735968
41. Nasr, M., Rando, J., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Tramer, F., & Lee, K. (2025). *SCALABLE EXTRACTION OF TRAINING DATA FROM ALIGNED, PRODUCTION LANGUAGE MODELS*.
42. Nealey, T., Daignault, R. M., & Cai, Y. (2015). Trade Secrets in Life Science and Pharmaceutical Companies. *Cold Spring Harbor Perspectives in Medicine*, *5*(4), a020982–a020982. https://doi.org/10.1101/cshperspect.a020982
43. Oh, J., Lee, S., & Lee, S. (2011). Advanced evidence collection and analysis of web browser activity. *Digital Investigation*, *8*, S62–S70. https://doi.org/10.1016/j.diin.2011.05.008
44. Okolica, J., & Peterson, G. L. (2011). Extracting the windows clipboard from physical memory. *Digital Investigation*, *8*, S118–S124. https://doi.org/10.1016/j.diin.2011.05.014
45. Ozcan, O., Pickernell, D., & Bacon, E. (2025). Identifying trade secrets: Strategic process and challenges in the UK. *Technology Analysis & Strategic Management*, *0*(0), 1–18. https://doi.org/10.1080/09537325.2025.2489155
46. Pahune, S., Akhtar, Z., Pahune, S., & Akhtar, Z. (2025). Transitioning from MLOps to LLMOps: Navigating the Unique Challenges of Large Language Models. *Information*, *16*(2). https://doi.org/10.3390/info16020087
47. Pedro, R., Coimbra, M. E., Castro, D., Carreira, P., & Santos, N. (2025). Prompt-to-SQL Injections in LLM-Integrated Web Applications: Risks and Defenses. *In Proceedings of the IEEE/ACM International Conference on Software Engineering (ICSE 2025)*.
48. Perry, N., Srivastava, M., Kumar, D., & Boneh, D. (2023). Do Users Write More Insecure Code with AI Assistants? *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2785–2799. https://doi.org/10.1145/3576915.3623157

49. Prinz, K. D. (2025). Managing the legal risks of artificial intelligence on intellectual property and confidential information. *Consulting Psychology Journal*, *77*(2), 169–179. https://doi.org/10.1037/cpb0000287
50. Rathod, V., Nabavirazavi, S., Zad, S., & Iyengar, S. S. (2025). Privacy and Security Challenges in Large Language Models. *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, 00746–00752. https://doi.org/10.1109/CCWC62904.2025.10903912
51. Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003
52. Sai, S., Yashvardhan, U., Chamola, V., & Sikdar, B. (2024). Generative AI for Cyber Security: Analyzing the Potential of ChatGPT, DALL-E, and Other Models for Enhancing the Security Space. *IEEE Access*, *12*, 53497–53516. https://doi.org/10.1109/ACCESS.2024.3385107
53. Sajjadi Mohammadabadi, S. M., Kara, B. C., Eyupoglu, C., Uzay, C., Tosun, M. S., & Karakuş, O. (2025). A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges, and Societal Implications. *Electronics*, *14*(18), 3580. https://doi.org/10.3390/electronics14183580
54. Saka, A., Taiwo, R., Saka, N., Salami, B. A., Ajayi, S., Akande, K., & Kazemi, H. (2024). GPT models in construction industry: Opportunities, limitations, and a use case validation. *Developments in the Built Environment*, *17*, 100300. https://doi.org/10.1016/j.dibe.2023.100300
55. Salminen, A., Jauhiainen, E., & Nurmeksela, R. (2014). A life cycle model of XML documents. *Journal of the Association for Information Science and Technology*, *65*(12), 2564–2580. https://doi.org/10.1002/asi.23148
56. Shvetsova, O., Katalshov, D., Lee, S.-K., Shvetsova, O., Katalshov, D., & Lee, S.-K. (2025). Innovative Guardrails for Generative AI: Designing an Intelligent Filter for Safe and Responsible LLM Deployment. *Applied Sciences*, *15*(13). https://doi.org/10.3390/app15137298
57. Sovrano, F., Hine, E., Anzolut, S., & Bacchelli, A. (2025). Simplifying software compliance: AI technologies in drafting technical documentation for the AI Act. *Empirical Software Engineering*, *30*(4), 91. https://doi.org/10.1007/s10664-025-10645-x
58. Staab, R., Vero, M., Balunovic, M., & Vechev, M. (2024). *BEYOND MEMORIZATION: VIOLATING PRIVACY VIA INFERENCE WITH LARGE LANGUAGE MODELS*.
59. Standing, C., & Kiniti, S. (2011). How can organizations use wikis for innovation? *Technovation*, *31*(7), 287–295. https://doi.org/10.1016/j.technovation.2011.02.005
60. Starov, O., & Nikiforakis, N. (2017). Extended Tracking Powers: Measuring the Privacy Diffusion Enabled by Browser Extensions. *Proceedings of the 26th International Conference on World Wide Web*, 1481–1490. https://doi.org/10.1145/3038912.3052596
61. Taeihagh, A. (2025). Governance of Generative AI. *Policy and Society*, *44*(1), 1–22. https://doi.org/10.1093/polsoc/puaf001
62. Torraco, R. J. (2005). Writing Integrative Literature Reviews: Guidelines and Examples. *Human Resource Development Review*, *4*(3), 356–367. https://doi.org/10.1177/1534484305278283
63. Wan, L., Wang, K., Wang, H., & Bai, G. (2024). Is It Safe to Share Your Files? An Empirical Security Analysis of Google Workspace. *Proceedings of the ACM Web Conference 2024*, 1892–1901. https://doi.org/10.1145/3589334.3645697
64. Wang, Z., Liu, T., Liu, Y., Zio, E., & Guan, X. (2025). Data Inference: Data Security Threats in the AI Era. *Engineering*, *52*, 29–33. https://doi.org/10.1016/j.eng.2025.08.007
65. Waters-Lynch, J., Allen, D. W. E., Potts, J., & Berg, C. (2025). *Shadow User Innovation: Governing Covert Generative-AI Use for Dynamic-Capability Renewal* (SSRN Scholarly Paper No. 5281695). Social Science Research Network. https://doi.org/10.2139/ssrn.5281695
66. Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing*, *52*(5), 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x
67. Williams, A., Fox, G., Amon, M. J., Tanni, T. I., & Solihin, Y. (2025). The GenAI networked privacy problem at work- How privacy knowledge and perceptions predict Generative AI disclosure in professional contexts. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3706599.3719923
68. Yang, B., Dang, J., Liu, H., & Jin, Z. (2025). Advancing LLM-Generated Code Reliability: A Hybrid Approach for Hallucination Detection. *IEEE Transactions on Software Engineering*, 1–17. https://doi.org/10.1109/TSE.2025.3640641

69. Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Trans. Knowl. Discov. Data*, *18*(6), 160:1-160:32. https://doi.org/10.1145/3649506
70. Yang, R., Fu, M., Tantithamthavorn, C., Arora, C., Vandenhurk, L., & Chua, J. (2025). RAGVA: Engineering retrieval augmented generation-based virtual assistants in practice. *Journal of Systems and Software*, *226*, 112436. https://doi.org/10.1016/j.jss.2025.112436
71. Zhan, Q., Liang, Z., Ying, Z., & Kang, D. (2024). *InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents* (No. arXiv:2403.02691). arXiv. https://doi.org/10.48550/arXiv.2403.02691
72. Zhang, A. (2025). Information Retrieval in the Age of Generative AI: A Mismatch That Matters. *Legal Reference Services Quarterly*, *44*(3), 297–306. https://doi.org/10.1080/0270319X.2025.2536920
73. Zhang, H., Huang, J., Mei, K., Yao, Y., Wang, Z., Zhan, C., Wang, H., & Zhang, Y. (2025). *AGENT SECURITY BENCH (ASB): FORMALIZING AND BENCHMARKING ATTACKS AND DEFENSES IN LLM-BASED AGENTS*.
74. Zolkifli, N. N., Ngah, A., & Deraman, A. (2018). Version Control System: A Review. *Procedia Computer Science*, *135*, 408–415. https://doi.org/10.1016/j.procs.2018.08.19

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:001.895]:502/504
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601095S

Upotreba naprednih tehnologija u konceptima bezbednosti životne sredine

Slobodan Simić[1]

[1]Security Research Center, Republic of Srpska, Bosnia and Herzegovina, slobodansimicss@yahoo.com

Apstrakt: Ubrzani razvoj u svetu velikim delom se zasniva na tehničko-tehnološkim inovacijama i dostignućima. Uzročno-posledične veze se utvrđuju u mnogim sferama života, posebno u segmentima koji su od suštinskog značaja za opstanak živih bića. Evidentno je da je tradicionalni pristup razmatranju ekološke bezbednosti u velikoj meri prevaziden, da je ekološka dinamika pojačana usled stanja životne sredine i da je neophodno početi sagledavanje segmenata ekološke bezbednosti i sastojaka životne sredine uz upotrebu naprednih tehnologija. Razloga za ove tvrdnje ima mnogo i oni su ustanovljeni u rasponu od fizičke zaštite životne sredine do prikupljanja, obrade, analize i distribucije velike količine informacija o stanju prirodne sredine u realno kratkom vremenu. Na ovom kontekstualnom nivou postaje celishodno korišćenje veštačke inteligencije u širokom spektru, od zaštite biodiverziteta do monitoringa životne sredine u kompatibilnom spektru aktivnosti. Softverske aplikacije postaju alat u rukama naučnih i stručnih radnika u oblasti ekološke bezbednosti. Višedimenzionalna i višekriterijumska analiza koju veštačka inteligencija poseduje u svojoj genezi omogućava praćenje životne sredine u planskim, organizacionim i realizacionim aktivnostima, unapređuje strateško odlučivanje i usmeravanje aktivnosti na globalnom, regionalnom i lokalnom nivou. Istraživanja u ovoj oblasti imaju značajnu osnovu jer će naredni period doneti izazove, rizike i pretnje koji, zbog svog intenziteta, potrebe za prikupljanjem i obradom podataka i zahteva akcionog pristupa, prevazilaze mogućnosti da čovek, ljudsko biće, blagovremeno razume različite uticaje na ljude, životinje i biljke. Gore pomenute konstrukte dodatno potkrepljuje činjenica da će veštačka inteligencija pomoći u obradi uticaja na živa bića iz svemira.
Cilj ovog rada je imperativ korišćenja veštačke inteligencije u oblasti ekološke bezbednosti, prema ranjivosti životnih ciklusa komplikovanih globalnim promenama u prirodnoj sredini.

Ključne reči: veštačka inteligencija, ekološka bezbednost

Use Of Advanced Technologies in Concepts of Environmental Security

Abstract: Accelerated development in the world is largely based on technical-technological innovations and achievements. Cause-and-effect relationships are determined in many spheres of life, especially in segments that are essential for the survival of living beings. It is evident that the traditional approach to considering environmental safety has largely been overcome, that ecological dynamics have increased due to the state of the environment, and that it is necessary to start looking at the segments of ecological safety and the constituents of the environment with the use of advanced technologies. There are many reasons for these statements and they are established in the range of physical environmental protection to the collection, processing, analysis and distribution of a large amount of information about the state of the natural environment in a realistically short time. In this contextual level, it becomes expedient to use artificial intelligence in a wide range, from biodiversity protection to environmental monitoring in a compatible spectrum of activities. Software applications become a tool in the hands of scientific and professional workers in the field of environmental security. The multi-dimensional and multi-criteria analysis that artificial intelligence possesses in its genesis enables monitoring of the environment in planning, organizational and implementation activities, improves strategic decision-making and directing activities at the global, regional and local level. Research in this field has a significant basis because the next period will bring challenges, risks and threats that, due to their intensity, the need to collect and process data and the requirements of an action approach, exceed the possibilities for man, a human being, to understand in a timely manner the different impacts on people, animals and plants. The above-mentioned constructs are additionally supported by the fact that artificial intelligence will help in processing the impact on living beings from space.

The goal of this work is the imperative use of artificial intelligence in the field of environmental security, according to the vulnerability of life cycles complicated by global changes in the natural environment.

Keywords: artificial intelligence, ecological security

1. Introduction

Scientific and professional workers agree that crisis moments have come for certain natural communities and that their further endangerment will produce serious consequences for the ecosystem of the planet. Coral reefs, changes in the course and intensity of the main ocean currents, the reduction of ice sheets, especially in the south and north poles and mountain massifs, and the systematic destruction of the Amazon forest are global problems, but we must not forget the derogation of the taiga and steppes, endangering the survival of certain species of plants and animals. Through ecological security, the state of natural communities is looked at and efforts are made to protect this value on the planet through planning and preventive activities. The introduction of scientific and technological achievements focused through the use of artificial intelligence gives the opportunity to look at potential threats from a primarily scientific and practical point of view and take steps to protect the environment. Artificial intelligence has not only made it possible to collect and process data on individual habitats, but is also becoming an indispensable tool in monitoring and assessing the state of biodiversity. This is particularly reflected in the most extreme conditions, geographical locations that are rarely monitored by humans, places where the application of laws is sometimes difficult on this basis, as well as outer space that is not considered by international law.

2. Theoretical approach to ecological security and the need for the use of artificial intelligence

It is evident that the living and working environment on Earth is changing and that the indirect and direct impacts of humans on the environment are becoming increasingly destructive. Industrialization and globalization in the late 19th and early 20th centuries improved living conditions, the flow of goods and capital, and service activities, but also brought the roots of confrontations, social disturbances, and harmful effects on the environment.

Scientific and research analyses and reports from related sciences and scientific disciplines warn us that life on the Planet is becoming more complex due to the inadequate and unplanned use of ecological and energy resources on the globe. This is supported by the statement that the conclusions of the COP (Conference of the Parties) 29 held in Baku, Azerbaijan, did not bring improvements except in conceptual ideas and declarative statements. These conclusions are summarized in the following:

- One of the key objectives of COP29 was to define a new collective quantified target for climate finance, replacing the current target of $100 billion per year. This new target takes into account the growing needs of developing countries, especially those most affected by the consequences of climate change, such as droughts, floods and storms;
- The issue of loss and damage, which includes financial support for countries facing the irreversible consequences of climate change, also took center stage. The Loss and Damage Response Fund, established at the previous COP, received new commitments and pledges from developed countries, but the funds are still not proportional to the estimated needs, which could reach as much as $580 billion per year by 2030;
- Nationally Determined Contributions (NDCs) were presented and improved in Baku, with an emphasis on increasing the ambition for emission reductions by 2030 and 2035. These plans aim to limit global warming to 1.5°C, but current commitments by many countries are still insufficient to achieve this goal;
- COP29 highlighted the need for further international cooperation, particularly in light of geopolitical challenges and uncertainty about the engagement of the United States after the elections (https://rcel.rs/sr/cop29-kljucni-zakljucci/).

This year, in 2025, six thematic units were dominant at COP 30 in Brazil:

- How to prevent accelerated global warming;
- How to protect communities from climate change;
- How to produce material goods for a trillion dollars;
- How to increase the strength of creative solutions according to climate change;
- How to ensure a fair and inclusive transition;
- How to revive the conclusions from Paris. (https://pocketproject.org/.).

Although, evidently, strategic ideas are determined, certain practical steps are difficult to implement by adopting strategies, policies and regulations. Also, standard operating procedures for environmental protection are not rarely implemented in conditions of difficult financing. National Ecosystem Assessments synthesize key knowledge about biodiversity and ecosystem services to enable full consideration of natural values in decision-making. International entities are mostly burdened with financing other items in the budget, while non-governmental organizations and associations at the world, regional and local level depend on donor funding. Also, it is visible from the strategic guidelines and conclusions that there are no special expenditures for innovation and technical-technological development, which points to the possibility that certain approaches are viewed monopolistically and that these segments are designated only for the privileged. I would therefore consider it expedient to point out that technical-technological development in this field should be determined according to the principle of "accessibility for all" because we all depend on this Planet and considered at the world level.

Even in the conclusions of COP 30, the more comprehensive and expedient use of software packages, monitoring of environmental constituents in different natural environments or laboratory achievements in the field of environmental forensics was not distinctively presented. The aforementioned conclusions point, however, to the readiness for a more active approach in the field of environmental safety:

- National Climate Plans;
- Adaptation;
- Finance;
- Nature;
- People-Centered Action;
- Trade;
- Sectoral Action (https://www.wri.org/insights/cop30-outcomes-next-steps)

Analyzing the available materials presented, it is evident that there is no direct mention of the use of new technologies or, decisively, artificial intelligence anywhere. Indirect content indicates that artificial intelligence is presented within the main macroeconomic areas with an emphasis on its key role in the transition to a circular and sustainable economy.

A significant shift has certainly been made in the United Nations programs, where an innovative and inspiring approach to the protection of living communities is affirmed through various content.

National ecosystem assessments are supported by UNEP-WCMC's NEA Initiative. Leveraging the expertise of the Sub-Global Assessment Network (SGAN), global activities encompass:

- A capacity-building programme that includes regular workshops, a dedicated website hosting knowledge and resources, fellowship programmes and exchange visits
- A series of global training workshops with country partners and up-to-date training materials on national ecosystem assessments made available online
- Knowledge-sharing and case studies disseminated through different channels, such as webinars, workshops and side events at international and regional biodiversity conferences
- Supporting national engagement with international processes, IPBES and CBD (https://www.besnet.world/national-ecosystem-assessment/).

3. Artificial intelligence as an imperative for monitoring segments of environmental security

The application of technical protection systems for the purpose of environmental safety is not a novelty. Technical systems (communication, integrated, computer-supported, etc.) in many ways give a more accurate and stable picture of the ecological reflection of a geographical climate and indicate certain anomalies and the need for protection.

The state of the technical system, in this contextual level, represents a set of data that provide complete information about the behavior of the system at a given moment in time and given environmental conditions, the need to adjust the system, i.e. projecting its behavior in the period starting from that moment. The indeterminacy of the state implies the degree of realization of the given conditions and procedures characteristic of certain states. The entropy of the system is a quantity that determines the measure of the determinacy of the system, and is based on the stochastic behavior of the system. The basic states of the system are determined by changing the parameters of the force function in time, under the influence of different magnitude, direction and direction, whereby:

- Changing the parameters of the objective function within the allowed limits determines the state of the system "satisfies", which means that the system successfully performs the criterion function;
- Changing the parameters of the set criteria function beyond the limit of permissible deviations determines the state of the system "does not satisfy", which means that the system does not successfully perform the function set by the criteria. The condition does not satisfy means the condition of cancellation (Adamović, Gavrić, Grbavac, 2009:13).

However, the natural environment is much more sensitive and reacts depending on many factors. Therefore, it is necessary to carry out: transformation of certain technical systems (eg updating software packages due to climate change), adaptation due to dynamics and changes in plant, animal and human communities; reconstruction due to the conditions in which certain technical systems work (in desert conditions, at the North or South Pole, under the surface of the sea, etc.).

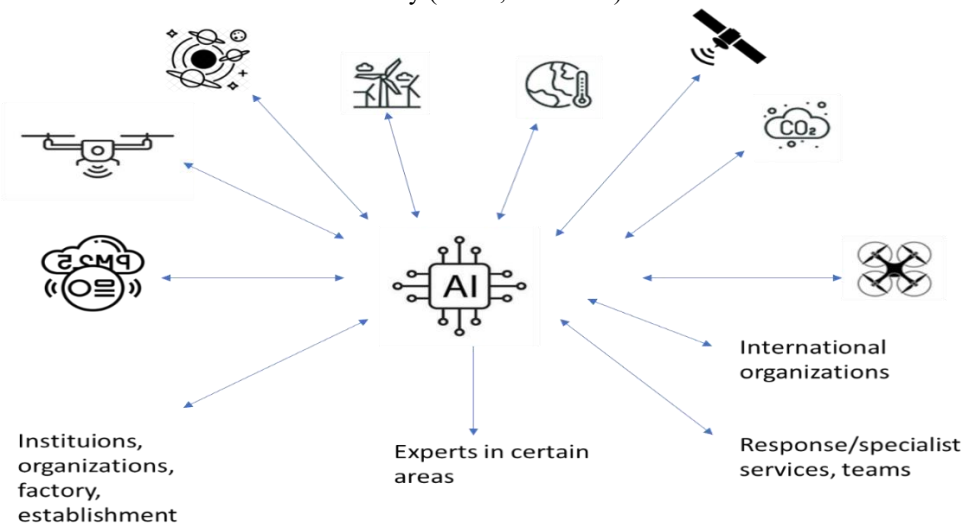
For example, there is a desire to invest in technical-technological solutions using the full range of application of modern technologies through the application of artificial intelligence.

Based on the use of artificial intelligence to predict climate change, (Kyungmee and Boulanin 2023) list several key points:

- Understanding and forecasting the impact of climate hazards;
- Managing vulnerabilities and exposure to climate change;
- Entry points for artificial intelligence in response to climate change security risks;
- Tools for mapping climate hazards and socio-economic stressors;
- Managing climate change-related disaster risks;
- Examples of policy interventions addressing climate change-related security risks;
- Challenges associated with artificial intelligence.

Essentially, artificial intelligence integrates a much broader spectrum of interdependent information and communication devices whose collected information is processed and analyzed. Figure 1 shows a simplified representation of such a model.

Figure 1: Simplified representation of the application of artificial intelligence in the field of environmental safety (Simic, 2025:168).



Dassault Systèmes connects virtual twins of physical and digital systems in a collaborative virtual world to simulate complex risk scenarios, explore prevention plans, and orchestrate the optimal deployment of resources:

- Cybersecure AI platforms (Thales) manage autonomous drone operations and orchestrate tactical missions, analyzing multi-source data (drones, satellites, etc.) in real time to better detect risks and anticipate their evolution - even from mobile, decentralized command centers like the R4;
- The Flux Vision (Orange) analysis tools and mission planning tools (HawAI.tech) optimize drone flight paths taking into account all mission constraints;

- A crisis management solution (Atos) integrates prevention plans, monitoring, and simulation data to organize emergency responses (https://www.thalesgroup.com/en/news-centre/press-releases/software-republique-unveils-vision-4rescue-integrated-technological).

4. Threat to environmental security in specific ecological communities with controversy on the application of artificial intelligence and international law

Contemporary risks to the constituents of environmental security are viewed on a multivariate basis. These points of view are based on the facts and assumptions that in the future phenomena and processes that threaten a stable security environment, and thus the environment, will come from outer space. The readiness to respond to these challenges depends on the possibility of collecting a huge number of undefined and unexplained data, but also on the need for international legal regulation of the use of outer space in order to influence the consequences for the natural environment (see more - https://www.unoosa.org/oosa/en/ourwork/spacelaw/index.html).

Given that there are specific conditions in space that have not yet been fully explored, certain modalities for software packages and analyzes implemented through artificial intelligence can be used from data, for example, from sub-Saharan Africa and from Antarctica and the Arctic. The collection, processing and analysis of data as well as its use has been greatly improved by the use of artificial intelligence.

Arctic weather is inseparable from security. Polar nights, light and winds began to be subject to the power and interests of states and other actors on land, sea, air and space. Free navigation patrols, subsea traffic, increased air traffic and communications networks depend on reliable forecasts and warnings, but these activities affect flora and fauna. The proximity of people to equipment undoubtedly affects plant and animal communities in addition to climate change. In this environment, risk awareness and assessment of the survival of individual species is a key factor for enabling their survival in harsh conditions.

Artificial intelligence (AI) and machine learning (ML) technologies are reshaping core practices and infrastructure worldwide, but in the Arctic, this transformation is not merely disruptive – it is decisive. The region is at once climatically unstable, geopolitically contested and operationally fragile. Accurate weather, water, and climate (WWC) services form the backbone of safety, mobility, and sovereignty in the High North. As Arctic stakeholders turn increasingly to AI/ML to enhance decision-making support, they are rapidly building dependencies on systems in an institutional vacuum. In a region where the margin for error is slim and the consequences of failure severe, these shifts demand urgent policy attention. At the same time, international law is struggling to keep pace. No treaty regime governs algorithmic decision-making in the Arctic, and the patchwork of relevant norms were not designed for this convergence of technical innovation, commercial incentives, and environmental exigency. This Policy Brief situates the algorithmic transformation of Arctic WWC services within its broader legal and security context. It concludes by outlining three domains for policy intervention: international organisations, international law and international security (Linch, Norchi, 2025:3).

International law facilitates patterns of authority and control that purport to impose stability, predictability, and continuity in an otherwise unorganised global arena. The Arctic WWC value cycle relies on a data chain whose every link is entangled in overlapping sovereignty, sharing, and privacy regimes underpinned by international law – the authoritative decisions of the world community expressed in conventions and custom. It supplies the scaffolding for these vital services. Unlike Antarctica, there is no dedicated treaty regime for the Arctic; instead, the region is governed by generally applicable and certain specialised legal instruments. As a general matter, the international law that applies to other regions of the world is equally applicable to the Arctic. States are under an obligation to refrain “from the threat or use of force against the territorial integrity or political independence of any state” (UN Charter, art. 2(4)) and to “settle their international disputes by peaceful means” (UN Charter, art. 2(3)). The law of the sea, the law of treaties, the law of state responsibility, international human rights law, the law of armed conflict and every other branch of international law apply in the Arctic as in other regions of the world. Are there points of convergence between Arctic law and international law applicable to AI/ML in the WWC value cycle? (Linch, Norchi, 2025:9).

5. Conclusion

Changes in the segments of environmental safety are now more frequent and with greater intensity, and the question arises whether we can monitor them with valid quantitative and qualitative indicators. There is no doubt that advanced technology established in artificial intelligence can help in many spheres of environmental protection and provide answers that will enable decision makers to make a more purposeful assessment and guide taking steps to protect the construct of the natural environment. The possibility of collecting a large amount of

information, its processing and analysis and providing reliable indicators to decision makers for environmental protection will significantly increase the use of artificial intelligence in the natural and space environment.

The development of artificial intelligence software solutions in this area will certainly imply an even greater integration of information and communication systems, but also an increasing application of multi-layered advanced neural networks and evolutionary algorithms. This will be especially significant in the extreme conditions of existence of living beings and the effects of natural threats that may come from outer space. The tripartite nature of ecological security, the use of artificial intelligence and the (international) legal aspect will obviously be determined through an integrative, multi-dimensional and multi-dependent perspective.

References

1. Адамовић, Ж., Гаврић, М., Грбавац, Ж, (2009), *Сигурност функционисања техничких система*, Академија инжињерства одржавања, Београд.
2. Kyngmee, K. and Boulanin V., (2023), *Artificial Intelligence for Climate Security*, Stockholm International Peace Research Institute, Sweden.
3. Linch, A., Norchi, C.,(2025), Algorithm North: Weather, Security an International Law, *In: GCSP Policy Brief 21*, Geneva.
4. Simic, S., (2025), Impact of artificial intelligence on environmental security segments, In: *11th International Scientific Professional Conference, "Security and Crisis Management-Theory and Practice"*, Belgrade.
5. Симић, С., (2025), *Еколошка безбједност*, Регионална асоцијација за безбједност и кризни менаџмент, Београд.

Internet presentations

1. https://www.thalesgroup.com/en/news-centre/press-releases/software-republique-unveils-vision-4rescue-integrated-technological (Accessed 05.12.2025.);
2. https://www.wri.org/insights/cop30-outcomes-next-steps (Accessed 04.12.2025.);
3. https://www.besnet.world/national-ecosystem-assessment/ (Accessed 04.12.2025.);
4. https://rcel.rs/sr/cop29-kljucni-zakljucci/ (Accessed 08.12.2025.);
5. https://pocketproject.org/ (Accessed 09.12.2025.);
6. https://www.unoosa.org/oosa/en/ourwork/spacelaw/index.html (Accessed 13.12.2025.).

Original Scientific Paper/Original naučni rad UDC/UDK: 341:004.8
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601101S

Digitalni *jus pacis*: međunarodna saradnja i pravni temelji mira u doba veštačke inteligencije

Troy Smith[1], Mikhail Byng[2]

[1] Ministry of Homeland Security, Trinidad and Tobago, dr.troy.smith@outlook.com
[2] University of the West Indies, Trinidad and Tobago, byngmikhail@gmail.com

Abstract: Rad razmatra transformativni uticaj veštačke inteligencije (VI) na globalnu bezbednost, kao i postojeće praznine u zakonodavnim okvirima za upravljanje sve izraženijim pretnjama u ovoj oblasti, naročito ograničenu sposobnost da se na adekvatan način reguliše tzv. „algoritamsko ratovanje". U radu se uvodi koncept *digitalnog jus pacis* kao novog pravno-etičkog okvira usmerenog na očuvanje mira u eri veštačke inteligencije. Ovaj koncept se nadovezuje na *jus ad bellum* i *jus in bello*, uz poseban naglasak na sprečavanje digitalne eskalacije, obezbeđivanje odgovornosti i zaštitu ljudskog dostojanstva u sajber prostoru. Uzimajući u obzir geopolitički i strateški kontekst u kojem se ove tehnologije razvijaju, rad predlaže tri ključna stuba algoritamskog mira: (1) zakonit dizajn (ljudski nadzor i proporcionalnost); (2) kooperativno upravljanje (regionalno i globalno usklađivanje); i (3) moralno uzdržavanje (digitalna humanitarna etika). Cilj rada je da se koncept digitalnog *jus pacis* razvije i pozicionira kao temeljni princip savremenog međunarodnog prava, čime bi se postavila osnova za budući Sporazum o miru u oblasti veštačke inteligencije ili za njegovu integraciju u Globalni digitalni sporazum Ujedinjenih nacija (GDC).

Ključne reči: veštačka inteligencija; algoritamsko ratovanje; digitalni *jus pacis*; globalno upravljanje; regulatorni okvir

Digital Jus Pacis: International Cooperation and Legal Foundations for Peace in the Age of Artificial Intelligence

Abstract: This paper aims to explore the transformative effect of artificial intelligence (AI) on global security and the existing gaps within legislative frameworks for managing the growing threats in this area, particularly the inability to manage "algorithmic warfare." Introducing the concept of a digital jus pacis, a new legal-ethical framework for preserving peace in the AI era. This proposed concept builds upon jus ad bellum and jus in bello, focusing on preventing digital escalation, ensuring accountability, and protecting human dignity in cyberspace. Whilst accounting for the geopolitical and strategic context within which this new technology is being developed, the article proposes three pillars for algorithmic peace, these include: 1) Lawful design (human oversight and proportionality); 2) Cooperative governance (regional and global alignment); 3) Moral restraint (digital humanitarian ethics). The goal is to develop and position the concept of digital jus pacis as a core principle of modern international law, which lays the foundation for a future AI Peace Compact or integration into the UN Global Digital Compact (GDC).

Keywords: Artificial Intelligence; Algorithmic Warfare; Digital Jus Pacis; Global Governance; Regulatory Framework

1. Introduction

Artificial intelligence (AI) is rapidly transforming global security, from autonomous weapons systems (AWS) and AI-driven security and intelligence analysis to cognitive warfare and information operations. The growth of AI and its integration into critical sectors within both the public and private sectors increases the importance of understanding this new technology and its potential impact on the world. These developments have already begun to reshape how conflict is waged, how threats are perceived, and how states and non-state actors compete in both physical and digital domains. The speed, opacity and scale of these AI-enabled capabilities also expose deep gaps in existing legal and regulatory frameworks, particularly those governing the lawful use of force, accountability,

and international cooperation. It appears that AI is advancing at such a rapid rate that regulatory frameworks are unable to adjust adequately or in a timely fashion to prevent the unfolding of worst-case scenarios or less desirable outcomes.

Classical *Just War* doctrines, such as *jus ad bellum* and *jus in bello*, which were primarily developed in relation to kinetic warfare, struggle to address algorithmic targeting, weaponised data, and AI-driven escalation dynamics (Bode & Bhila, 2024; Smith, 2025a). At the same time, new global governance initiatives are emerging to regulate AI more broadly, including the European Union AI Act, the Council of Europe's Framework Convention on AI, the United Nations Secretary-General's High-Level Advisory Body on AI, the United Nations Independent International Scientific Panel on Artificial Intelligence, and regional frameworks in Latin America and the Caribbean (European Parliament, 2024; ECLAC, 2024; United Nations High-Level Advisory Body on AI, 2024). However, the regulation of AI qualified by the need to present stymying innovation has not yet reached a nexus of the required understanding and applicability of controls for AI use in conflict.

Furthermore, whilst much of the academic and policy attention is placed on the strategic competition between the United States (US) and China in relation to technological progress in the domain of advanced technologies, middle powers and smaller states are actively engaging in these areas, further heightening the need for frameworks to guide the development and application of these new technologies. Countries such as Singapore, Norway, the United Kingdom, Turkey, Israel and others are actively making rapid progress in the field, further emphasising the need for collective and global action (Ratska & Bitzinger, 2023). What normative guidelines, legislative frameworks, and laws exist to address the paucity of guardrails regarding these new technologies? Furthermore, what recommendations can be put forth to address the absence of regulations and rules that provide a common ground for states and even private sector entities to engage AI tools? The importance of the answers to these questions is not in the development of country or region-specific solutions but in the applicability on the global scale to maintain peace through alignment with existing anti-war, just war, and humanitarian frameworks

Despite the proliferation of AI ethics guidelines, policy statements, and emerging regulatory instruments, there is an absence of coherent peace-oriented legal frameworks specifically designed to prevent AI-driven conflict/escalation prior to the outbreak of hostilities. Against this backdrop, the concept of *digital jus pacis* provides a means to extend peace-oriented legal thinking into the digital and algorithmic realm. Rather than focusing solely on the conditions under which force may be used (*jus ad bellum*) or how it should be constrained once conflict has begun (*jus in bello*), *digital jus pacis* foregrounds the law and norm-building required to prevent AI-driven escalation, protect human dignity, and preserve the informational and cognitive foundations of peace. Notably, international cooperation is increasingly subjugated to the narrow realist political and economic considerations of nations which maintain the capacity to affect the direction of development in the field. Hence, the need to emphasise international cooperation, particularly in selective domains of high impact. This is essential even in times of international flux, increased geopolitical tensions and political uncertainty, as a lack of guardrails in these areas will likely lead to further uncertainty and costly unintended consequences (Dookeran & Byng, 2025). In response, this paper seeks to address the existing gap by advancing *digital jus pacis* as a distinct and necessary complement to existing just war doctrines and contemporary AI governance regimes.

2. Key Definitions and Conceptual Frameworks

2.1. Jus ad Bellum, Jus in Bello and Jus Ante Bellum

At the core of the discussion on the ethics of warfare is the concept of "Just War," which provides a lens for assessing the justification for and conduct of warfare (Hidalgo, 2025). Perhaps at its core, the concept aims to reduce the likelihood of escalation, damage, or incidental harm that could result from war, and provide an avenue to return to a peaceful state, rather than justifying any war as "just" (Hidalgo, 2025; Smith, 2025a).

Jus ad bellum governs the conditions under which states may resort to force, focusing on self-defence, proportionality and the authority to use force under the UN Charter. *Jus in bello* (international humanitarian law) regulates how force must be exercised once conflict has begun, including the principles of distinction, proportionality and precaution.

However, AI challenges both bodies of law. Algorithmic escalation can blur thresholds for armed attack; cyber operations may fall below traditional conceptions of "force"; and cognitive warfare can destabilise societies without firing a shot. In this context, some scholars and practitioners have proposed *jus ante bellum*: a set of norms and obligations that focus on prevention, risk reduction, and pre-emptive transparency to avoid conflict spirals in the first place (Nishimoto, 2025). This doctrine resonates with the concept of jus pacis presented in this paper.

2.2. Digital Jus Pacis

The proposed concept of *digital jus pacis* seeks to extend existing doctrines into the digital domain. It is not simply an AI-specific law of armed conflict, nor merely another set of ethical guidelines. Digital *jus pacis* is proposed not as a replacement for existing international legal regimes, nor as an immediately binding body of law, but as a normative-legal framework. Its introduction seeks to embed peace-preserving principles into the design, deployment and governance of AI systems that affect international security by providing a framework that guides state behaviour, informs treaty development, and shapes the interpretation of existing obligations in contexts where AI-enabled capabilities risk destabilising peace before the threshold of armed conflict is reached.

Three pillars, which will be further explained later in this paper, underpin this framework:

1. Lawful design
2. Cooperative governance
3. Moral restraint

In this way, *digital jus pacis* treats AI as a domain in which legal, technical and normative architectures must converge to prevent algorithmic escalation and preserve human dignity.

3. The Geopolitical and Strategic Context of AI

A consensus exists that the US and China are in direct competition regarding progress in several areas, including advanced technologies, particularly AI. The vast investments by these two countries in both the commercial and public sectors, the increasing sensitivities regarding commercial espionage, and the emphasis on AI supply chains (particularly regarding rare earth minerals essential to the computing power required by AI tools), illustrate the high value placed on AI. However, other influential states are actively investing to ensure they remain at the cutting edge of new advancements in AI. Their collective actions, including he normative guardrails (or absence thereof) that they institute in particular, will likely prove highly consequential in determining the growth of the sector. As early as 2018, France, for example, commenced an investment scheme in which the government allocated $1.85 billion (USD) to AI technology, with similar efforts underway in the United Kingdom and Australia (Barsade & Horowitz, 2018). Other countries, including Türkiye, Norway, and Singapore, are investing substantially in their own AI infrastructure. These countries are attempting to utilise their existing relative advantages, such as strategic geographic location, foreign direct investment, and highly skilled labour forces, to leverage this new technology (Erai Türkiye, 2026; Barsade & Horowitz, 2018).

Similarly, the Israeli military, for example, has been actively utilising AI since 2014 in real-life applications, particularly in its 2014 conflict with Hamas, and in its more recent military activities in the Gaza Strip. The use of AI to analyse vast amounts of video footage and live feeds from hotspots around the Israeli border, throughout occupied territory and active battlefield operations, provides their security apparatus with the capacity to identify patterns, engage in predictive analytics, and take operational and tactical action accordingly (Lappin, 2017). The Israeli Defence Force (IDF) has signalled its intent to utilise AI tools not only for operational decision-making but also for strategic-level decision-making. This action will likely further enhance its military capacity by automating much of the time-consuming data collection and information sorting critical to high-level decision-making.

The application of AI to diverse sectors within the public and private domains will likely shape the geopolitical landscape of the 21st century. At the core of AI's hardware ecosystem are rare-earth elements (REEs). These elements are indispensable to key components and inputs which create the infrastructure upon which AI tools operate (Roy, 2025). Notably, these REEs are primarily extracted and refined by China, placing the country in a highly valuable position, at least in the short term, to determine the direction in which these resources are directed. Roy (2025) estimates that Beijing maintains 60% of control over extraction and 85% of refining capacity, notwithstanding the pre-eminence of private sector Taiwanese and US companies in creating the most advanced chips in the industry. The potential for geopolitical conflicts centred on access to these critical minerals may therefore become an increasingly significant feature of the 21st century's strategic landscape, a point already signalled by China's increasingly strategic posture toward Taiwan, which raise questions as to whether Beijing will eventually seek formal annexation or continue its consolidation of influence through diplomatic manoeuvring, economic leverage, and the mobilisation of international support.

For smaller states and middle powers, digital *jus pacis* offers more than an ethical framework. It provides a strategic instrument and foundational way of thinking that can shaping norms and future discussions in a domain traditionally dominated by technologically advanced actors. By embedding peace-preserving principles into AI

governance early, these states can exercise normative influence disproportionate to their material capabilities, reduce strategic vulnerability, and mitigate the risks of being norm-takers in algorithmic security environments. A concept well aligned with the global governance framework advanced by the established of the United Nations and actioned in initiative such as the Global Digital Compact and the Open-ended Working Group on security of and in the use of information and communications technologies.

4. Existing Legal and Governance Frameworks

4.1. International Humanitarian Law (IHL) and State Positions on Autonomous Weapons Systems (AWS)

International Humanitarian Law (IHL) applies to all means and methods of warfare, including those that rely on new technologies such as AI. The core principles of distinction, proportionality and precaution remain central, and states remain legally accountable for violations, regardless of the tools used (Viveros Alvarez, 2024). However, the integration of AI into targeting and decision-making processes creates practical challenges for compliance. Questions arise as to whether complex machine-learning models can reliably implement distinction; how proportionality assessments should be conducted when harm estimates are algorithmically generated; and how responsibility should be assigned when harm results from interactions between autonomous systems and human operators.

Several states and coalitions have issued political declarations on responsible military use of AI, setting out non-binding principles such as meaningful human control, reliability, and human accountability. These include national policy statements, regional initiatives and joint declarations in multilateral fora (Tréhu & Ricart, 2024; United Nations Secretary-General's High-Level Advisory Body on AI, 2024). While valuable, these efforts remain fragmented and lack the force of binding law.

4.2. Regional and Global AI Governance Instruments

A mix of binding regulations, soft-law instruments, and multistakeholder initiatives characterises the emerging global AI governance regime. Key examples include:

- European Union AI Act
- Council of Europe Framework Convention on AI
- UNESCO Recommendation on the Ethics of AI
- UN High-Level Advisory Body on AI and Global Digital Compact

These instruments demonstrate growing recognition that AI governance must be both principled and practically implementable. Yet none of them is specifically designed to address the full spectrum of security-related risks posed by AI or to serve as a comprehensive peace-oriented framework for algorithmic warfare and cognitive manipulation.

5. Pillars of Digital Jus Pacis

5.1. Pillar 1: Lawful Design – Human Oversight and Proportionality

The first pillar of *digital jus pacis* is lawful design: embedding legal and ethical constraints directly into AI systems used in defence and security. This goes beyond ex post review and calls for ex ante alignment of technical architectures with international law and human rights standards (Viveros Alvarez, 2024). An approach, which reframes the regulation of military and security-related AI from a predominantly logistical or international relations challenge into a question of global governance and acceptable state behaviour in the development and use of AI-enabled capabilities.

Key components include:

- **Meaningful human control.** AI systems involved in the use of force must be designed so that human operators can understand, scrutinise and override machine outputs. This entails user-centred interface design, decision-support that explains underlying reasoning, and operational doctrines that preserve human authority over critical choices (NIST, 2023). With a leaning towards Human-in-the-Loop (HITL) approaches compared to Human-on-the-Loop (HOTL) and Human-over-the-Loop (HOOTL) approaches. The key difference lies in the level and frequency of human involvement, ranging from constant collaboration (HITL) to distant supervision (HOOTL).

- **Proportionality and distinction by design.** Systems should incorporate constraints and safeguards that reflect legal obligations—for instance, conservative thresholds for target confirmation, multi-sensor cross-checking, and conservative defaults when uncertainty is high.
- **Explainability and transparency.** Black-box models may be incompatible with contexts where legal accountability requires traceability of decisions. Where opaque models are used, additional auditing layers and post-hoc explanations may be necessary (Smith & Crampton, 2024).
- **Fail-safe mechanisms and kill-switches.** War algorithms should include robust mechanisms for safe shutdown, especially when anomalous behaviour is detected or communication is lost.

By designing systems with legal and ethical constraints in mind, states can reduce the risk that AI deployments will inadvertently violate IHL, human rights, or emerging *jus ante bellum* norms.

5.2. Pillar 2: Cooperative Governance – Regional and Global Alignment

The second pillar emphasises cooperative governance and information exchange. Given that AI supply chains, data flows and security risks are transnational, no state can achieve *digital jus pacis* in isolation (Tréhu & Ricart, 2024; Global Partnership for Sustainable Development Data, 2025).

This pillar has four main elements:

1. **Global multilateral instruments.** A future AI Peace Compact could build on existing initiatives by codifying principles for responsible military AI, algorithmic transparency, and restrictions on AI in nuclear command and control.
2. **Regional frameworks and model laws.** Regional organisations can contextualise global principles by developing model legislation and guidelines tailored to specific legal traditions, threat environments and levels of capacity (ECLAC, 2024).
3. **Information exchange and transparency mechanisms.** Effective cooperation requires trusted channels for sharing information about AI incidents, vulnerabilities, deployment practices and best practices.
4. **Specialised institutions and monitoring.** Dedicated bodies, such as an international observatory on military AI or a multilateral AI safety council, could coordinate data collection, conduct independent assessments and support verification.

Cooperative governance thus centres on building a networked architecture of norms, institutions, and information flows that collectively foster restraint and reduce the risk of AI-driven conflict.

5.3. Pillar 3: Moral Restraint – Digital Humanitarian Ethics

The third pillar emphasises moral restraint by integrating human rights, intergenerational equity, environmental sustainability, and cultural pluralism into AI security governance.

- **Human rights integration.** Many scholars argue that human rights norms provide a robust ethical foundation for AI governance, particularly in terms of dignity, equality, and ensuring accountability (Jones, 2023). Factors that are already at the forefront of discussion on ethical design, especially in public governmental forums.
- **Intergenerational equity.** Decisions about AI systems, especially in nuclear, cyber and cognitive domains, can have long-term implications for future generations.
- **Environmental sustainability.** AI training and deployment can carry high environmental costs, including energy consumption and resource utilisation.
- **Cultural pluralism and openness.** Debates about open-source AI and data sharing must consider the risk of exacerbating digital divides and reinforcing dominant cultural narratives.

Moral restraint thus anchors *digital jus pacis* in a broader ethical horizon, recognising that peace in the AI era is not merely the absence of war but the presence of just, sustainable and inclusive digital orders.

6. Conclusion

AI has ushered in a new era of algorithmic warfare, cognitive manipulation and data-driven security practices. These developments strain existing legal frameworks and expose gaps in accountability, transparency and restraint. Nevertheless, they also offer an opportunity: to rethink how law, technology and ethics interact in the preservation of peace. D*igital jus pacis*, grounded in lawful design, cooperative governance and moral restraint,

can provide a valuable framework for addressing AI's security implications. This concept connects classical doctrines of *jus ad bellum* and *jus in bello* with emerging ideas of *jus ante bellum* and contemporary debates on AI ethics and governance.

To move from concept to practice, states and international organisations must deepen cooperation and information exchange, align regional and global instruments, support capacity-building in the Global South, and design AI systems that are legally and ethically constrained by default.

Literature

1. Barsade, I., & Horowitz, M. C. (2018, August 16). *Artificial intelligence beyond the superpowers*. *Bulletin of the Atomic Scientists*. https://thebulletin.org/2018/08/the-ai-arms-race-and-the-rest-of-the-world/
2. Bode, I., & Bhila, I. (2024, September 3). *The problem of algorithmic bias in AI-based military decision-support systems*. International Committee of the Red Cross – Law and Policy Blog. https://blogs.icrc.org/law-and-policy/2024/09/03/the-problem-of-algorithmic-bias-in-ai-based-military-decision-support-systems/
3. Dookeran, W., & Byng, M. (2025). Geopolitical realignment in the twenty-first century: A case for Trinidad and Tobago's strategic shift from non-alignment to multi-alignment. *Horizons: Journal of International Relations and Sustainable Development*, 30, 262–272. https://www.jstor.org/stable/48829698
4. Erai Turkey. (2025). *How AI in Turkey's industrial revolution is driving innovation and growth*. https://eraiturkey.com/2024/08/ai-in-turkeys-industrial-revolution/
5. Global Partnership for Sustainable Development Data. (2025). *A step in the right direction: UN establishes new mechanisms to advance global AI governance*. https://www.data4sdgs.org/news/step-right-direction-un-establishes-new-mechanisms-advance-global-ai-governance
6. Jones, K. (2023). *Human rights should be at the heart of AI and technology governance*. Carnegie Council for Ethics in International Affairs. https://www.carnegiecouncil.org/media/article/human-rights-ai-technology-governance
7. Lappin, Y. (2017). *Artificial intelligence beyond the superpowers*. *Bulletin of the Atomic Scientists*. https://thebulletin.org/2018/08/the-ai-arms-race-and-the-rest-of-the-world/
8. Nishimoto, J. (2025). Artificial intelligence and nuclear weapons: A critical assessment of risks and benefits. *Texas National Security Review*, 8(3). https://tnsr.org/2025/06/artificial-intelligence-and-nuclear-weapons-a-commonsense-approach-to-understanding-costs-and-benefits/
9. Ratska, M., & Bitzinger, R. A. (2023). *The AI wave in defence innovation: Assessing military artificial intelligence strategies*. Routledge.
10. Smith, B., & Crampton, N. (2024). *Global governance: Goals and lessons for AI*. Microsoft On the Issues Blog. https://blogs.microsoft.com/on-the-issues/2024/09/23/global-governance-goals-and-lessons-for-ai/
11. Smith, T. (2025a). *Cyber crisis management in the AI era: Confronting disinformation and hybrid threats* (EU CyberNet Expert Series No. 5). https://www.eucybernet.eu/wp-content/uploads/2025/10/eu-cybernet-expert-blog-series-smith-no5-2025.pdf
12. Smith, T. (2025b). Just war theory in the cyber age: Ethical implications for modern-day security. *SPOTLIGHT on Crime and Public Safety*, 5(2), 4.
13. Tréhu, J., Ricart, R. J., & German Marshall Fund of the United States. (2024). *Global AI governance: Key steps for transatlantic cooperation*. https://www.gmfus.org/news/global-ai-governance-key-steps-transatlantic-cooperation
14. United Nations. (2024). *Global Digital Compact*. https://www.un.org/techenvoy/global-digital-compact
15. United Nations High-Level Advisory Body on AI. (2024). *Governing AI for humanity: Final report*. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_es.pdf
16. Viveros Alvarez, J. (2024, September 4). *The risks and inefficiencies of AI-based military targeting*. International Committee of the Red Cross – Law and Policy Blog. https://blogs.icrc.org/law-and-policy/2024/09/04/risks-inefficiencies-ai-military-targeting

Original Scientific Paper/Original naučni rad UDC/UDK: 327:378]:005.7
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601107M

Primena poslovne arhitekture za složenost upravljanja u transnacionalnim univerzitetskim savezima

Senne De Moor[1], Renata Petrevska Nechkoska[2]

[1]Faculty of economics and business administration, Ghent University, Belgium, senne.de.moor@9altitudes.com
[2] Faculty of economics and business administration, Ghent University Belgium and University St. Kliment Ohridski, Bitola, North Macedonia, renata.petrevskanechkoska@ugent.be

Summary in Serbian: Izazovi u nedavno prisutnim univerzitetskim savezima koji rade na ostvarivanju evropskog stepena odnose se na dvostruku logiku: projektno zasnovan način rada vođen kratkoročnim prekretnicama i rezultatima, i ambicioznu viziju postajanja dugoročnog, misijom vođenog obrazovnog ekosistema. Ovo stvara trenje između administrativne koordinacije i strateške transformacije. Prema evaluaciji Komiteta CULT o Inicijativi evropskih univerziteta (EUI), savezi često ostaju zarobljeni u „zamci projekata“, fokusirajući se na ispunjavanje ugovornih rezultata, a ne na dizajniranje struktura upravljanja koje traju i nakon finansiranja. Da bi prevazišli ovaj kratkoročni cilj, univerzitetskim savezima su potrebni modeli upravljanja koji mogu podržati dugoročnu institucionalnu transformaciju dok upravljaju neposrednim operativnim zahtevima. Oni moraju da se kreću kroz fragmentirane pravne, administrativne i kulturne pejzaže dok teže ambicioznim ciljevima interoperabilnosti, ko-kreacije i inovacija. Da bi podržali takvu institucionalnu transformaciju, neki autori predlažu Arhitekturni okvir za sisteme visokog obrazovanja (HES) zasnovan na pogledu. Ovaj pristup prilagođava principe arhitekture preduzeća (EA) - posebno TOGAF standard - akademskim ekosistemima i nudi strukturirane smernice za usklađivanje upravljanja, strategije i komunikacije preko granica. Sistemi visokog obrazovanja funkcionišu kao složena socio-tehnička okruženja sastavljena od različitih aktera - studenata, fakulteta, administrativnog osoblja, kreatora politike - koji deluju u različitim nacionalnim kontekstima. Naše istraživanje istražuje literaturu, kao i nekoliko evropskih univerzitetskih saveza i jedan detaljno, i pokušava da argumentuje obrazloženje za poslovnu arhitekturu u visokom obrazovanju. Naš rad je usmeren ka referentnoj arhitekturi ili „meta-modelu“ koji bi mogao da vodi dizajn saveza i efikasno upravljanje širom Evrope.

Keywords : Arhitektura preduzeća, Evropski univerzitetski savezi, upravljanje, otkrivanje, interoperabilnost

Applying Enterprise Architecture for governance complexities in transnational university alliances

Abstract in English: The challenges in the lately present university alliances working towards European Degree are around the dual logic: a project-based mode driven by short-term milestones and deliverables, and an aspirational vision of becoming a long-term, mission-driven educational ecosystem. This creates friction between administrative coordination and strategic transformation. According to the CULT Committee’s evaluation of the European Universities Initiative (EUI), alliances often remain locked in a “project trap,” focusing on meeting contractual outputs rather than designing governance structures that endure beyond the funding. To move beyond this short-termism, university alliances require governance models that can support long-term institutional transformation while managing immediate operational demands. Thet must navigate fragmented legal, administrative, and cultural landscapes while pursuing ambitious goals of interoperability, co-creation, and innovation. To support such institutional transformation, some authors propose a view-based Architecture Framework for Higher Education Systems (HES). This approach adapts enterprise architecture (EA) principles - particularly the TOGAF standard - to academic ecosystems and offers structured guidance for aligning governance, strategy, and communication across borders. Higher Education Systems function as complex socio-technical environments composed of diverse actors - students, faculty, administrative staff, policymakers - operating under different national contexts. Our research explores literature, as well as several European university alliances and one in depth, and attempts to argument the rationale for Enterprise Architecture in Higher Education. Our work is towards a reference architecture or “meta-model” that could guide alliance design and effective governance across Europe.

Keywords: Enterprise Architecture, European University Alliances, governance, discovery, interoperability

1. Introduction

European university alliances, as the relatively new universe of pan-European co-creation in the higher education, operate within a dual logic: a project-based mode driven by short-term milestones and deliverables, and an aspirational vision of becoming a long-term, mission-driven educational ecosystem. This creates friction between administrative coordination and strategic transformation. According to the CULT Committee's evaluation of the European Universities Initiative (EUI), alliances often remain locked in a "project trap," focusing on meeting contractual outputs rather than designing governance structures that endure beyond the funding cycle (Craciun et al., 2023) To move beyond this short-termism, university alliances require governance models that can support long-term institutional transformation while managing immediate operational demands. One promising approach lies in the application of enterprise architecture principles tailored to the higher education context. Transnational university alliances must navigate fragmented legal, administrative, and cultural landscapes while pursuing ambitious goals of interoperability, co-creation, and innovation. To support such institutional transformation, (Rouvrais & Petersen, 2024) propose a view-based Architecture Framework for Higher Education Systems (HES). This model adapts enterprise architecture (EA) principles - particularly the TOGAF standard - to academic ecosystems and offers structured guidance for aligning governance, strategy, and communication across borders. The accelerating pace of global interdependence shaped by challenges such as climate change, technological disruption, migration, and the reconfiguration of international cooperation has fueled the rise of cross-border partnerships across sectors. These challenges increasingly transcend the capacities of individual nations, institutions, or sectors to address alone. As a result, transnational collaboration has emerged not only as a pragmatic response but as governance innovation in itself.

At the heart of these arrangements lies the concept of the ecosystem: a loosely coupled, interdependent network of actors, resources, and institutions that co-evolve around a shared value proposition. First introduced in a business context by James F. Moore, ecosystems are defined as "an economic community supported by a foundation of interacting organizations and individuals - the organisms of the business world who co-evolve their capabilities and roles and tend to align themselves with the directions set by one or more central companies" (Moore, 1993). In more recent scholarship, (Adner, 2017) reframes ecosystems as structures of interdependent actors whose alignment is necessary for the realization of value, emphasizing that it is not the actors themselves, but the interdependencies among them that define the ecosystem (Adner, 2017). This notion has since expanded beyond business into innovation systems (Jackson, 2011), environmental governance (Ostrom, 2015), and higher (Benneworth et al., 2017). Common across these domains is the shift from control to coordination, from hierarchy to co-creation, and from formal integration to adaptive alignment (Emerson et al., 2012; Nechkoska, 2019). Ecosystems emphasize distributed agency, where multiple stakeholders interact continuously in complex environments, co-producing value and collectively responding to uncertainty. One initial framework that helps unpack how ecosystems function transnationally is Collaborative Environmental Governance (CEG). As Ulibarri et al. emphasize, collaboration is particularly crucial in environmental contexts where problems are dynamic, data is uncertain, and resources are fragmented (Ulibarri et al., 2023. In such settings, transnational collaboration becomes both a necessity and an innovation, fostering ecosystems that span political, ecological, and institutional boundaries. The CEG literature offers useful insights into how such collaborations are initiated, how power and trust evolve, what governance structures support long-term functionality.

The European Universities Initiative (EUI) represents a particularly novel and ambitious application of ecosystem logic within the field of higher education. Launched in response to President Macron's Sorbonne speech in 2017 and supported through Erasmus+ and Horizon Europe funding, the EUI seeks to transform European higher education by building long-term, integrated alliances between universities across Europe. These alliances - such as EU-CONEXUS, Una Europa, and 4EU+ - are expected to function as transnational ecosystems for education, research, and innovation. Yet, as in environmental governance, establishing such ecosystems within higher education presents unique challenges: national legal frameworks differ significantly, institutional autonomy varies across systems, cultural norms, decision-making styles and languages introduce friction and coordination is often stretched across a vast web of actors with uneven power and priorities. These dynamics closely mirror what Ulibarri et al. observe in their synthesis of global collaborative governance regimes: that no universal blueprint exists for building successful transnational collaborations (Ulibarri et al., 2023). Each context requires customized practices, adaptive leadership, and tailored facilitation. "You can't apply the same methods or tools in every setting," they note, "you have to understand the local political, social, and environmental realities." For European university alliances, this means moving beyond technocratic management to cultivate co-creation, trust, and continuous sense-making among diverse stakeholders.

Our methodology deploys systematic literature review, analysis of different European university alliances on different criteria relevant for this research, and use of a case study of one European university alliance COLOURS, using Social Network Analysis, to map and model the governance contexts alliances operate in. Across the entire workflow, we used the principles of Enterprise Architecture standard – TOGAF to address the potentials of EA to transnational governance complexities in Higher Education ecosystems.

2. Literature review

This literature review builds on these insights by exploring what makes transnational ecosystems function effectively, especially when they span national, cultural, legal, and institutional boundaries. It positions European University Alliances as a compelling site of inquiry - not just as policy experiments, but as complex adaptive systems that require careful management of the human and organizational interfaces that sustain them. Specifically, it investigates how challenges related to communication infrastructure shape the collaborative capacity of these alliances.

By drawing lessons from **cross-sectoral transnational collaborations** in environmental and governance contexts, we transition to the specific characteristics of European University Alliances, before unpacking the role of communication practices and facilitation in enabling or impeding their evolution. Ultimately, the goal is to bridge abstract ecosystem theory with the practical realities of managing people, relationships, and shared meaning in one of Europe's most ambitious cross-border higher education experiments, with focus on governance complexities.

While the majority of literature on Collaborative Environmental Governance (CEG) draws from European or North American contexts, valuable insights can also be gained from large-scale, complex governance models within single political systems. The case of the Yangtze River Basin in China offers a particularly instructive example of how collaborative environmental governance functions across diverse jurisdictions, each with different levels of capacity, motivation, and autonomy. Although intra-national, the cross-provincial dynamics of the Yangtze governance system closely mirror many of the institutional and relational complexities found in transnational ecosystem collaborations. As foundational structures for collaboration Xia et al. identify a framework of internal and external factors that jointly shape the effectiveness of collaborative environmental governance (Xia et al., 2024). This framework reflects the necessity of balancing formal structures with soft relational conditions - an equilibrium similarly pursued in the design of European university alliances and cross-border environmental networks. The factors can be divided into: External (structural) factors: Legal, institutional, technical factors and internal (relational) factors: perceptual, relational, interactivity, efficacy Together, these dimensions highlight the dual importance of **hard governance architecture and soft system qualities in managing complex ecological networks.** Notably, these findings closely mirror those in EU-based ecosystems, where success is often contingent on policy frameworks and cultural/institutional alignment.

For transnational ecosystems, where legal harmonization is often slow and cultural diversity high, the insight that internal relational factors - such as trust, communication quality, mutual understanding, and belief in the efficacy of joint action - are not merely supportive of governance effectiveness but constitute its core engine, is pivotal. Governance structures must invest in trust-building, sense-making, and continuous dialogue, not just formal agreements. The internal **"infrastructure of trust"** becomes the invisible institution upon which visible structures depend. This study provides three effective collaboration models (pathways) using fuzzy-set Qualitative Comparative Analysis (fsQCA), they identified three collaborative "success pathways". These provide a nuanced view of how different configurations of factors can enable effective governance. Each model offers a transferable governance archetype relevant to transnational collaborations: Technology Empowers Relationship Driving, Institution Reinforces Interactive Driving And Internal-External Interactive Driving.

This aligns closely with Emerson and Nabatchi's (2015) view of Collaborative Governance Regimes (CGRs) as adaptive systems, shaped by changing inputs, stakeholder dynamics, and feedback over time. In their "Integrative Framework," they emphasize that no static structure is sufficient; governance must emerge through continuous alignment of shared purpose, institutional arrangements, and collaborative dynamics (Emerson & Nabatchi, 2015). Similarly, Ulibarri et al. (2023) reinforce the idea that governance must be tailored to the specific social, political, and ecological setting, supported by case-based evidence from the Collaborative Governance Case Database (CGCD) (Ulibarri et al., 2023). These insights have significant implications for the governance of transnational ecosystems where conditions differ widely across jurisdictions. Attempting to impose uniform standards or governance models can backfire if they do not match the readiness levels, incentives, and capacities of local actors. Instead, governance models must be modular, reconfigurable, and sensitive to emerging conditions - a principle

echoed in tactical management approaches like the Denica method (Petrevska Nechkoska, 2019). In practical terms, this suggests that governance design should:

- Allow for multiple entry points (e.g., legal, technical, cultural).
- Prioritize facilitative leadership that adapts strategies to shifting dynamics.
- Create mechanisms for continuous reflection and adjustment, such as feedback loops, learning cycles, and real-time dialogue.

Thus, **ecosystem governance** becomes less about enforcing uniformity and more about cultivating coherence in diversity - aligning diverse actors around shared purpose through adaptable structures and participatory processes. Effective collaboration in complex ecosystems relies not only on legal structures or formal agreements, but equally on relational factors such as trust, shared urgency, and the capacity for joint decision-making. These lessons highlight the importance of governance models that accommodate institutional diversity and functional complementarity, especially in transnational contexts like higher education. This is where the concept of hybrid governance becomes especially relevant.

Hybrid governance refers to arrangements that combine centralized policy support with decentralized execution. This model provides a flexible framework for organizing collaboration where formal authority is fragmented or limited, and where actors differ in legal status, cultural background, or strategic priorities. Rather than enforcing uniform solutions, hybrid systems rely on a blend of formal rules, informal norms, voluntary codes, partnerships, and shared platforms to guide collective action. As Gunningham (2016) notes, this shift reflects a broader transformation in governance theory - from hierarchical, top-down regulation to pluralistic, networked architectures capable of navigating complexity. Hybrid governance thus becomes a crucial enabler in contexts like transnational university alliances, where institutional misalignment, legal fragmentation, and cultural diversity render uniform governance approaches ineffective (Gunningham & Holley, 2016). Crucially, this model reconceptualizes the role of central institutions. Rather than functioning as top-down regulators, supranational bodies such as the European Commission or intergovernmental environmental regimes act as platform providers. They enable localized experimentation through funding, convening power, and legitimacy, while allowing bottom-up adaptation and co-creation. This form of facilitative centralism supports resilient and adaptive ecosystems that balance coherence with autonomy. However, the effectiveness of such systems depends on several enabling conditions:

- Trust among actors, especially in the absence of strong enforcement mechanisms
- Accountability frameworks that ensure transparency without stifling flexibility
- Intermediary organizations to bridge interests and sustain collaboration
- Shared definitions of problems and outcomes to enable alignment and learning

These findings reinforce central claims in the collaborative governance literature (Emerson & Nabatchi, 2015; Ulibarri et al., 2023) and empirical cases such as the Yangtze River Basin and EU-CONEXUS alliance. Across these contexts, the message is consistent: effective transnational ecosystems emerge not from rigid uniformity but from adaptive structures that combine centralized support with decentralized initiative, institutional scaffolding with relational trust, and hybrid mechanisms with shared purpose.

While hybrid governance provides the conceptual foundation, its effectiveness hinges on how it is operationalized through global governance mechanisms. **Global governance** refers to the collective efforts of international institutions, states, civil society actors, and other stakeholders to address cross-border issues. It encompasses processes such as agenda setting, policy implementation, and monitoring and enforcement - all of which are vital to structuring transnational collaboration. Important aspects to be incorporated to achieve hybrid governance are:

- Agenda Setting: Establishing Global Environmental and Institutional Priorities
- Policy Implementation: Coordinating Actions Across Jurisdictions
- Monitoring and Enforcement: Balancing Flexibility and Accountability

Together, the perspectives of hybrid and global governance underscore the interdependence between conceptual adaptability and operational structure. Transnational ecosystems require governance models that embrace diversity, decentralization, and experimentation, while also embedding clear mechanisms for coordination, monitoring, and accountability. Hybrid governance offers the conceptual map; global governance provides the institutional tools to navigate it. By linking the relational foundations of collaboration with structured enforcement, these combined models illuminate a path toward more resilient, inclusive, and functional transnational ecosystems.

3. Enterprise Architecture approach to transnational governance

Higher Education Systems function as complex socio-technical environments composed of diverse actors - students, faculty, administrative staff, policymakers - operating under different national contexts. (Rouvrais & Petersen, 2024) argue that a tailored Enterprise Architecture (EA) approach enables universities to manage this complexity by improving sense-making, systemic design, and collaborative alignment.

Their Architecture Development Method (ADM) outlines interconnected views for managing transformation:

A. Vision & Principles: Establishes shared strategic goals and values.
B. Business Architecture: Clarifies institutional roles, processes, and stakeholder responsibilities.
C. Information Systems Architectures: Help choose integration or new approaches or combinations
D. Technology Architecture: Extremely challenged due to the alliances' built by different universities from different regions with all different contextual factors
E. Opportunities & Solutions: Identifies shared goals and collaborative initiatives.
F. Migration Planning: Provides phased, structured implementation pathways.
G. Implementation & Governance: Establishes operational governance mechanisms and quality assurance.
H. Change Management: Supports institutional learning and adaptation.

These views are not linear steps but iterative processes - a reflection of the recursive, emergent properties common in Complex Adaptive Systems (CAS), as also described by (Petrevska Nechkoska et al., 2023). In the further discussions, we will be using these views for all the modeling discussions and choices.

4. The necessary strategic alignment in university alliances – analysis of alliances

Applying the ADM helps alliances overcome strategic challenges like governance complexity and legal disparities. For instance, EU-CONEXUS, facing national legal discrepancies complicating joint program delivery, leveraged Business Architecture (View B) by clearly defining institutional responsibilities in a cross-institutional agreement, facilitating smoother joint degree implementation. Additionally, CHARM-EU successfully implemented the Migration Planning (View F) approach, deploying a structured roadmap to incrementally integrate hybrid classrooms and align educational technology standards across its partners. This phased implementation approach improved stakeholder buy-in, resource allocation clarity, and reduced integration risks.

4.1. Towards interoperability of governance and communication

Effective governance is essential yet challenging for transnational alliances. According to the LERU report (Lievens, 2024), alliances struggle to establish stable, transparent governance. Addressing this, Implementation & Governance (View G) was utilized by Una Europa, which formalized governance structures and accountability mechanisms, improving operational transparency and strategic coherence across diverse institutional cultures. Furthermore, the Opportunities & Solutions (View E) played a pivotal role in helping CIVICA - the European University of Social Sciences - structure effective communication channels across multiple institutional levels. Recognizing the inherent complexity of transnational collaboration, CIVICA established a multi-tiered communication framework designed to bridge gaps between academic, administrative, and technical staff, ensuring that strategic objectives were translated into operational practices consistently across partners (*About CIVICA | Civica*, n.d.).

At the macro level, CIVICA instituted Alliance Coordination Committees that included representatives from each member university's leadership, policy, and governance teams. These committees were tasked with aligning strategic goals, monitoring progress on joint initiatives, and serving as a forum for resolving institutional discrepancies. Simultaneously, at the micro level, specialized working groups and task forces were formed, focusing on specific operational areas such as curriculum development, digital infrastructure, and mobility schemes. These groups maintained direct lines of communication with their counterparts across partner institutions, fostering peer-to-peer knowledge exchange and agile problem-solving. To support these interactions, CIVICA leveraged collaborative digital platforms - including shared project management tools, regular virtual meetings, and a centralized knowledge repository - ensuring that all stakeholders, regardless of institutional affiliation or role, had access to up-to-date information and decision-making processes.

This structured yet flexible communication architecture proved critical during the joint development of micro-credentials, a process that demanded both academic rigor and administrative coordination. Given the diversity of

regulatory environments, institutional cultures, and stakeholder expectations, transparent communication ensured that course design, quality assurance, and credential recognition processes were co-developed in an inclusive and coherent manner. Moreover, this approach aligns closely with the Una Europa Diversity Council's recommendations on intersectional and inclusive communication within transnational educational networks. By embedding principles of transparency, accessibility, and participatory engagement into its communication strategy, CIVICA not only enhanced operational efficiency but also fostered a sense of shared ownership among its member institutions. In essence, Opportunities & Solutions (View E) functioned as an enabling layer of governance, translating high-level strategic ambitions into coordinated, inclusive, and actionable collaboration practices - crucial for navigating the multi-institutional complexity of joint educational innovation.

4.2. Toward a Meta-Model for European Alliances

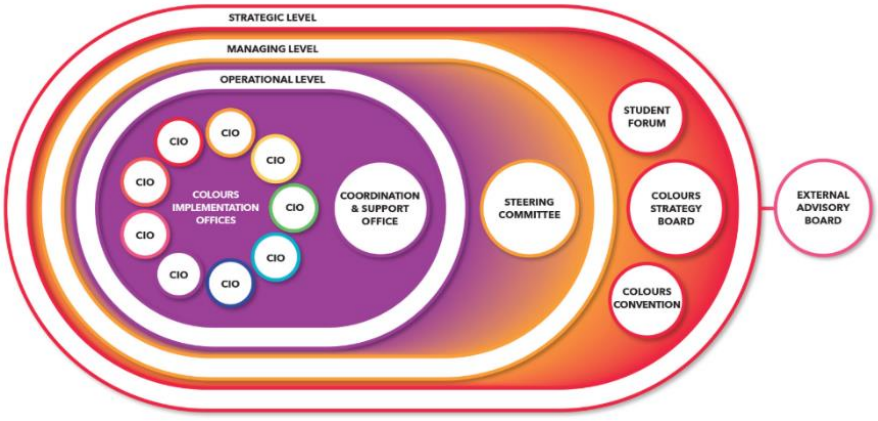
The combined views in this framework offer more than just a descriptive tool - they lay the groundwork for a reference architecture or "meta-model" that could guide alliance design across Europe. Such a model would support modular design of curricula, joint infrastructures, and strategic coordination across national boundaries. This vision is echoed in EU policy discourse: both the CULT Committee's 2023 study on the European Universities Initiative and the LERU 2024 paper call for more coherence, resilience, and legal clarity in alliance development.

By applying the Rouvrais & Petersen framework (TOGAF), alliances can identify structural bottlenecks, align their governance with institutional missions, and prototype new forms of academic cooperation. As alliances move toward long-term institutionalization, EA-based models provide the foundation for shared strategic foresight and agile transformation. While frameworks like the one proposed by Rouvrais and Petersen offer valuable tools for aligning strategy, structure, and transformation in university alliances, they cannot fully resolve the lived complexities of governance on their own. Empirical evidence from existing alliances highlights persistent tensions - between inclusivity and efficiency, standardization and autonomy, vision and viability - that architectural models alone cannot neutralize. This section complements the architectural perspective by synthesizing insights from case studies, policy evaluations, and operational experiences across European University Alliances (EUAs). It foregrounds unresolved governance challenges and offers an evidence-based overview of evolving models, practical tensions, and strategic recommendations.

4.3. Case study of a European University Alliance: COLOURS

The COLOURS European University Alliance has adopted a multi-level governance structure designed to ensure inclusivity, agility, and representativeness across all stakeholders. This architecture reflects the alliance's values of co-creation, transparency, and democratic participation, aligning closely with the principles of the European Universities Initiative. Governance is organized into three interrelated levels - Strategic, Managing/Coordination, and Operational - each with distinct roles and bodies that work in synergy to steer the alliance's development and implementation.

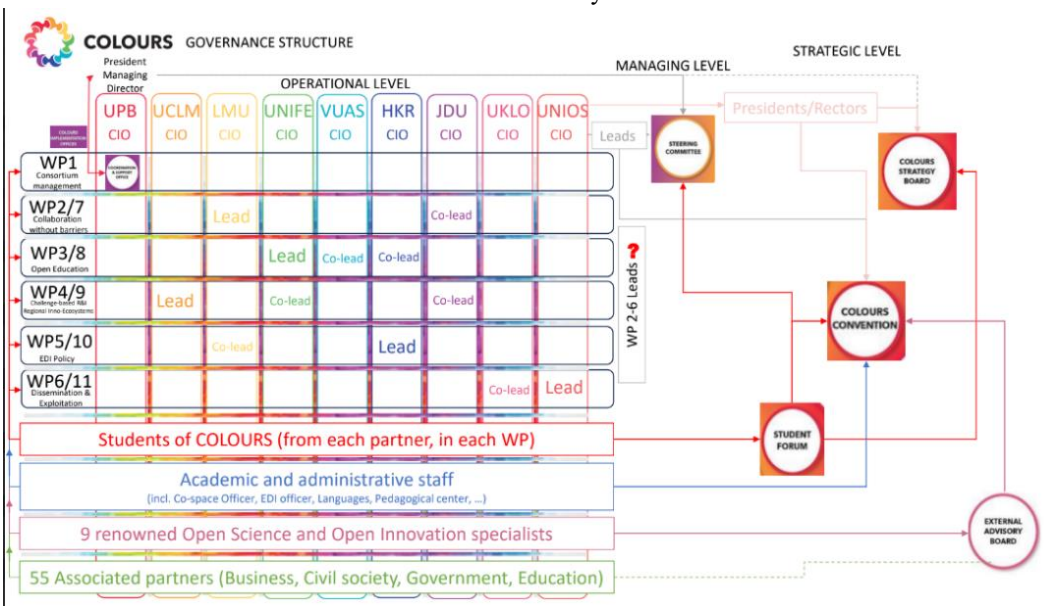
Figure 1. High level governance structure of the COLOURS European university alliance (strategic, tactical, operational and external level)



The alliance is organized into eleven major work packages (WPs), each addressing specific thematic or functional areas of collaboration: WP1: Consortium Management and Decision Making – Development, Testing,

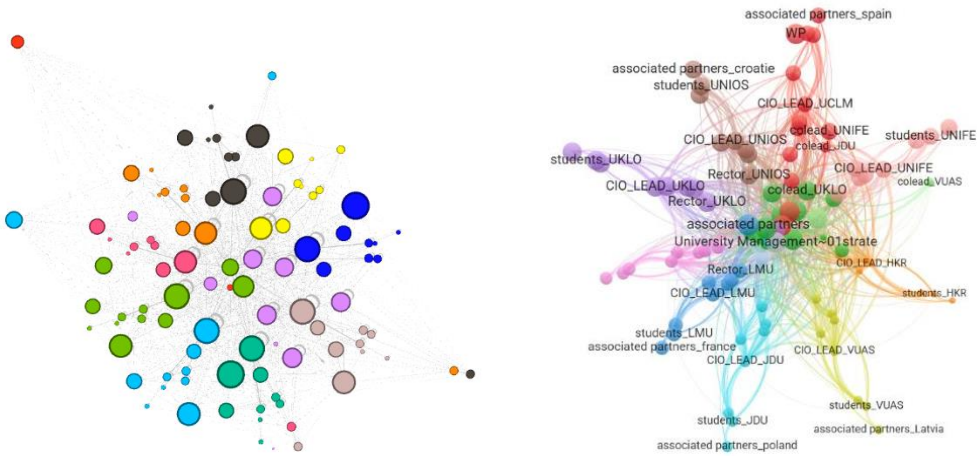
Implementation and Long-Term Sustainability, WP2/WP7: Collaboration without Barriers, WP3/WP8: Open Education (OE), WP4/WP9: Challenge-based Research & Innovation in Regional Innovation Ecosystems, WP5/WP10: Equality, Diversity, and Inclusion (EDI) Policy, WP6/WP11: Dissemination and Exploitation. Notably, each work package is led or co-led by different universities, ensuring distributed ownership and shared responsibility across the alliance. This sets the scene for multi-layered, highly complex governance ecosystem.

Figure 2. Governance structure of the COLOURS alliance across the work packages and transnational stakeholders ecosystem



The visual distinction of institutional boundaries depicted through the use of 9 different colors for each partner university within the network, providing a meaningful frame of reference for interpreting the results of cluster analysis and centrality metrics. It is particularly useful when assessing whether communication patterns are institution-bound or extend across organizational borders.

Figure 3. Governance structure of the main roles in COLOURS alliance (nodes), and their links



This visualization is particularly insightful and was generated using the Force Atlas layout in Gephi. This layout is especially suited for social network analysis, as it simulates physical forces: nodes repel each other while edges act like springs, pulling connected nodes together. The Force Atlas layout was run with inertia (0.1), repulsion strength (200), attraction strength (25), maximum displacement (10), gravity (10), speed (3.0), auto stabilization enabled with strength (80) and sensitivity (0.2), attraction distribution enabled, and without adjusting for node

size. As a result, highly connected nodes are drawn into clusters, while less connected or peripheral actors are pushed outward.

Table 3. Network metrics translated into strategic governance insights

Metric	**Value**	**Interpretation**
Average Degree	20.745	On average, each actor is connected to ~21 others. This suggests a fairly active communication network.
Avg. Weighted Degree	343.164	The average strength/intensity of connections is high. This could indicate frequent collaboration or communication. (Opsahl et al., 2010a)
Network Diameter	3	The maximum distance between any two nodes is 3. This is very low and shows a highly cohesive alliance. No one is more than 3 steps removed from anyone else. (S. Borgatti & Halgin, 2011)
Density	0.19	About 19% of all possible connections exist. This is moderate, suggesting good interconnectivity, though not saturated - still room to grow in cross-institutional links. (Provan & Kenis, 2008)

A network diameter of 3 indicates that the maximum number of steps required to connect any two nodes within the alliance is remarkably low. This suggests a high level of structural cohesion and communication efficiency across the network. In social network theory, a small diameter is associated with faster information dissemination, improved coordination, and a greater capacity for collective responsiveness (S. Borgatti & Halgin, 2011). Particularly in collaborative settings such as transnational university alliances, where cross-border complexity can hinder information flow, a compact diameter supports agile decision-making and reinforces the potential for networked governance (Provan & Kenis, 2008). From a design perspective, this reflects a well-integrated network structure, enabling actors in the alliance to access each other with minimal intermediaries - an asset for both formal coordination and informal knowledge exchange.

The network density of 0.19 indicates that approximately 19% of all possible connections within the alliance are currently active. While this reflects a baseline level of collaboration, it also reveals substantial untapped potential for strengthening interconnections, particularly between different institutions or work packages that may currently operate in relative isolation (specific nodes will be discussed). In the context of inter-organizational networks, moderate density is not inherently negative - it often signals a balance between efficiency and flexibility (Provan & Kenis, 2008). However, in alliances aiming for deep integration and joint strategic alignment, higher density can facilitate shared understanding, trust-building, and responsiveness across units. Thus, a density of 0.19 may highlight opportunities to intentionally foster cross-node collaboration, bridge communication gaps, and reinforce alliance-wide cohesion through structured interactions or co-creation initiatives.

The high average weighted degree of 343.16 suggests that, beyond simply being connected, many actors in the alliance engage in frequent or intense communication exchanges. This implies the presence of a core group of individuals or units functioning as operational hubs, characterized by strong, recurring interactions. When considered alongside the low network diameter (3) and moderate density (0.19), this pattern indicates a network in which information can circulate efficiently through a few key nodes, but where collaboration may still be concentrated rather than evenly distributed. Such concentration can be both a strength and a vulnerability. On one hand, it facilitates swift coordination through central actors; on the other, it creates a risk of communication bottlenecks or overload, particularly if those central actors are not supported by a distributed structure. These high-weight actors - especially if they also exhibit high betweenness or closeness centrality - should be flagged as potential core operators and may warrant targeted support, delegation mechanisms, or buffer roles to ensure resilience and sustainability within the alliance's communication ecosystem.

This interpretation is further supported by insights from the conducted interviews, which revealed a persistent reliance on a fragmented set of communication tools across the alliance. Platforms such as GoFAST, Google Drive, traditional email, and even informal channels like WhatsApp are concurrently in use. While each of these tools may serve distinct functional purposes - such as document collaboration, synchronous updates, or informal check-ins - the absence of a unified digital infrastructure exacerbates the risks associated with centralization. Specifically, when communication flows are already dependent on a limited set of highly active nodes, the lack of platform standardization introduces additional friction in coordination, document tracking, and institutional memory.

4.4. Interpreting Network Patterns Through Enterprise Architecture Principles

The communication dynamics within the COLOURS alliance reflect more than just informal collaboration - they reveal a latent functional architecture that maps closely onto the TOGAF inspired Architecture Development Method (ADM) (Rouvrais & Petersen, 2024). At its core, View B: Business Architecture is designed to clarify who does what in an organization, by explicitly mapping roles, interdependencies, and critical processes. In COLOURS, key actors (such as Work Package leads and CIO leads) demonstrate consistently high levels of degree, eigenvector, and betweenness centrality - indicating that they serve as operational and informational hubs within the alliance. WP leads from UCLM and UNIOS act as functional anchors around which coordination tasks and knowledge flows are organized, while CIOs from HKR, LMU, UNIFE, and others bridge institutional and thematic clusters. The Managing Director, with the highest betweenness score, functions as a cross-functional process integrator, a role precisely described in View B's logic. Meanwhile, CoSpace Officers though less central globally - are embedded within tightly knit local clusters, supporting internal cohesion and continuity. This role-based architecture is further complicated by the lack of integration of actors such as student forums and associated partners, whose peripheral positions in the network highlight an absence of formal role embedding and process alignment - gaps that Business Architecture seeks to make visible and address. One solution could be the explicit architectural recognition of these actors within the alliance's business process models - assigning them specific coordination roles, feedback loops, or inclusion in structured communication routines. However, such inclusion cannot be operationalized effectively without also addressing the underlying digital fragmentation that characterizes COLOURS' current communication ecosystem. The use of multiple, uncoordinated tools - such as GoFAST, Google Drive, WhatsApp, and various institutional platforms - creates friction, reduces transparency, and limits institutional memory. In line with View F: Migration Planning, COLOURS must adopt a phased digital integration strategy that aligns technical systems with strategic roles. This could begin with a unified communications audit and stakeholder mapping, followed by the collaborative selection of one or two core platforms to support critical workflows such as project coordination and document management. These platforms should be co-designed with end-users to ensure usability and buy-in, and managed by designated change agents - such as CIO leads or Co-Space Officers - tasked with onboarding, aligning policies, and capturing feedback throughout the transition. In this way, digital convergence and role integration become mutually reinforcing, enabling the alliance to move beyond symbolic inclusion toward operational cohesion and strategic interoperability. Finally, the tension between institutional affiliation and emergent functional clustering, as revealed in the modularity analysis, directly speaks to View G: Implementation & Governance. This view addresses how alliances translate strategic ambitions into operational structures and accountability mechanisms. The presence of high centrality among operational leads - particularly WP leads - without clear institutional integration suggests a gap between functional influence and formal governance visibility. Interestingly, such a gap appears less pronounced for CIOs, who do operate within a formal Steering Committee structure within COLOURS. This is a promising governance mechanism that aligns well with View G's emphasis on structured oversight and role clarity. However, this committee was not modelled as a single node in the network analysis, meaning its collective coordination function is not directly visible in the results. In contrast, WP leads - despite their central communicative and strategic role - currently lack a comparable governance body. This asymmetry highlights a critical opportunity for governance evolution: the establishment of a cross-institutional WP Lead Council or Coordination Forum. Such a structure would provide a formal layer of accountability, peer coordination, and strategic alignment - making explicit what is currently emergent and informal. Embedding this forum within the alliance's governance architecture and reflecting it in documentation and communication flows would strengthen implementation coherence, reduce dependence on individual initiative, and bring the COLOURS alliance closer to the architecture-aligned governance envisioned in View G. By overlaying these ADM components onto the observed network patterns, it becomes evident that COLOURS is not merely an ad hoc collaboration, but an evolving, adaptive system with the early features of architecture-informed governance. This underscores the potential of Enterprise Architecture not just as a technical schema, but as a meta-model that bridges Social Network Analysis insights with role-based coordination, phased transformation, and long-term strategic alignment across institutional and national borders.

5. Conclusion

This work explored how governance design and communication and collaboration architectures intersect within the settings, contexts and needs of European university alliances, offering an integrated analysis grounded in social network analysis (SNA), enterprise architecture, and collaborative governance theory. The findings reveal that internal communication within our case study alliance is not evenly distributed, but instead reflects a latent architecture centered on a limited number of high-performing roles (especially WP leads and CIO leads). These

individuals occupy central positions in the network, showing high degree, betweenness, and eigenvector centrality, and function as key operational and strategic nodes. However, this centralization creates systemic vulnerabilities. Actors such as student forums, associated partners, and some peripheral CIO leads are marginalized (due to the high intensity of workflows) in the alliance's structural core. Their limited connectivity, paired with high local clustering, signals functional isolation and raises concerns about inclusivity, institutional learning, and long-term resilience. These patterns can be exacerbated by a fragmented digital infrastructure, with multiple uncoordinated tools undermining transparency and institutional memory. To address these gaps, we propose a set of architectural and facilitative interventions, as well as Enterprise Architecture reasoning as foundation for the governance complexity. This will help consolidate strategic and operational roles into a unified, multi-level governance body, which appears to be needed by alliances and builds on TOGAF's View G (Implementation & Governance), operationalizing alignment across roles, institutions, and platforms. It offers a structure for data-informed decision-making, policy coordination, and cross-role learning -anchored legally through institutional statutes and Memorandums of Understanding. Complementing this structural intervention are targeted facilitation mechanisms: such as Peer Mentorship Programs and Rotating Coordination Clinics, as well as intercultural training platforms. These interventions promote adaptive governance, distributed leadership, and inclusive co-creation. In sum, this work demonstrates that Alliances must go beyond project management logic to function as resilient, collaborative ecosystems. By diagnosing communication structures, designing adaptive governance models, and embedding facilitation as a strategic resource, Alliances can evolve into more integrated and future-proof transnational entities with sustainable, beyond project effectiveness.

Literature

1. Altbach, P. G., & Knight, J. (2007). The internationalization of higher education: Motivations and realities. *Journal of Studies in International Education*, *11*(3–4), 290–305. https://doi.org/10.1177/1028315307303542
2. Ansell, C., & Gash, A. (2007). *Collaborative governance in theory and practice*. Oxford University Press. https://doi.org/10.1093/jopart/mum032
3. Australian National University. (2014). *Association of Pacific Rim Universities*. https://www.anu.edu.au/association-of-pacific-rim-universities
4. Basel Myhub & Paderborn University. (2024). *COLOURS quality assurance plan*. Colours.
5. Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, *3*(1), 361–262. https://doi.org/10.1609/icwsm.v3i1.13937
6. Benneworth, P., Pinheiro, R., & Karlsen, J. (2017). Strategic agency and institutional change: Investigating the role of universities in regional innovation systems (RISs). *Regional Studies*, *51*(2), 235–248. https://doi.org/10.1080/00343404.2016.1215599
7. Benzinger, B., Canellas Lardies, J., Lapuenta, J., & Knoth, A. (2025). *Creating seamless learner experiences: Towards achieving interoperability in European University Alliances*. EPiC Series in Computing. EUNIS 2024 Annual Congress, Athens. https://easychair.org/publications/paper/bJKT
8. Borgatti, S. P., & Halgin, D. (2011). On network theory. *Organization Science*, *22*. https://doi.org/10.1007/978-1-4419-5513-5_2
9. Borgatti, S. P., Mehra, A., Brass, D., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892–895. https://doi.org/10.1126/science.1165821
10. Brooks, R., & Rensimer, L. (2025). Higher education actors' responses to the Ukraine-Russia conflict: An analysis of geopolitical spatial imaginaries. *Journal of Education Policy*, *40*(1), 111–129. https://doi.org/10.1080/02680939.2024.2334945
11. Charret, A., & Chankseliani, M. (2023). The process of building European university alliances: A rhizomatic analysis of the European Universities Initiative. *Higher Education*, *86*(1), 21–44. https://doi.org/10.1007/s10734-022-00898-6
12. Cleays, A.-L., Pruvot, E., Jørgensen, E., & Jørgensen, T. (2022). *The European Universities Initiative and system level reforms*. European University Association. https://www.eua.eu/publications/briefings/the-european-universities-initiative-and-system-level-reforms.html
13. Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, *94*, S95–S120.
14. Corlew, L. K., Keener, V., Finucane, M., Brewington, L., & Nunn-Crichton, R. (2015). Using social network analysis to assess communications and develop networking tools among climate change

professionals across the Pacific Islands region. *Psychosocial Intervention*, *24*(3), 133–146. https://doi.org/10.1016/j.psi.2015.07.004

15. Craciun, D., Kaiser, F., Kottmann, A., & Van der Meulen, B. (2023). *The European Universities Initiative: First lessons, main challenges and perspectives* (IPOL_STU(2023)733105). European Parliament, Policy Department for Structural and Cohesion Policies. https://www.europarl.europa.eu/RegData/etudes/STUD/2023/733105/IPOL_STU(2023)733105_EN.pdf
16. Diaz Olson, D. (2024, July). *Culture clash: Investigating interpersonal challenges between international and Dutch students* [Bachelor's thesis, University of Twente]. University of Twente Student Theses. https://essay.utwente.nl/101753/
17. Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press. https://www.cs.cornell.edu/home/kleinber/networks-book/
18. Emerson, K., Nabatchi, T., & Balogh, S. (2012). An integrated framework for collaborative governance. *Journal of Public Administration Research and Theory*, *22*(1), 1–29. https://doi.org/10.1093/jopart/mur011
19. Enders, J., & Fulton, O. (Eds.). (2002). *Higher education in a globalising world: International trends and mutual observations*. Springer.
20. EU Conexus. (2025). *ROADMAP on creating a sustainable governance and cooperation structure for a European University Alliance*.
21. EuniQ. (2020). *Developing a European approach for comprehensive QA of (European) university networks*.
22. European Commission. (2025, January 24). *European Universities alliances and their partners*. European Education Area. https://education.ec.europa.eu/education-levels/higher-education/european-universities-initiative/map
23. Feiel, S., Frühauf, S., Pichler, L., Kircher, V., Kosciuszko, A., & Egger, J. (2021). *EURECA-PRO, the European University on Responsible Consumption and Production: An alliance for sustainability*. https://doi.org/10.3217/978-3-85125-842-4-21
24. Gunningham, N., & Holley, C. (2016). Next-generation environmental regulation: Law, regulation, and governance. *Annual Review of Law and Social Science*, *12*, 273–293. https://doi.org/10.1146/annurev-lawsocsci-110615-084651
25. Haeckel, S. H. (1999). *Adaptive enterprise: Creating and leading sense-and-respond organizations*. Harvard Business School Press.
26. Jackson, D. J. (2011). *What is an innovation ecosystem?* National Science Foundation.
27. Kumar, A., Singh, S. S., Singh, K., & Biswas, B. (2020). Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and Its Applications*, *553*, 124289. https://doi.org/10.1016/j.physa.2020.124289
28. Lievens, P. (2024). *Maximizing the impact of European Universities alliances*. LERU.
29. O'Malley, A. J., & Marsden, P. V. (2008). The analysis of social networks. *Health Services & Outcomes Research Methodology*, *8*(4), 222–269. https://doi.org/10.1007/s10742-008-0041-z
30. Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, *32*(3), 245–251. https://doi.org/10.1016/j.socnet.2010.03.006
31. Ostrom, E. (2015). *Governing the commons*. Cambridge University Press.
32. Parlett, M., & Hamilton, D. (1972). *Evaluation as illumination: A new approach to the study of innovatory programs* (Occasional Paper No. 9). University of Edinburgh. https://eric.ed.gov/?id=ED167634
33. Petrevska Nechkoska, R. (2019). *Tactical management in complexity: Managerial and informational aspects*. Springer Cham.
34. Petrevska Nechkoska, R., Manceski, G., & Poels, G. (2023). *Facilitation in complexity: From creation to co-creation, from dreaming to co-dreaming, from evolution to co-evolution*. Springer. https://doi.org/10.1007/978-3-031-11065-8
35. Provan, K. G., & Kenis, P. (2008). Modes of network governance: Structure, management, and effectiveness. *Journal of Public Administration Research and Theory*, *18*(2), 229–252. https://doi.org/10.1093/jopart/mum015
36. Puranam, P., Raveendran, M., & Knudsen, T. (2012). Organization design: The epistemic interdependence perspective. *Academy of Management Review*, *37*(3), 419–440. https://doi.org/10.5465/amr.2010.0535

37. Rensimer, L., & Brooks, R. (2024). The European Universities Initiative: Further stratification in the pursuit of European cooperation? *Compare: A Journal of Comparative and International Education*, 1–19. https://doi.org/10.1080/03057925.2024.2307551
38. Rouvrais, S., & Petersen, S. A. (2024). An architecture framework for higher education. *Proceedings of the 16th International Conference on Computer Supported Education*, *2*, 739–746. https://doi.org/10.5220/0012738900003690
39. Ulibarri, N., Imperial, M. T., Siddiki, S., & Henderson, H. (2023). Drivers and dynamics of collaborative governance in environmental management. *Environmental Management*, *71*(3), 495–504. https://doi.org/10.1007/s00267-022-01769-7
40. Una Europa. (2022, November 7). *Engaging with diversity in European universities*. https://www.una-europa.eu/knowledge-hub/engaging-with-diversity-in-european-universities
41. Xia, Y., Tian, Z., & Ding, C. (2024). Collaborative governance in action: Driving ecological sustainability in the Yangtze River basin. *Frontiers in Environmental Science*, *12*. https://doi.org/10.3389/fenvs.2024.1463179

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:004.738.5.056(497-15)
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:10.01.2026.
DOI: 10.5937/SJEM2601119S

Улога вештачке интелигенције у јачању хибридних претњи на Западном Балкану

Marko Savković[1], Igor Novaković[2]

[1]ISAC Fund, Belgrade, Serbia. E-mail: marko.savkovic@isac-fund.org
[2]ISAC Fund, Belgrade, Serbia. E-mail: igor.novakovic@isac-fund.org

Сажетак: Вештачка интелигенција (ВИ) све више утиче на безбедносни пејзаж Западног Балкана, региона који је обележен крхким политичким окружењем и оспораваном управом. У овом раду се истражује како ВИ појачава хибридне претње као што су сајбер напади, олакшава мешање у изборни процес и омогућава дестабилизацију од стране државних и недржавних актера. Путем машинског учења и аутоматизованих система, злонамерни актери могу да креирају кампање дезинформација, искористе рањивости у дигиталној инфраструктури и манипулишу јавно мњење. Истраживање поставља поменуте појаве у шири контекст хибридних претњи, наглашавајући међудејство између технолошких иновација и системских слабости демократских институција. Анализом недавних случајева и нових трендова, у раду се тврди да безбедносни изазови које покреће ВИ на Западном Балкану нису само техничке природе, већ дубоко политички, те захтевају координисане одговоре који комбинују мере сајбер безбедности, регулаторне оквире и јачање отпорности на друштвеном и институционалном нивоу.

Кључне речи: вештачка интелигенција (ВИ), сајбер безбедност, изборне интерференције, хибридне претње, кампање дезинформација, Западни Балкан, политичка стабилност

AI's Role in Amplifying Hybrid Threats in the Western Balkans

Abstract: Artificial Intelligence (AI) is increasingly shaping the security landscape of the Western Balkans, a region marked by fragile political environments and contested governance. This paper examines how AI amplifies hybrid threats such as cyberattacks, facilitates election interference, and enables destabilization by both state and non-state actors. Through machine learning and automated systems, malicious actors can create disinformation campaigns, exploit vulnerabilities in digital infrastructure, and manipulate public opinion. The study places these developments within the broader context of hybrid threats, highlighting the interplay between technological innovation and systemic weaknesses in democratic institutions. By analyzing recent cases and emerging trends, the paper argues that AI-driven security challenges in the Western Balkans are not merely technical but deeply political, requiring coordinated responses that combine cybersecurity measures, regulatory frameworks, and resilience-building at societal and institutional levels.

Keywords: artificial intelligence (AI), cybersecurity, election interference, hybrid threats, disinformation campaigns, Western Balkans, political stability

1. Introduction

Hybrid threats are reshaping Europe's security landscape. They are blurring lines between war and peace, internal and external security, and state and non-state action (Hoffman, 2007; NATO, 2014). Strategies employed combine cyber operations, disinformation, political influence, economic pressure, and legal or institutional exploitation all while staying below the threshold of open conflict (EU Commission, 2016; Kofman & Rojansky, 2015). By expanding the reach and effectiveness of these tactics, AI has become their key enabler (Brundage et al., 2018).

Numerous and recent examples show how through machine learning hostile actors produce and spread disinformation quickly, tailor their messages, exploit digital weaknesses of intended targets, and manipulate public debate (Buchanan, 2020; Nemitz, 2018). In effect, trust, legitimacy and democratic integrity of one society

institutions are eroded (Risse, 2020). Yet within the policy and scholarly discussion, what we understand as deeply political nature of AI-enabled hybrid threats remains overlooked.

This new dynamic is clearly present in the Western Balkans, a region with weak democratic institutions, contested governance, polarization, and uneven progress toward Euro-Atlantic integration (Bieber, 2018; Keil & Perry, 2015). At the same time, societies and their respective governance systems have been digitalizing rapidly, often without strong regulation or safeguards. This has heightened exposure to manipulation (Freedom House, 2023) and has created favorable conditions for state and non-state actors to influence politics by using AI tools.

Within this paper we examine how AI has amplified hybrid threats in the Western Balkans, focusing on cyber operations, election interference, and disinformation. In our understanding AI is a "force multiplier" that, rather than creating new types of hybrid activity, exploits existing vulnerabilities (Brundage et al., 2018). It increases the speed, reach, precision, and deniability of political interference. Thus, rather than being merely technical, threats are made fundamentally political (Arkan, 2025).

The paper contributes to wider security debates by linking hybrid-threat theory (Hoffman, 2007) with emerging research on AI, highlighting Western Balkan-specific conditions, and calling for responses that combine cybersecurity, regulation, institutional reform, and societal resilience (Nemitz, 2018; Risse, 2020).

Methodologically, it uses a conceptual policy-analysis approach (Browne et al., 2019), drawing on cases and observable trends rather than full empirical testing - appropriate given fast-moving AI technologies and the opaque nature of hybrid operations (Buchanan, 2020). The aim is to identify how technological innovation interacts with political vulnerability.

2. Conceptual framework

Rather than viewing AI as a new or separate threat, this analysis places it within existing hybrid strategies, highlighting how AI strengthens political, informational, and cyber tactics already present in modern conflict. For instance, Romansky et al. (2024) argue that emerging hybrid threat strategies increasingly rely on "exploiting economic dependencies and manipulating societal polarization to undermine state resilience". Furthermore, malicious actors have weaponized digitalization and distorted information environments.

As for hybrid threats, we accept their common definition as "coordinated activities that combine military and non-military tools, involve both state and non-state actors, and use a mix of overt and covert methods to pursue strategic aims" (EU Commission, 2016; NATO, 2014). As such, we find them in the "grey zone between peace and war", exploiting legal and institutional gaps to achieve political effects, but without escalation (Mazarr, 2015). "Core goal" is not in territorial control but shaping perceptions, weakening institutions, and eroding social cohesion (Kofman & Rojansky, 2015), gradually, over time.

Hybrid threats have three main features. They are multidimensional, combining cyber operations, information manipulation, and economic pressure. They rely on plausible deniability, complicating attribution (who has done it?) and collective action (what can be done about it?). And they are highly context-specific, exploiting media fragmentation, polarization, and limited institutional capacity (Hoffman, 2007; Risse, 2020). These traits make them especially effective in regions with weak democratic and institutional consolidation or contested statehood (as in the case of Kosovo), such as Western Balkans.

Hybrid tactics are not new: propaganda, subversion, and covert influence have long been part of international politics. They "simply reflect long-standing methods of influence and coercion seen throughout earlier conflicts" (NATO, 2024). What distinguishes contemporary hybrid threats is the integration of digital technologies, which greatly increases the speed, reach, and coordination of such activities (Buchanan, 2020). As a result, modern strategies increasingly target the informational dimensions of security, challenging traditional, state-centric and military-focused ideas of threat and defense (Nemitz, 2018).

3. Artificial intelligence as a force multiplier in hybrid threats

AI is not a single technology but a set of methods. Machine learning, natural language processing, and automated decision-making can support many kinds of hybrid activity (Brundage et al., 2018). Which is why in this paper AI is not understood as a standalone threat, but a "force multiplier" (Hynes et al., 2025) that expands the scale, effectiveness, and adaptability of operations.

AI strengthens hybrid tactics in three main ways. First, it enables large-scale automation. Tasks such as content generation, data analysis, or cyber reconnaissance, previously requiring extensive human labor, can now be carried

out quickly and cheaply (Buchanan, 2020). In disinformation campaigns, AI systems generate tailored content, boost selected narratives, and fine-tune messages based on user reactions, lowering barriers to sustained influence operations (Ryan-Mosley, 2023). Second, AI increases speed and adaptability. Machine learning tools process big datasets in real time, detect emerging trends, and shift tactics quickly. In political or electoral contexts, this allows hybrid actors to exploit crises or uncertainty and shape public debate before fact-checkers or institutions can respond (Nemitz, 2018). Third, AI improves targeting and personalization. By analyzing behavior and social networks, AI can identify specific groups and deliver messages that match their identities or grievances. This makes hybrid operations more effective and less visible, especially in polarized societies (Risse, 2020).

AI does not cause instability by itself; its impact depends on political choices, institutional strength, regulation, and overall societal resilience (Brundage et al., 2018). Strong institutions can limit its effects, while weak governance allows AI-enabled hybrid threats to deepen existing vulnerabilities.

4. The Western Balkans as a hybrid threat environment

A core vulnerability in the Western Balkans is the fragility of democratic institutions (Nix & Tamisiea, 2025). Although democratic frameworks exist, their effectiveness, independence, and public trust remain uneven and often contested (Freedom House, 2023). Electoral oversight, media regulation, judicial autonomy, and accountability mechanisms are regularly exposed to political pressure (European Commission, 2025), reducing institutions' capacity to detect or counter covert interference. In such conditions, hybrid threats aimed at weakening institutional legitimacy can have massive effects.

Political polarization further heightens these weaknesses. The region continues to face deep ideological, ethnic, and identity-based divisions rooted in unresolved conflicts and competing historical narratives (Bieber, 2018). These divides create fertile ground for manipulation, allowing hybrid actors to amplify existing grievances, undermine political opponents, and weaken trust in institutions. AI-driven tools intensify this process through highly targeted and emotionally charged messaging that is difficult to detect or attribute.

The region's uneven digital transformation adds another layer of risk. While governments, political actors, media, and citizens have rapidly adopted digital platforms, this shift has not been matched by investments in cybersecurity, regulatory oversight, or digital literacy (Buchanan, 2020; Nemitz, 2018). Digital infrastructures and information ecosystems remain exposed to manipulation and abuse. This gap between fast technological adoption and weak institutional capacity makes the environment particularly conducive to AI-enabled hybrid threats.

The media landscape represents an additional structural vulnerability. Although pluralism formally exists, many outlets operate under economic pressure, political influence, or opaque ownership (Freedom House, 2023). Disinformation spreads easily, especially via social media. AI-generated content, including synthetic text, audio, and imagery, heightens these challenges by increasing the volume and credibility of misleading narratives. For example, An AI- generated fake interview video featuring a Bosnian politician circulated in August 2025, using synthetic speech to make it appear as though he made statements he never actually said. This unlabeled AI- manipulated clip spread across Bosnia and Herzegovina and the wider Western Balkan information space, creating public confusion as fact- checkers later confirmed its artificial origin (IFEX, 2025).

External influence also shapes the region's hybrid-threat environment. The Western Balkans sit at the intersection of European, transatlantic, and global power dynamics. Prolonged uncertainty around Euro-Atlantic integration creates incentives for external actors to rely on indirect, deniable forms of influence (Kofman & Rojansky, 2015). Hybrid strategies are attractive because they allow impact without open confrontation. AI further lowers engagement costs, increases deniability, and enables sustained operations across multiple domains.

Despite the dominant narrative portraying the Western Balkans as a "passive target" in reality it is a space where domestic political actors, media networks, and social groups actively amplify and disseminate foreign disinformation, enabling external actors to exploit existing ethnic and political divisions (Strategic Analysis, 2023). Hybrid threats in the region are not solely externally driven: local media platforms, political parties, and interest groups often themselves participate in spreading manipulated content, blending economic interests, corruption, and partisan competition, thus making the region a producer, not only a consumer, of hybrid activities (EUISS, 2024).

Taken together, these factors make the Western Balkans a highly permissive environment for hybrid threats, with AI acting as an amplifier rather than a root cause. The next section examines why such threats are deeply political.

5. Why AI-driven hybrid threats are political

Debates on AI and security often focus on technical risks, such as system robustness, cybersecurity, or platform governance. While they are important, these concerns overlook the political nature of AI-enabled hybrid threats. Also, it is often human error that makes cyber-attacks possible. As shown throughout this paper, AI acts as a force multiplier for existing influence strategies because it reshapes relations of power, legitimacy, and accountability within democratic systems (Brundage et al., 2018; Nemitz, 2018).

First, AI-driven hybrid threats target political authority and democratic legitimacy. Disinformation erodes trust; election interference aims to delegitimize the process; and cyber operations often signal state weakness (Benkler et al., 2018; Mazarr, 2015; Risse, 2020; Buchanan, 2020). As noted before, AI increases scale, speed, and deniability, but the core effect is political disempowerment: weakening public belief in democratic institutions' ability to mediate conflict.

Second, AI-enabled hybrid tactics are context-dependent. Their impact varies with institutional strength, media structures, and social cleavages. As explained, vulnerabilities in the Western Balkans stem from contested governance, polarized media, and uneven regulation (Bieber, 2018; Freedom House, 2023). AI does not act autonomously; rather, it amplifies existing power asymmetries and governance gaps (Keil & Perry, 2015; Nemitz, 2018).

Third, treating AI-driven hybrid threats as political reframes, or allows for, attribution and accountability. Hybrid tactics thrive on deniability and legal ambiguity (Hoffman, 2007; NATO, 2014). AI heightens these challenges by enabling cross-border, semi-automated operations that evade clear attribution. The core issue becomes one of governance: existing rules on political communication, campaign finance, data protection, and platform responsibility lag behind technological realities (Nemitz, 2018).

Fourth, having a political lens helps us see the cumulative nature of AI-enabled hybrid threats. Their most damaging effects emerge gradually (Benkler et al., 2018; Mazarr, 2015). In the Western Balkans, targeted influence and disruption locks societies into circles of low trust, incentivizing domestic actors to use similar tactics (Bieber, 2018; Risse, 2020).

Finally, recognizing the political character of these threats has direct policy implications. Technical measures, such as cybersecurity upgrades, authentication tools, detection systems, are necessary but insufficient. Lasting mitigation requires institutional and societal strategies: for instance, transparent rules for digital campaigning, independent oversight bodies, media reforms that strengthen professionalism and ownership transparency, civic and media literacy, and regional coordination to prevent regulatory gaps (EU Commission, 2016; Freedom House, 2023). Effective responses depend as much on parliaments, regulators, courts, and newsrooms as on technical infrastructure. All this, however, is not enough without whole of society campaigns – for instance, on digital hygiene.

Viewed this way, the Western Balkans offer a relevant case: AI accelerates hybrid tactics not because technology is determinative, but because political structures are contested and institutions remain fragile. Understanding AI as a political amplifier shifts attention toward governance configurations that shape both vulnerability and resilience.

6. Evidence from the Western Balkans

Once amplified by AI, disinformation will only reinforce existing skepticism toward democratic institutions and media across the region. During elections, protests, or disputes, such narratives that question the credibility of electoral bodies, courts, or independent media circulate widely, without offering coherent alternatives. In North Macedonia's 2018 name-change referendum, for example, coordinated online campaigns amplified claims of foreign manipulation and institutional bias, contributing to low turnout and public distrust. The referendum's 37% turnout was influenced by boycott campaigns and by accusations of Russian attempts to influence the outcome (New Eastern Europe, 2018). The effect across the region is cumulative: declining epistemic trust, normalization of suspicion, and weakening confidence in democratic procedures (Benkler et al., 2018; Risse, 2020). Because it exploits pre-existing distrust, AI will be especially effective here.

Rather than manipulating vote counts, malicious actors seek to undermine the perception of fairness. The online information space surrounding the 2020 parliamentary elections in Montenegro featured coordinated disinformation efforts, including narratives alleging electoral fraud, external influence, and questioning the legitimacy of institutions, which contributed to a climate of distrust and heightened perceptions of procedural

uncertainty (ENEMO, 2020). The long-term impact is the delegitimization of elections as instruments of democratic accountability (Mazarr, 2015). AI tools could enhance such strategies through micro-targeting and rapid adaptation.

Cyber operations in the region follow the same political logic. High-profile incidents, such as the 2022 coordinated cyberattacks on Albania's government services, produced lasting political fallout. In retaliation for Albania's hosting of the Mujahedin- e- Khalq (MEK), an exiled Iranian opposition group, state-linked actors carried out a cyber-attack against country's government that destroyed data and disrupted essential public services, including the e- Albania portal, and leaked Albanian government data, such as emails from senior officials (Foreign, Commonwealth and Development Office, 2022). Automation has the potential to make such incidents more frequent.

In Serbia, a well- documented illustration of the interplay between political contention and information manipulation emerged during the mass anti- government protests that followed the November 2024 collapse of the Novi Sad railway station canopy, widely perceived as emblematic of country's systemic corruption and institutional failure. As student- led demonstrations expanded into the largest civic mobilization in Serbia in decades, government actors and pro- government media deployed a coordinated narrative strategy aimed at delegitimizing the protests, framing them as a "foreign- orchestrated color revolution" rather than a domestic movement for accountability (Đorđević, 2024). This messaging was amplified through synchronized primetime broadcasts, quasi- expert commentary, and extensive use of bot networks disseminating identical pro- government talking points across social media, reflecting a hybrid information environment in which state- aligned outlets and digital assets operate in tandem (Ibid, 2024). Independent monitoring further indicates that these narrative operations were accompanied by intensified coercive measures: the CIVICUS Monitor and BIRN documented increased arrests of protesters, intimidation of student leaders, and smear campaigns portraying civil society actors as "foreign mercenaries," often linked to alleged Western efforts to destabilize Serbia (Baletić, 2024). Parallel analyses by Freedom House and the Reuters Institute observe a marked deterioration in online freedoms during this period, including the targeting of activists with spyware, the dominance of government-aligned media ecosystems, and the strategic marginalization of independent broadcasters, even as students relied heavily on social media platforms to mobilize and circulate uncensored information (Freedom House, 2024; Milivojević, 2025). Serbia's protest cycles are affected by hybrid governance practices, wherein information manipulation, media capture, and coercive policing mutually reinforce one another to contain dissent and reframe it as externally driven subversion rather than legitimate democratic action (Stojanović, 2025).

Another illustrative example concerns the 2023 case involving Željko Mitrović, owner of TV Pink, who publicly promoted what he described as an "AI- based" system capable of generating satirical political content. In practice, however, the system was used to fabricate video and audio representations of opposition politicians, placing statements in their mouths that they had never made. The generated videos were broadcast in Pink's main news programs, provoking significant public and political criticism. President Aleksandar Vučić himself condemned the practice as "unfair" and inappropriate for a democratic society, explicitly stating that artificial intelligence should not be used to falsify individuals' speech or likeness. (Insajder, 2023; Nova.rs, 2023). This event demonstrates how AI- labeled technologies can be operationalized within hybrid media systems to distort political communication. It also illustrates the strategic ambiguity of "AI" as a discursive tool: while Mitrović presented the content as benign satire enabled by advanced technology, its effect was the production of politically consequential deepfakes.

Finally, second example concerns a 2024 incident in which an audio recording surfaced allegedly capturing Damir Zobenica, high- ranking SNS official and Vice- President of the Assembly of Vojvodina, giving detailed instructions to activists on how to provoke incidents during civic protests. When the recording was published by opposition figure Marinika Tepić, President Vučić responded by stating that Zobenica had informed him the audio was produced using artificial intelligence, implicitly suggesting that the recording was a deepfake. (N1, 2024). Independent audio- forensics experts quickly challenged the claim, arguing that the recording exhibited natural speech patterns inconsistent with current AI- generated Serbian- language voice synthesis. Audio engineer Dejan Tomka stated that producing such an authentic- sounding recording using existing AI tools would be "impossible," pointing to emotional cues, breathing, hesitation patterns, and tonal consistency that contemporary models cannot reliably reproduce. (N1, 2024b.) This case exemplifies a different but equally significant political use of AI: invoking the idea of artificial intelligence as a rhetorical shield to deflect allegations.

At the regional level, external influence interacts with internal vulnerability. Prolonged uncertainty around Euro-Atlantic integration (while Albania, Montenegro and North Macedonia are NATO members, Serbia, Bosnia and Herzegovina and Kosovo are not) and geopolitical competition create incentives for interventions that can

later be denied. Region- wide disinformation and influence campaigns have been documented across Bosnia and Herzegovina, Kosovo, North Macedonia, and Serbia; malign narratives circulate seamlessly across borders.

Serbia functions as a central hub for pro- Kremlin messaging, exporting narratives that delegitimize the EU and Western institutions into Montenegro, Bosnia and Herzegovina, and North Macedonia, where they are adapted to local grievances and ethnic divisions (Press Room, 2024). The same patterns emerge in the EUISS regional assessment, which highlights that disinformation about the EU, NATO, and Western actors travels with ease across Western Balkan borders, enabled by shared language ecosystems, high media interoperability, and cross-border political networks (EUISS, 2021).

Serbia's centrality within the regional information environment is structural rather than externally imposed, stemming from a combination of linguistic, media, political, and digital factors that enable narratives originating in Belgrade to circulate widely across neighboring states. The shared linguistic space of the Western Balkans allows Serbian- language media content to flow easily into Montenegro, Bosnia and Herzegovina, and parts of Kosovo and North Macedonia, giving Serbian broadcasters a built- in cross- border audience. This influence is reinforced by the dominance of pro- government outlets, such as Pink and Happy, across regional cable packages, which means that Serbian political narratives routinely spill into adjacent media markets and shape public discourse beyond Serbia's borders (Milivojević, 2025). Additionally, political networks aligned with Belgrade, including parties and officeholders in Republika Srpska, Montenegro, and segments of North Macedonia, help reproduce and legitimize Serbian state- aligned messaging across their own domestic spheres.

These activities are not driven predominantly by Russian (or other foreign) actors; rather, they rely on domestic political elites, aligned media conglomerates, and cross- border networks of local influencers, who deploy Russia- compatible narratives because they serve their own power consolidation strategies. In examining Serbia's contemporary media ecosystem on Facebook, it is notable that the principal pro- government pages do not exhibit operational or organizational ties to Russian state- aligned outlets; their messaging strategies remain largely domestically rooted and aligned with ruling- party communication priorities. By contrast, a parallel ecosystem of grey- zone platforms, such as IN4S and Srbin.info, maintains clear ideological affinities with Kremlin- aligned narratives and frequently amplifies content compatible with Russian geopolitical messaging. These outlets have historically served as vectors for disinformation campaigns and influence operations, and until at least 2022 were often sharply critical of President Vučić, portraying him as insufficiently aligned with Moscow's strategic interests. This dual structure, state- aligned Facebook pages on one side and Russian- connected grey- zone portals on the other, reflects what the ISAC Fund's Vulnerability Index – Serbia identifies as a fragmented information landscape particularly susceptible to malign foreign influence, especially in the domain of Kremlin- linked information operations (ISAC Fund, 2021).

This is the context in which AI enters the fray. In October 2024, the European Centre of Excellence for Countering Hybrid Threats (Hybrid CoE) organized a major wargame in Vienna, Austria, focused specifically on hybrid threats aimed at the Western Balkans. Over four days (7–10 October), national teams, including delegations from Serbia, Montenegro, and North Macedonia, participated in a simulation built around a destructive earthquake scenario. Within the exercise, malign actors used "novel AI tools" to generate fabricated emergency announcements and AI- produced visuals portraying institutional failure, with the goal of destabilizing the regional information environment and eroding public trust in state authorities (Hybrid CoE, 2024).

In Albania, the 2024 disinformation landscape documented by SEE Check showed a marked expansion in the use of AI- generated manipulated visuals spreading across Facebook, Instagram, and anonymous online platforms. The Tirana- based fact- checking organization Faktoje verified more than 1,000 misleading claims during 2024, many of which relied on AI- fabricated or AI- enhanced imagery designed to mimic legitimate news reporting. These synthetic visuals contributed to a significant erosion of public confidence in Albanian media, institutions, and democratic processes (SEE Check, 2025).

These developments occurred alongside high- profile AI- driven political interference elsewhere in Europe, most notably during the 2023 Slovak parliamentary elections, where a deepfake audio recording targeted pro- European candidate Michal Šimečka. Scholars and analysts highlighted the Slovak incident as a test case for the type of AI- driven hybrid operations likely to be replicated in the Western Balkans, given similar low- trust political environments and susceptibility to pro- Russian disinformation (de Nadal & Jančárik, 2024).

Taken together, these cases illustrate how the Western Balkans has already entered a new phase of hybrid- threat exposure in which AI is not hypothetical, but operational. Technical defenses, cybersecurity upgrades, content moderation, detection tools are necessary but insufficient. Durable resilience depends on institutional reform, regulatory clarity, transparent media ownership, stronger professional standards, and societal capacity to resist

manipulation (EU Commission, 2016; Freedom House, 2023). Effective responses must therefore address the political foundations of democratic governance, not just technical vulnerabilities.

7. Policy responses

If AI-driven hybrid threats are treated only as technical problems, responses will remain reactive and fragmented. As shown earlier, AI amplifies hybrid threats by exploiting political, institutional, and societal vulnerabilities - conditions especially visible in the Western Balkans. Effective mitigation therefore requires a governance-centered approach that links cybersecurity with regulatory reform, institutional strengthening, and societal resilience.

Technical defenses remain essential, particularly for public institutions, electoral bodies, and critical infrastructure. Several Western Balkan governments have begun to modernize their cyber frameworks: North Macedonia's National Cybersecurity Strategy (2023) strengthens Computer Emergency Response Team (CERT) functions and mandates incident reporting, while Serbia's 2022 Law on Information Security introduces stricter risk-assessment requirements for operators of essential services (Government of North Macedonia, 2023; CERT Serbia, 2022). AI can improve anomaly detection and automated threat monitoring, but must operate within systems that ensure transparency and accountability (Buchanan, 2020).

Regulatory upgrades are equally necessary. Legal frameworks governing political advertising, data protection, and platform accountability still lag behind the realities of AI-enabled manipulation. Some progress exists - for example, Montenegro's 2023 media-law amendments introduced stronger transparency requirements for online political advertising, while Bosnia and Herzegovina has moved toward GDPR (General Data Protection Regulation) – inspired data-protection standards (Montenegro Ministry of Culture & Media, 2023; BiH Personal Data Protection Agency, 2022).

Yet regulatory reforms must avoid over-securitization. Treating too many social, political, and informational challenges primarily as security threats produces distortions that ultimately weaken, rather than strengthen, democratic resilience. Overly restrictive rules risk being used to control speech or target political opponents, already visible in the region through selective takedown practices and ambiguous "fake news" provisions. Safeguarding fundamental rights and ensuring independent oversight remain central (Nemitz, 2018).

Institutional resilience is just as important as technical security. Independent electoral commissions, media regulators, data-protection authorities, and courts provide the backbone of democratic legitimacy under persistent hybrid pressure. Practical examples include the Central Election Commission of Kosovo, which monitors online political advertising during elections, and North Macedonia's Agency for Audio and Audiovisual Media Services (AAVMS), which conducts transparency audits of broadcasters and online portals (CEC Kosovo, 2021; AAVMS, 2022).

Media ecosystems remain a significant vulnerability. Greater transparency in media ownership, sustainable funding for independent journalism, and support for fact-checking organizations can reduce susceptibility to AI-enhanced manipulation. Initiatives such as the Balkan Fact-Checking Network (BFCN), RCC-supported regional disinformation-monitoring programs, and EU IPA III assistance for investigative journalism illustrate practical steps already underway (Freedom House, 2023; RCC, 2022).

Societal resilience depends heavily on civic and media literacy. Public-education programs, such as Serbia's "Check Before You Share," UNDP Montenegro's media-literacy pilots, or EU-supported digital-literacy schools in North Macedonia, improve citizens' ability to identify manipulation and understand algorithmic dynamics (UNDP Montenegro, 2023). In contexts marked by polarization and disengagement, such initiatives strengthen civic agency and reduce vulnerability to hybrid interference.

8. Regional and European coordination

Governments in the Western Balkans face limits in addressing these challenges through purely national measures. Recent regional initiatives show how information sharing, joint training, and coordinated response can strengthen collective resilience. For example, the Western Balkans Cyber Capacity Centre (WB3C), launched in 2023 by France, Slovenia, and regional partners, provides shared cyber incident reporting channels, analytic workshops, and AI-related training modules for national CERT teams (WB3C, 2023). The Regional Cooperation Council (RCC) has also facilitated cooperation against disinformation through regional exchanges among fact-checking organizations and joint monitoring of cross-border influence campaigns (RCC, 2022). Regular bilateral cybersecurity exercises, such as the Serbia–North Macedonia joint cyber drills held since 2021, simulate

coordinated attacks on government platforms and critical infrastructure, improving detection and response capacity (CERT Serbia, 2022). These mechanisms help mitigate asymmetries in capacity and reduce opportunities for hybrid actors to exploit weaker states or institutions.

At the European level, deeper integration of the Western Balkans into EU security, digital, and regulatory frameworks should provide a critical long-term resilience strategy. The EU Hybrid Fusion Cell enhances situational awareness by collecting, analyzing, and sharing information on hybrid threats among EU members and Western Balkan partners (EU INTCEN, 2016). The region has increasingly participated in training programs organized by the European Centre of Excellence for Countering Hybrid Threats (Hybrid CoE), which supports table-top exercises and scenario planning on hybrid and AI-enabled interference (Hybrid CoE, 2021). The EU–Western Balkans Cybersecurity Blueprint Exercise mirrors the EU's Blue OLEx process, helping partner states practice coordinated regional responses to large-scale cyber incidents (ENISA, 2023).

Regulatory alignment is also advancing. Through the accession process, Western Balkan governments have begun approximating elements of the Digital Services Act (DSA), Digital Markets Act (DMA), and the emerging AI Act, including rules on platform accountability, transparency of political advertising, and algorithmic governance (EU Commission, 2021; 2023). Importantly, EU support extends beyond technical standards: initiatives under IPA III, the EU Rule of Law missions, and media-sector assistance programs provide institutional reform, independent regulatory capacity, and support for professional journalism (EU Parliament, 2022). These combined measures help anchor Western Balkan states within rule-based European norms and strengthen their ability to counter hybrid and AI-enabled threats.

9. Conclusion

This paper has examined how artificial intelligence amplifies hybrid threats in the Western Balkans, arguing that AI-driven security challenges are fundamentally political rather than merely technical. By situating AI within established hybrid-threat frameworks, the analysis shows that AI acts as a force multiplier - expanding the scale, speed, adaptability, and deniability of existing tactics of influence and disruption. Its impact depends not only on technological capabilities, but on the quality of democratic governance, institutional resilience, and societal trust.

The conceptual framework clarifies that AI is not an autonomous source of instability. As shown across disinformation, election interference, and cyber operations, AI does not create new forms of hybrid activity; it intensifies pre-existing political dynamics, especially in contexts marked by polarization, weak oversight, and contested legitimacy. The Western Balkans illustrate how such conditions generate permissive hybrid-threat environments, where AI-enabled tools produce cumulative political effects even when individual incidents appear limited.

Empirically and analytically, the paper contributes to international relations and security studies in three ways. First, it connects hybrid-threat scholarship with emerging work on AI and political security, emphasizing interaction effects rather than technological determinism. Second, it provides a region-specific perspective that treats the Western Balkans not as exceptional, but as a setting where broader European and global dynamics appear in concentrated form. Third, it reinforces a process-oriented understanding of hybrid threats, showing how repeated and adaptive interference gradually erodes democratic legitimacy.

Policy implications follow directly. Technical responses, such as cybersecurity upgrades, detection tools, platform governance are necessary but insufficient. Durable resilience requires strategies that strengthen democratic institutions, clarify regulatory authority over digital political spaces, support more transparent and credible media ecosystems, and invest in societal capacity to recognize and counter manipulation. For the Western Balkans, closer alignment with EU regulatory frameworks and deeper regional cooperation are especially important for reducing asymmetries that hybrid actors exploit.

More broadly, the findings suggest that AI-driven hybrid threats challenge not only security policy, but democratic governance in the digital age. As AI capabilities continue to expand, the gap between technological change and democratic oversight risks widening. Addressing this gap is ultimately political: it depends on reinforcing public accountability, institutional legitimacy, and democratic control over the infrastructures that shape political communication.

References

1. AAVMS. (2022). *Annual transparency report on media services*. Agency for Audio and Audiovisual Media Services.
2. Arkan, Z. (2025). Conclusion: Hybrid threats, shared stories - Narratives of security in NATO and the EU. In *European security and hybrid threats* (pp. 67–69). Springer.
3. Baletić, K. (2024). *Protesters' arrests fuel human rights concerns in Serbia, report says*. Balkan Insight.
4. Benkler, Y. (2023). Disinformation, social media, and the democratic crisis. *Journal of Democracy, 34*(1), 50–64.
5. Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.
6. Bieber, F. (2018). Patterns of competitive authoritarianism in the Western Balkans. *East European Politics, 34*(3), 337–354.
7. BiH Personal Data Protection Agency. (2022). *Guidelines on GDPR alignment*.
8. Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy, 3*, e32.
9. Browne, J., Coffey, B., Cook, K., Meiklejohn, S., & Palermo, C. (2019). A guide to policy analysis as a research method. *Health Promotion International, 34*(5), 1032–1044.
10. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., … Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. Future of Humanity Institute, University of Oxford.
11. Buchanan, B. (2020). *The hacker and the state: Cyber-attacks and the new normal of geopolitics*. Harvard University Press.
12. CEC Kosovo. (2021). *Election integrity and online campaign monitoring report*. Central Election Commission of Kosovo.
13. CERT Serbia. (2022). *Annual report on cybersecurity exercises in regional cooperation*. Ministry of Information and Telecommunications.
14. CERT Serbia. (2022). *Information security annual report*. Ministry of Information and Telecommunications.
15. de Nadal, L., & Jančárik, P. (2024). Beyond the deepfake hype: AI, democracy, and "the Slovak case". HKS Misinformation Review.
16. Đorđević, T. (2024). *When the regulators are raised: How does the propaganda machine "extinguish" the crisis in Serbia?* Istinomer.
17. ENEMO. (2020). *International Election Observation Mission: Montenegro parliamentary elections, 30 August 2020 – Final report*. European Network of Election Monitoring Organizations.
18. ENISA. (2023). *EU–Western Balkans cybersecurity blueprint exercise report*. European Union Agency for Cybersecurity.
19. European Commission. (2016). *Joint framework on countering hybrid threats: A European Union response* (JOIN (2016) 18 final).
20. European Commission. (2021). *Proposal for a Regulation on a Single Market for Digital Services (Digital Services Act)* (COM (2020) 825 final).
21. European Commission. (2022). *Strengthened Code of Practice on Disinformation*.
22. European Commission. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union*.
23. European Commission. (2024). *Regulation (EU) 2024/900 on the transparency and targeting of political advertising. Official Journal of the European Union*.
24. European Commission. (2025). *Rule of Law Report 2025: Candidate countries - Albania, Montenegro, North Macedonia, and Serbia*.
25. European Parliament. (2022). *IPA III programming document – Governance and rule of law*.
26. EU Institute for Security Studies. (2021). *The Western Balkans and EU–NATO cooperation: How to counter foreign interference and disinformation?*
27. European Union Institute for Security Studies. (2024). *Countering cyber-enabled hybrid interference in the Western Balkans: A scenario-based approach*.
28. EU Intelligence and Situation Centre. (2016). *EU Hybrid Fusion Cell: Mandate and functions*.
29. Foreign, Commonwealth & Development Office. (2022, September 7). *UK condemns Iran for reckless cyber-attack against Albania*. UK Government.
30. Freedom House. (2023). *Nations in Transit 2023: The repressive turn in the Western Balkans*.

31. Government of North Macedonia. (2023). *National Cybersecurity Strategy 2023–2028*.
32. Hoffman, F. G. (2007). *Conflict in the 21st century: The rise of hybrid wars*. Potomac Institute for Policy Studies.
33. Hybrid CoE. (2021). *Training and exercises catalogue*. European Centre of Excellence for Countering Hybrid Threats.
34. Hybrid CoE. (2024). Western Balkans in focus at countering disinformation wargame and conference in Vienna. Hybrid Centre of Excellence for Countering Hybrid Threats.
35. Hynes, P., Bew, R., Williamson, J. R., & Kenyon, M. (2025). *AI as a cybersecurity risk and force multiplier*. National Association of Corporate Directors.
36. IFEX. (2025, August 29). *Detecting fake content online in the Balkans*. Mediacentar Sarajevo.
37. Insajder. (2023). Vučić kaže da veštačka inteligencija ne sme da se koristi na način na koji to čini vlasnik TV Pink. https://www.insajder.net/prenosimo/vucic-kaze-da-vestacka-inteligencija-ne-sme-da-se-koristi-na-nacin-na-koji-to-cini-vlasnik-tv-pink
38. Keil, S., & Perry, V. (Eds.). (2015). *State-building and democratization in the Western Balkans*. Routledge.
39. Kofman, M., & Rojansky, M. (2015). *A closer look at Russia's "hybrid war"*. Kennan Institute, Wilson Center.
40. Mazarr, M. J. (2015). *Mastering the gray zone: Understanding a changing era of conflict*. U.S. Army War College Press.
41. Milivojević, S. (2025). Serbia. In *Digital News Report 2025*. Reuters Institute for the Study of Journalism.
42. Montenegro Ministry of Culture & Media. (2023). *Draft law on media – Amendments*.
43. N1. (2024). Vučić o navodnom snimku Zobenice: Poslao mi je poruku da je to veštačka inteligencija.
44. N1. (2024b). Veštačka inteligencija ili Zobenica "stasom i glasom": Stručnjak tvrdi – nemoguće da AI napravi tako autentičan snimak.
45. NATO. (2014). *Wales Summit Declaration*.
46. NATO. (2024). *Hybrid threats and hybrid warfare: Reference curriculum*. NATO Headquarters.
47. Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A, 376*(2133), 20180089.
48. New Eastern Europe. (2018, October 1). *The Macedonian name change referendum*. New Eastern Europe.
49. Nix, S., & Tamisiea, M. (2025). *Democracy at a crossroads: Rule of law and the case for US engagement in the Balkans*. Atlantic Council.
50. Nova.rs. (2023). "To je nefer, to ne sme da se radi…": Vučić o „novom projektu "Željka Mitrovića.
51. Press Room (DISA). (2024, December 18). *Influence and disinformation campaign monitoring in the Western Balkans (MEDIWEB report)*.
52. Regional Cooperation Council. (2022). *Countering disinformation in the Western Balkans – Regional assessment*.
53. Regional Cooperation Council. (2022). *Countering disinformation in the Western Balkans*.
54. Risse, T. (2020). *Governance without a state? Policies and politics in areas of limited statehood*. Columbia University Press.
55. Romansky, S., Hoenig, A., Meessen, R., & Kruijver, K. (2024). *New technologies, changing strategies: Five trends in the hybrid threat landscape*. The Hague Centre for Strategic Studies.
56. Ryan-Mosley, T. (2023). How generative AI is boosting the spread of disinformation and propaganda. *MIT Technology Review*.
57. SEE Check. (2025). Disinformation report: Albania in 2024. SEE Check / Faktoje. https://seecheck.org/index.php/2025/06/05/disinformation-report-albania-in-2024/
58. Strategic Analysis. (2023). *Hybrid threats in the Western Balkans: State and non-state perspective*. Strategic Analysis Think Tank.
59. UNDP Montenegro. (2023). *Media and digital literacy programme outcomes*. United Nations Development Programme.
60. Western Balkans Cyber Capacity Centre. (2023). *Programme overview*. Government of France & Government of Slovenia.

Instructions for Authors

The Journal Committee strives to maintain the highest academic standards. The submitted papers should be original and unpublished until now. Also, it is forbidden that papers are in the process of reviewing in some other publication.

The papers would be subjected to check. The paper should fit the outlined academic and technical requirements.

Paper Types

Original unpublished scientific paper:

- Original scientific paper;
- Plenary lecture and paper presented at the conference;
- Review paper;
- Scientific review; discussion.

Original unpublished professional paper:

- Original professional paper;
- Contribution
- Book review.

Papers may be written in Serbian and English for authors from Serbia and the region or English for authors from other countries.

Submitted papers must be in alignment with guidelines for authors. In case they have not followed these guidelines, they would be reviewed for correction.

All manuscripts are subject to *double blind review*, i.e. the process of double "blind" anonymous reviewing. The papers must not contain any references which may indicate the author(s).

Paper Submission

Authors should send their papers via email casopis@fim.rs in .doc or .docx format.

The application consists of two separate attachments:

- Attachment 1, which contains the following data: the title of paper, author's name (without professional title), institution and address (email, postal address, phone number), as well as the asterisk next to the author in charge of correspondence;
- Attachment 2, which contains the paper with the following elements: paper title, abstracts, key words, the middle part of the paper, tables, graphs, references and attachments.

Authors, who pass the *double-blind* anonymous review, will receive the document called the Author's Statement of Originality, which will be filled in, underlined, scanned and sent to the email: casopis@fim.rs.

Paper content

All papers should contain: introduction, which elaborates on the aim and subject of the research, main hypothesis, work methods and paper structure; middle part of the paper where research is outlined (it is further divided into sub-headings) and conclusion, which represents summed up results and implications for further research.

Author's rights

After accepting the paper and signing up the Author's Statement of Originality, the author signs the statement according to the Author's Rights of the Journal.

Author's editions

Authors of published papers will receive one print version of the paper for their personal usage.

Paper submissions:

Papers should be submitted via email: casopis@fim.rs.

Uputstvo za autore

Uredništvo časopisa nastoji da održi visok akademski standard. Radovi, koji se podnose, treba da budu originalni i do sada neobjavljeni. Takođe, radovi ne smeju da se nalaze u postupku recenzije u nekom drugom časopisu. Radovi će biti podvrgnuti proveri. **Tekst rada mora da odgovara akademskim i tehničkim zahtevima.**

Tip rada

Originalni naučni rad, koji nije objavljen:

- Originalni naučni rad;
- Plenarno predavanje i rad prezentovan na konferenciji;
- Pregledni rad;
- Naučna kritika, odnosno polemika.

Originalni stručni rad, koji nije objavljen:

- Stručni rad;
- Informativni prilog;
- Prikaz knjige.

Jezici radova mogu biti srpski i engleski za autore iz Srbije i engleski za autore sa drugih govornih područja.

Podneti radovi moraju biti usaglašeni sa uputstvom za autore. U slučaju da nisu usaglašeni, biće vraćeni na ispravljanje.

Svi rukopisi podležu tzv. *double blind* recenziji, odnosno procesu dvostruko „slepe“, anonimne recenzije. Tekst rada ne sme da sadrži bilo kakve reference koje mogu da ukažu na autora/e rada.

Prijava radova

Autori treba da pošalju svoje radove elektronski, putem i-mejla casopis@fim.rs u vidu priloga u .doc ili .docx formatu.

Prijava se sastoji iz dva odvojena priloga:

- Prilog 1, koji sadrži sledeće podatke: naslov rada, imena autora (bez titula i zvanja), institucija/e i adresa/e (i-mejl, poštanska adresa, broj telefona), kao i zvezdicu kod imena autora koji je zadužen za korespondenciju;
- Prilog 2, koji sadrži rad sa sledećim elementima: naslov rada, apstrakt/i, ključne reči, središnji deo rada, slike, tabele, grafikoni, reference, prilozi;

Autorima, koji prođu dvostruko anonimnu recenziju, biće poslat dokument Izjave autora o originalnosti rada, koji će popuniti, potpisati, skenirati i poslati na i-mejl casopis@fim.rs.

Sadržaj rada

Svi rukopisi treba da sadrže: uvod, koji čine cilj i predmet istraživanja, osnovna hipoteza, metode rada i struktura rada; središnji deo rada u kome se prikazuje istraživanje (dalje podeljen na potpoglavlja) i zaključak, koji predstavlja sumiranje rezultata istraživanja kao i implikacije za dalja istraživanja.

Autorska prava

Po prihvatanju rada i potpisivanje izjave o originalnosti, autor potpisuje izjavu kojom prenosi autorska prava na Časopis.

Autorski primerci

Autori publikovanih radova će dobiti primerak štampane verzije časopisa za lično korišćenje.

Dostavljanje radova:

Radovi se dostavljaju putem i-mejla casopis@fim.rs.

List of Reviewers/Spisak recenzenata

Editorial Board concluded this issue on Janaury 30, 2026.
Uređivački odbor je zaključio ovaj broj 30. januara 2026.

ISSN: 2466-4693

Contact/Kontakt:
Serbian Journal of Engineering Management
Editorial Board/Uredništvo
School of Engineering Management/Fakultet za inženjerski menadžment
Bulevar vojvode Mišića 43
11000 Beograd
casopis@fim.rs
Tel. +381 11 41 40 425

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

005:62

SERBIAN Journal of Engineering Management /
glavni i odgovorni urednik Vladimir Tomašević. - Vol.
1, no. 1 (2016)- . - Beograd : Univerzitet "Union -
Nikola Tesla", Fakultet za inženjerski menadžment,
2016- (Beograd : Draslar Partner). - 30 cm

Polugodišnje.
ISSN 2466-4693 = Serbian Journal of Engineering
Management
COBISS.SR-ID 224544524