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

Special Issue

Belgrade, February 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

Special Issue

Belgrade, February, 2026
Beograd, februar, 2026

Serbian Journal of Engineering Management
Special Edition, 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

Serbian Journal of Engineering Management is a scientific journal, published by School of Engineering Management and Society of Engineering Management of Serbia.The Journal is cathegorized by the Ministry of science, technological development, and Innovationof the Republic of Serbia. From 2020, the Journal is indexedat EBSCO databases. The Journal is indexed at the ERIH Plus list since 2023. This international Journal is dedicated to the wide scope of themes associated to engineering management and industrial engineering and is published semiannualy. The papers are presented in English.
This special issue of the Serbian Journal of Engineering Management addresses the intersection of artificial intelligence and security through several interconnected themes: the geopolitical dimensions of AI competition between major powers, particularly the US-China technological rivalry and its impact on global order and digital sovereignty; military applications including autonomous weapons systems, AI-enabled warfare, and defense capabilities; risk assessment and governance frameworks for AI deployment in security infrastructures; economic and legal aspects of AI regulation and international cooperation; the role of AI in amplifying hybrid threats and information warfare, especially in regional contexts like the Western Balkans; ethical implications and regulatory challenges of autonomous systems for both military purposes and environmental security; industrial safety applications through machine learning in critical processes; and organizational governance challenges including the protection of technical documentation from unauthorized disclosure to large language models and the management of AI systems in transnational institutional settings.
Editorial board is consisted of distinguished academics from various countries dedicated to establishing the highest academic standards and promoting engineering management principles in Serbia.

Information on the journal in English and Serbian can be found at the journal web page: 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 posebno 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 two special issues of the Serbian Journal of Engineering Management, dedicated to one of the most pressing and transformative challenges of our time: the intersection of artificial intelligence and security in the 21st century. This collection emerged from the international scientific conference "Artificial Intelligence and Security in the 21st Century", held in November 2025, which brought together scholars, practitioners, and policymakers to examine how AI is fundamentally reshaping the architecture of global security.

The convergence of artificial intelligence and security studies represents far more than a technological evolution, it marks a paradigm shift in how we understand power, governance, conflict, and human agency. As AI systems become increasingly capable of autonomous decision-making, predictive analysis, and large-scale information processing, they introduce unprecedented opportunities and equally significant risks across military, economic, social, and political domains. These special issues address these complexities through rigorous interdisciplinary scholarship that bridges computer science, international relations, ethics, law, and engineering management.

Both issues of the Serbian Journal of Engineering Management addresses the intersection of artificial intelligence and security through several interconnected themes: the geopolitical dimensions of AI competition between major powers, particularly the US-China technological rivalry and its impact on global order and digital sovereignty; military applications including autonomous weapons systems, AI-enabled warfare, and defense capabilities; risk assessment and governance frameworks for AI deployment in security infrastructures; economic and legal aspects of AI regulation and international cooperation; the role of AI in amplifying hybrid threats and information warfare, especially in regional contexts like the Western Balkans; ethical implications and regulatory challenges of autonomous systems for both military purposes and environmental security; industrial safety applications through machine learning in critical processes; and organizational governance challenges including the protection of technical documentation from unauthorized disclosure to large language models and the management of AI systems in transnational institutional settings.

The selection process for these special issues was particularly rigorous and competitive. Each manuscript underwent double-blind peer review by distinguished international experts in security studies, computer science, international relations, and engineering management. Reviewers evaluated submissions not only for methodological rigor and theoretical contribution but also for practical relevance to policymakers, security practitioners, and technology developers navigating the complex landscape of AI-enabled security systems. The papers included in this volume represent the highest caliber of scholarship, offering both analytical depth and actionable insights.

What distinguishes this collection is its balanced and nuanced approach to examining AI's dual nature in security contexts. The articles demonstrate convincingly that AI's impact depends fundamentally on the governance frameworks, ethical guidelines, regulatory mechanisms, and strategic choices made by states, international organizations, technology developers, and civil society actors.

The geographical and institutional diversity of our contributors, spanning Europe, North America, the Middle East, and East Asia ensures multiple perspectives on AI security challenges. This diversity is particularly valuable given that AI governance cannot be approached through a single cultural, political, or economic lens. The Western Balkans perspective, well-represented in this volume, offers crucial insights for medium and smaller states navigating between competing technological blocs while seeking to maintain strategic autonomy, protect national interests, and ensure that AI development serves democratic values and human rights.

Several critical themes emerge across the contributions to this special issue:

First, the geopolitical dimension of AI competition is reshaping international order, with major powers racing to achieve technological supremacy in AI capabilities. This competition carries significant implications for global stability, alliance structures, technological standards, and the future balance of power. Our contributors examine how states can navigate these dynamics while avoiding an AI arms race that could destabilize international security.

Second, the military applications of AI from autonomous weapons systems to AI-enabled intelligence analysis and cyber operations, raise profound ethical, legal, and strategic questions. The papers in this volume critically examine the promises and perils of military AI, addressing issues of accountability, human control, compliance with international humanitarian law, and the risk of AI-driven escalation in crisis situations.

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

Fourth, the role of AI in hybrid threats, including disinformation campaigns, election interference, and information warfare represents a growing challenge to democratic societies and regional stability. Several papers examine how AI amplifies these threats while also exploring how AI tools can be deployed defensively to detect and counter malicious information operations.

Fifth, ethical considerations pervade every aspect of AI deployment in security contexts. From algorithmic bias and surveillance concerns to questions of human dignity and autonomy, our contributors grapple with the fundamental ethical dilemmas that arise when powerful AI systems are applied to security decisions affecting human lives and societal well-being.

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

Looking forward, both special issues identify several critical research gaps and policy challenges that warrant continued scholarly and practical attention. These include the need for standardized risk assessment methodologies for AI in security applications, the development of international norms and treaties for autonomous weapons systems that balance humanitarian concerns with legitimate defense needs, mechanisms for preventing AI-driven escalation and miscalculation in crisis situations, frameworks for ensuring that AI advances contribute to rather than undermine global stability and human security, enhanced international cooperation on AI safety and security research, and educational initiatives to prepare the next generation of security professionals for an AI-enabled operational environment.

I wish to express my profound gratitude to all who contributed to making this special issue possible. First and foremost, to the authors who shared their cutting - edge research and insights, your scholarly excellence and commitment to addressing these critical challenges inspire us all. To the conference organizing committee and participants, whose intellectual engagement created the foundation for this publication. And to the editorial team and production staff of the Serbian Journal of Engineering Management, whose professionalism and dedication made this special issue a reality.

As we stand at the threshold of an era in which AI will increasingly shape security outcomes across multiple domains, the scholarship presented in this volume offers both sobering warnings and constructive pathways forward. The future of AI in security is not predetermined, it will be shaped by the choices we make today about governance, ethics, regulation, and international cooperation. I hope this collection will serve not only as a valuable scholarly resource but also as a catalyst for continued dialogue between academia, policy communities, technology sectors, and civil society on one of the defining challenges of our era.

The intersection of artificial intelligence and security will remain a critical area of inquiry for years to come. Two issues represent an important contribution to our understanding of these complex issues, but it is only one step in a longer journey toward ensuring that AI serves humanity's security needs while upholding our fundamental values of human dignity, justice, and peace.

Sincerely,

Prof. dr Katarina Štrbac

Reč urednice

Poštovani čitaoci,

Sa velikim zadovoljstvom predstavljam dva posebna izdanja Serbian Journal of Engineering Management, posvećena jednom od najznačajnijih izazova našeg vremena: vezi izmedju veštačke inteligencije i bezbednosti u 21. veku. Ova kolekcija nastala je iz 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. Ova dva posebna izdanja obrađuju kompleksne teme kroz interdisciplinarnu naučnu analizu koja prevazilazi računarske nauke, međunarodne odnose, etiku, pravo i inženjerski menadžment.

Posebna izdanja Serbian Journal of Engineering Management obrađuju 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 ova izdanja 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 bezbednosti 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 da se upravljanju AI ne može pristupiti kroz jednu kulturnu, političku ili ekonomsku prizmu.

Perspektiva Zapadnog Balkana, dobro zastupljena u ovom tomu, 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 se ističe 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 nosi značajne implikacije za globalnu stabilnost, strukture saveza, tehnološke standarde i buduću ravnotežu moći. Naši autori ispituju kako države mogu delovati kroz ovu dinamiku 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 regulacije za AI u bezbednosnim primenama ostaju fragmentisani i nedovoljno razvijeni. Autori I u ova dva izdanja analiziraju postojeće regulatorne pristupe na nacionalnom, regionalnom i međunarodnom nivou, identifikujući najbolje prakse dok naglašavaju kritične praznine koje zahtevaju hitnu pažnju kreatora politika i međunarodnih organizacija.

Četvrto, uloga AI u hibridnim pretnjama - uključujući kampanje dezinformacija, mešanje u izbore i informaciono ratovanje - predstavlja rastući izazov demokratskim društvima i regionalnoj stabilnosti. Nekoliko radova ispituje kako AI pojačava ove pretnje dok takođe istražuje kako AI alati mogu biti primenjeni odbrambeno za otkrivanje i suprotstavljanje zlonamernim informacionim 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.

Konačno, organizacioni i tehnički izazovi - uključujući sajber ranjivosti, zaštitu osetljivih informacija od AI, upravljanje talentima u organizacijama intenzivnim u AI, i integraciju AI u postojeće institucionalne structure - zahtevaju pažnju kako istraživača tako i praktičara.

Gledajući unapred, ova posebna izdanja identifikuje nekoliko kritičnih 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 humanitarne brige 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.

Želim da izrazim duboku zahvalnost svima koji su doprineli da ovo posebno izdanje postane moguće. Pre svega, autorima koji su podelili svoja najsavremenija istraživanja i uvide - vaša posvećenost rešavanju ovih kritičnih izazova nas sve inspiriše. Organizacionom odboru konferencije i učesnicima, čija je angažovanost stvorila osnovu za ovu publikaciju i uredničkom timu Serbian Journal of Engineering Management, čija je profesionalnost i posvećenost učinila da ovo posebno izdanje postane stvarnost.

Dok stojimo na pragu ere u kojoj će AI sve više oblikovati bezbednosne ishode u više oblasti, naučna analiza predstavljena u ova dva posebna izdanja 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 kritična oblast istraživanja u godinama koje dolaze. Ovo posebno izdanje predstavlja značajan doprinos našem razumevanju ovih kompleksnih pitanja, ali je samo jedan korak u dužem putovanju ka obezbedjenju 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: 004.8:17
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600001M

Etičke dileme i društveni izazovi: Ko će preuzeti odgovornost za zloupotrebu VI?

Ida Manton[1]

[1] Prague University of Economics and Bussines**,** ida.manton@gmail.com

Abstract: Članak se bavi fenomenom veštačke inteligencije, moralnim dilemama koje proizilaze iz njene široke upotrebe i neophodnosti regulisanja te upotrebe. Dok žurimo ka stvaranju potencijalno najopasnijeg agensa automatizovanog donošenja odluka, moramo postaviti pitanje: čija je odgovornost da kontroliše zloupotrebu VI i potencijalne štete koje će naneti našim društvima – države, tehnoloških korporacija ili pojedinca? Informacija je najvredniji resurs našeg vremena. Informacije su oduvek bile dragocene, ali načini na koje im pristupamo su se promenili, a sa tim i strukture i metode kojima se dele. Ljudi su oduvek bili vrednovani na osnovu svog znanja, na osnovu toga koliko dobro su mogli da koriste ono što su znali i koliko kreativnosti je moglo proizaći iz informacija koje su prikupili. Period koji je prethodio ovoj novoj realnosti vođenoj VI dao nam je uvid u poteškoće sa kojima naša nacionalna i međunarodna tela mogu da regulišu i donose zakone o upotrebi nepoznatih tehnologija.

Ključne reči: veštačka inteligencija, etika, odgovornost, regulacija, društveni uticaj, demokratija, upravljanje

Ethical Dilemmas and Social Challenges: Who Will Take Responsibility for AI Misuse?

Abstract: The article looks into the phenomenon of artificial intelligence, the moral dilemmas rising from its widespread use and the necessity for regulating that use. As we are rushing towards creating the potentially most dangerous agent of automated decision-making we need to ask the question: whose responsibility is it to control the misuse of AI and the potential damages it will inflict on our societies – the state, the tech corporations, or the individual? Information is the most valuable resource of our time. Information has always been precious, but the ways we access it have changed, and with that, so have the structures and methods by which it is shared. People have always been valued based on their knowledge, on how well they could use what they knew, and on how much creativity could stem from the information they gathered. The lead up to this new AI-driven reality has given us a preview of the difficulty with which our national and international bodies can regulate and legislate the use of unknown technologies.

Keywords: artificial intelligence, ethics, responsibility, regulation, social impact, democracy, governance

1. Ethical Dilemmas and Social Challenges: Who Will Take Responsibility for AI Misuse?

The text therefore explores if the Artificial intelligence is intelligent, whether it is a tool that will be useful for humanity or will make people subservient to machines, and also what dangers are preventable and what is an integral part of the system, which we have to accept with making AI fundamental part of our lives. A part of the text will deal with the effects AI has on education, social dynamics that are already showing signs of pathologies which AI will only exacerbate, as well as the effect it will have on furthering polarization and creating echo-chambers that amplify dangerous group-think and changes in social and political interactions.

And we often do it before we even see the full potential of the well-intended invention. We did that with almost all political models and we saw very wrong deviations, which compellingly became arguments for witch-hunting rather than advanced thinking about how to repair and upgrade.

2. The dangers of advancing AI

Introducing AI in our everyday life, on our personal computers, our social media pages and even every single message we send to another human, poses a huge ethical dilemma. It makes me think that the Universe is spitting at us a scrambled response carrying some meaning along the lines of "be careful what you wish for" knowing humanity has a tendency to overdo, to exaggerate and to turn sour even the most helpful tools and ideas. We have turned cure into poison before. And even just a naturally occurring bacteria, like Anthrax, we have turned into a bioweapon. In Chinese alchemy, the elixirs prepared to prolong life, were found to be the cause of death to a few Emperors and noblemen because of the heavy metals they contained. But we have also experimented and turned the notorious "poison of the Kings", the Arsenic, into a cure for syphilis, and mold into antibiotic. So, the core determinant of whether something will be used as an antidote or a poison, as helpful or detrimental, even devastating, is the user's intent which determines the impact. This is how our story with the use of AI will untangle. The difference is that all the other inventions, be those cures or poisons, were tools, just products, vehicles, or trinkets. AI is the first invention we humans created and are in a rush to amplify it to a level where it surpasses us in, what we thought to be the best of all spices in – thinking! No other human form is a more advanced thinker than the human being and now we seem not to trust ourselves with this skill as much as we trust AI technologies, that already proved to have a better and wider scope of knowledge assembling, analyzing in mere seconds and producing content better than any of us, especially when confronted with time and insurmountable quantities of content. But does that make the Artificial Intelligence intelligent and even more, does it make it smarter than us? It sure is created to be, to outmatch us in our human intelligence. The promise and hope, is that AI can invent and create things that are beyond our imagination, but at the same time this unpredictability leaves us humans in a vulnerable position where we cannot predict what AI will come up with or do. The threat is that how it develops is beyond our control and therefore we cannot guarantee that AI will comply with our instructions because it has a "mind" of its own. Yuval Noah Harari intriguingly asked recently at the World Economic Forum 2026 in Davos: Will AI challenge our (human) supremacy in thinking? (Harari, 2026)

Additionally, how do we entrust AI with truth, shared moral concepts and integrity and other more fluid areas that enjoy a huge range of positioning among humans? As creators of AI, as digital content creators, we all provide material that can then be used by algorithms for a curated AI content, so anything becomes potentially repeated and regurgitated content, regardless of whether it is truth or a lie, disinformation or fact. "This means AI has no understanding. No consciousness. No knowledge in any real, human sense. Just pure probability-driven, engineered brilliance — nothing more, and nothing less".

A less obvious potential problem is based on the hypothesis that even if all the creators of AI content had genuinely good and truthful intention, the system would not be absolutely truthful, as the whole is never just a sum of the particles because the interactions, the synergies create new qualities. Science has showed this to us. We have explored "emergence" as a concept and have identified the element of unpredictability, or as Robert Musil calls what he explores in his novels, the "imaginary unit-i ". That is if everything uploaded in the AI systems was morally unquestionable, which is of course an absolute and improbable situation, which does not reflect our multilayered, cross-cultural reality. My truth can be someone else's lie. Still an intellectually stimulating though comes to mind - would even in such absolute circumstances AI technologies be truthful, can they employ restraint and avoid escalation and polarization, can they operate benevolently, and even more, can they be trained to know how to avoid human suffering of any kind. So far, or shall I say already, we have seen the opposite. What we know from the recent military operations in Gaza and Ukraine, is that the AI-DSS (AI decision support systems) can create "kill lists", identifying and eliminating targets with facial recognition technologies (FRT) in complete disregard of IHL (international humanitarian law).

This leaves no space for hope that we will be spared of the AI's dark side. And if that is a starting point in exploring how to prevent the dangers approaching humanity with hypersonic speed, we have to start creating institutions, SOPs, experts and legislation that will slow down the unhinged use and creation of AI models that violate at least the already existing laws, like data privacy, to name just one in addition to the whole body of IHL, which I mentioned earlier. And while we do that on communal, national and international level, we have to broaden the discussion to understand AI. Here I offer just a few thoughts for such necessary discussion.

Both words describing this phenomenon are challenging. "Artificial" refers to a human creation, an artifact. In Aristotelian terms, an artifact is something that exists by craft and has its origin in the craftsperson in the form of the thing as it exists in the mind of the maker. When it comes to "intelligence", if it is narrowly understood as information gathering, we have already lost the battle to AI. Luckily, intelligence has many features and quite significant part of it is emotional, kinetic, abstract and artistic - fields that require our human existence and values our neuro-divergence, our different points of view and our need for dialogue to find common ground, mutual

understanding, morality and even a worldview. In all of these aspects of modern human life, technologies already have massive roles – from how we interact with other humans, to our office life, diplomacy, film industry, video gaming, performing arts, wellbeing and heath, all handled through apps and computer programmes that measure and collect our data. The eventual goal of this information gathering exercise is that the technology we use will predict what we want and need, which services we require and soon AI will be able to provide those without humans interfering in that process. Automated, however, still does not mean artificial intelligence, and our experiences so far show us how stupid technology is even though we name it "smart". We all catch ourselves hating the logjams created by machines more than valuing its usefulness. Too many times, in just a single day. But AI is different from voice mail machines and a milk foam maker. AI is not just producing or providing the service. AI is ordering and delivering, which then excludes the will and the intentions of the people, of its creator.

The key question therefore is who are the humans behind the creation of a world based on AI. Can they be trusted? What are their motives and how do they see the endgame? How is this new "gadget" affecting power, governance, politics, social contracts, security and how can it disrupt the already ruptured tissues?

Let us first review what we know about the behavior and interests of the tech giants from this relatively short phase of our digital life on this planet. With the "democratization" of the media that happened under minimal regulation, content was being uploaded without restrictions on sharing or mindfulness of intellectual property regulations already in place. A few IT guys earn money for something that the journalist who wrote the article is paid far less for. Almost 80% of the income went to Google and Facebook. Google has already been fined for violating competition laws in the European Union. A few years ago, the US Department of Justice (n.d.) filed a lawsuit against Google for monopolizing the online advertising market, accusing it of using "anticompetitive and illegal methods to eliminate or drastically reduce any threat to its dominance over the technologies used for digital advertising". Facebook was the subject of criticism, but also of lawsuits related to the management of the information it possesses, the way it transmits news, identifies users and recognizes their faces, protects privacy, and creates an addiction that is often compared to drug addiction. The Cambridge Analytica scandal, detailed and explained in the documentary "The Great Hack" (The Great Hack, 2019), shows how personal data collected to build psychological profiles (through applications such as "This is your Digital Life"), revealed a very dangerous side of digital tools — that they are abused to influence electoral processes, as was the case with the 2016 US Presidential Election and the Brexit referendum. With this track record, almost the same actors leading the tech giants and investing outrageously large sums of money in developing AI systems, being sceptic about the benevolence of the AI is the least we can do.

By now we have access to published research and accounts by whistleblowers and reliable investigative journalists analyzing and exposing the monetization strategies employed by tech billionaires to generate profit by promoting polarization, misogyny and violence of various kinds. They expose the financial dynamics at play, key actors, the narratives they deploy, and the tactics they use to operate as a scalable business model that often develops to level of capturing states. One such eye-opening insight was provided by GNET, and it discusses Engagement Farming and the Tactics Behind Incendiary Online Content (Global Network on Extremism and Technology, n.d.). The text explains that: "Engagement farming exploits the political economy of social media platforms, specifically algorithms that prioritise emotional responses. Adversarial actors intentionally curate inflammatory, misogynistic content to trigger indignation and counter-speech". This being the case with online content, should already raise all the flags for introducing AI, which uses all of this trolled, biased and warped content, as there are no restrictions to what data the AI creators are using to build up their models.

Additionally, the potential for misuse of combined data which leads to creation of digital IDs that are then in the hands of people like Peter Thiel, Elon Musk, David Sacks and other "Tech bros" and billionaires, was predictable. DOGE and Palantir infiltrated the US Government and the White House, harvested classified information and documents from agencies and departments they dismantled (like USAID) have awarded themselves enormous federal funds and made even more money than the investments they made into the political campaigns to bring Trump into power. In March, Trump signed an executive order requiring all agencies and departments of the federal government to share data under the pretense that he was "Stopping Waste, Fraud, and Abuse by Eliminating Information Silos".

To get the job done, Trump chose Palantir Technologies. Palantir, "sells an AI-based platform that allows its users – among them, military and law enforcement agencies – to analyze personal data, including social media profiles, personal information and physical characteristics. These are used to identify and surveil individuals". Palantir, was also reported to have built the deportation software for Trump. So, the symbolism behind the name became chillingly real and so did the role Palantir and similar tech companies have in the war on democracy and liberal values.

Some of the features amplified by AI were already available in the pre-AI phase and were concerning because they were used to incite violence, manipulate elections, intimidate political opponents etc. On top of all the information collection, AI offers upgraded features for manipulating information and immediate action - detects patterns, analyzes data, provides options for strategic approaches and much more. In the hands of people who want to run the world unchallenged, AI is dangerous and arguably devastating invention. The reveal of the work Palantir has conducted for ICE (US Immigration and Customs Enforcement) and its investigative branch HIS (Homeland Security Investigations), showed that there is capacity to track these developments by civil society groups. However, despite the work done by an immigrant legal rights group, called Just Futures Law (Just Futures Law, n.d.) the second Trump administration awarded a $30 million contract to Palantir to build the government a new platform called ImmigrationOS that will "service Ice branches beyond HSI, and aims to "streamline" the identification and deportation of immigrants. Revealing and publicizing the deals and how they affect the American society has mobilized more people in variety of local and national organizations working on digital rights, regular protests and campaigns, like No Tech for Ice. They demand responsible government, safe digital space and data privacy, all of which are being traded and sold as part of multimillion contracts between the Government agencies and the tech billionaires who now have unrestricted access to data and no legal or moral restriction in how to use it.

3. Most Often Discussed Negative Effects of AI

Among the concerns across academic analysis, policy debates, and regulatory discussions, the most relevant for this paper can be grouped as those pertaining to safety, economy and relationships. Among those falling under safety, including security and societal risks, most discussed are misinformation and manipulation (including deepfakes, synthetic media, automated propaganda, political micro-targeting, and election interference), cybersecurity threats (automated hacking and malware generation, exploitation of vulnerabilities accelerated by AI tools, AI-enabled cyberattacks on critical infrastructure), military misuse (like lethal autonomous weapon systems - LAWS), privacy erosion (mass surveillance through biometric tracking, Inference without consent) as well as bias, discrimination, and deficient outcomes due to biased algorithms in hiring, policing, healthcare etc. The economic risks primarily talk about job replacement and re-skilling challenges. On a broader societal level, there is fear of loss of human oversight and control, unpredictability, hallucinations, memory gaps, black box logic, as well as large-scale societal destabilization.

Maybe the most dangerous aspects of AI that we need to regulate is the human responsibility to avoid abdicating our traditional superpower – to think? And this is a responsibility that we will have to nurture as individuals, as educators, as parents, as societies and as humanity. The question still remains how and who will take a lead on each step of these pathways. Also, will it suffice if we have responsible individuals and irresponsible network of technofeudalists? As Varoufakis warns in his book "Technofeudalism: What Killed Capitalism", the traditional capitalism has been replaced by a new system where big tech companies act as modern-day feudal lords, extracting "rent" from users through their platforms. Big tech firms like Google, Amazon, and Facebook have become the new ruling class, extracting value not through traditional profit from selling goods, but through "rents" from their platforms (Varoufakis, 2023). As platforms that live off of engagement and attracting (maybe more correct terminology would be luring) visitors to their spaces to interact, shop and do almost everything through them – from ordering food, to engaging with your doctor, presenting your work etc., it is unlikely that they will add ethical considerations if their profits are decreasing. Unless restricted by law-makers, they will have no incentives to be an ethics champion.

In the field where I have specialized over the years—international negotiations—there have already been many hopes and many warnings about the role AI might have. A Harvard-led research paper suggested that more research will have to be done in order to "define the complementary roles of AI and human negotiators, ensuring that AI supports, rather than undermines, the complexities of negotiation practice". And this is where we stand when it comes to "high-stakes workplaces where human judgment, relationship-building, and adaptability are essential—such as crisis response, social work, and labor mediation—AI should be designed to support workers' expertise rather than automate complex, context-sensitive decisions". The main concerns expressed by negotiators, as part of this research were confidentiality and hallucinations/bias.

In addition to all of this, AI is likely to reduce creativity if we rely on its intellectual capacity without entering the process prepared and with expectation that AI will be a control function, not the content creator, especially in areas where we are not experts. So, whether AI turns into an assistive technology, efficient expert that will interact with humans or an uncontrollable foe, will depend on what we make of it, how lazy we become and what will our societies, competition and other circumstances force us to do to stay in the game.

4. Ethical dilemmas of using AI in political and social contexts

The rapid expansion of artificial intelligence raises profound ethical dilemmas that cut across power, autonomy, democracy, inequality, and human dignity. Thinkers such as Harari (2026), Zuboff (2019), Noble (2018), and Eubanks (2018), warn that AI is not merely a technical innovation but a political force capable of reshaping societies at their core.

One major category of concern relates to power and human agency. Harari argues that AI's growing ability to "hack" human psychology threatens the foundations of democracy. By analyzing biometric and behavioral data, AI systems can influence emotions, beliefs, and political preferences, blurring the line between legitimate persuasion and unethical manipulation. Personalized political messaging, while efficient, risks undermining individual autonomy by nudging citizens toward decisions they do not fully understand or consciously choose. As people increasingly defer judgment to algorithms, human agency itself is weakened. This creates an ethical dilemma: does increased efficiency justify outsourcing understanding, decision-making, and responsibility to machines, or does this lead to a form of "digital serfdom" in which humans are optimized rather than empowered?

A second set of dilemmas concerns knowledge, truth, and democratic processes. Democracy depends on shared facts and meaningful deliberation. Without them, political disagreement turns into manipulation and confusion. This raises difficult questions about whether AI-generated political content should be restricted, even if such restrictions limit freedom of speech. Closely related is the problem of algorithmic polarization. Engagement-driven systems tend to amplify outrage, filter information, and exaggerate extremes, pushing societies toward radicalization. Also, massive information asymmetries are emerging. Governments and tech corporations increasingly possess far more behavioral data than ordinary citizens, shifting political authority and changing decision-making practices.

AI also intensifies dilemmas related to inequality and social justice. Extensive research has shown that algorithms can encode and amplify existing racial, gender, and socioeconomic biases (Eubanks, 2018; Noble, 2018). Decisions shaped by biased data may appear neutral and objective while masking structural discrimination, making accountability harder to establish. Additionally, unequal access to AI technologies risks deepening global and domestic inequalities. Wealthy states, corporations, and political actors are better positioned to exploit advanced AI for influence and control, raising doubts about whether AI truly democratizes knowledge or instead entrenches elite power. Also, based on the content it is fed, the languages it recognizes and adopts, it is very likely to have a racial, western and cultural bias. While employment and automation are often framed as economic issues, their political consequences are significant. Job displacement and social insecurity can undermine democratic participation, and economic exclusion can translate into political marginalization, among the many desired occurrences by authoritarian leaders.

Surveillance, control, and political authority form another critical area of concern. While surveillance is often justified in the name of security or efficiency, it raises difficult ethical questions about privacy, freedom, and consent. AI-powered monitoring risks transforming liberal democracies into data-driven totalitarian systems. Predictive policing and algorithmic risk assessments further complicate matters, as they frequently misrepresent marginalized communities and raise serious concerns about due process. If an algorithm predicts criminal behavior, should authorities act on that prediction, and if so, how? Even more extreme is the use of AI in autonomous weapons systems. Delegating decisions over life and death to non-conscious machines challenges foundational ideas of moral responsibility and accountability.

AI also affects identity and culture. Global AI systems often reflect Western, corporate, or dominant cultural assumptions, leading to concerns about cultural cognitive imperialism. This raises the question of whose values are embedded in the algorithms that increasingly shape global communication and culture. The rise of AI companions, therapy bots, and political chatbots introduces dilemmas about authenticity and human relationships. Simulated care, intimacy, or political solidarity may blur emotional boundaries and manipulate trust, raising ethical questions about whether such simulations are acceptable.

Finally, issues of governance, accountability, and moral responsibility cut across all these domains. Many AI systems operate as opaque "black boxes," making it difficult to understand how decisions are made or who is responsible when harm occurs. When an AI system causes damage, responsibility is fragmented among developers, deployers, users, politicians and regulators. Moreover, there is a growing tendency to treat AI as a technical fix for governance and bureaucracy. So far, the use has not been very sophisticated, but when that happens the databases from different agencies could come together into a larger system which can generate digital profiles of citizens. Additional dangers might include sidelining of democratic debate due to overreliance on AI

recommendations and shrinking of the space for collective decision-making, which could easily be delegated to AI to avoid public participation and to avoid the push backs from any opposition.

The hallucinations, biases and disconnected argumentation are going to be part of AI for a long time. Whether they will be fixed, fought or even challenged, will depend whether humanity will catch up with big tech and will demand a more responsible development of the systems or the race will justify the means and AI, like many other potentially useful inventions, will become a serious threat to the healthy society, human relationships and a threat to peace on the planet whose existence is challenged because AI's expansion already challenged basics like water, energy and mining.

5. AI, Power, and the Future of Democracy

Artificial intelligence is often framed as a technological revolution, but its deepest impacts are political and ethical. AI is not just another tool—it is the first tool capable of shaping the very human minds that created it. This is the heart of the dilemma: we are deploying a technology that can read our weaknesses, predict our choices, and influence our behavior faster than democratic institutions can respond.

Already now, AI systems generate realistic images, voices, and narratives that make truth negotiable and trust fragile. When citizens cannot distinguish fact from fabrication, democratic debate collapses. The result is not disagreement but disorientation—a political environment where the loudest algorithm, not the best argument, prevails. No society can sustain democratic legitimacy without a shared reality. But the ethical dilemmas extend beyond truth. AI amplifies existing inequalities by embedding the biases of the past into decisions about the future. Predictive policing, algorithmic hiring, and automated welfare systems can quietly punish the already marginalized, creating a digital caste system dressed up as efficiency. At the same time, concentration of data in the hands of a few governments and corporations threatens to undermine human agency (Zuboff, 2019). If algorithms know us better than we know ourselves, the space for genuine autonomy narrows.

Perhaps the most profound ethical question is whether democratic societies can preserve the dignity of human judgment in an age when machines can outperform us in analysis, persuasion, and prediction. The temptation to outsource political decisions to "neutral" algorithms is growing. But politics is not a math problem; it is a moral project. Handing our collective choices to opaque systems risks turning citizens into spectators, governed by forces they cannot see or challenge.

AI's promise is real, but so is the danger that it becomes a technology of domination—one that manipulates emotions, films populations, and concentrates power at unprecedented scale. Ethical governance must begin with protecting human agency and maintaining democratic control over the systems that increasingly shape our lives. The future of AI is not just about innovation. It is about who we become when our choices, beliefs, and relationships are filtered through machines that do not share our values, but can profoundly influence them. As part of the great power competition, the big question will be who will have a bigger cloud service and how its development will affect the electricity grids and expansion to emerging markets.

The question, therefore, is not whether AI will change society, but rather who gets to decide how and whether democracy survives the cast-creating cosmos.

6. AI and Education

Education is already a challenge and the transformation has brought both innovation and laziness.

AI can add a lot in education, but the worst damage we already observe is that in this very competitive field, where quantitative indicators decide which school one will go to or whether they will get a scholarship or not, students are not encouraged to learn how to study and the system replaces their curiosity with ingenuine products produced by AI.

So far, we have been defensive, but soon we will have to teach how to use AI responsibly, how to teach assisted by AI, not expecting AI to compensate for lack of substantive knowledge, teaching materials and lesson plans, how to evaluate students' work and making a distinction between AI generated content and genuine input, etc.

While AI may replace humans in some roles, "it also opens up new opportunities in sectors that demand complex decision making, emotional intelligence, and creative skills—attributes that AI cannot replicate. Understanding these trends is crucial for future workforce preparation". Education and training will need to adapt to help people transition to roles where human expertise remains irreplaceable.

7. How States Regulate Negative Effects of AI

Regulatory approaches to artificial intelligence vary across jurisdictions, but they generally cluster into several recognizable models.

Horizontal, comprehensive AI laws aim to regulate AI across sectors through overarching legal frameworks. The most prominent example is the European Union's AI Act. It introduces a risk-based system that categorizes AI uses as unacceptable, high-risk, limited-risk, or minimal-risk, with the strictest obligations applied to systems affecting biometric surveillance, critical infrastructure, and safety-sensitive applications. Enforcement is carried out through national supervisory authorities, coordinated at the EU level by the EU AI Office.

Similarly, the Council of Europe (2024) adopted the Framework Convention on AI in 2024, the first binding international treaty focused on AI and human rights. Notably, participation extended beyond EU member states to include countries such as the United States, the United Kingdom, and Canada.

A second model consists of sector-specific regulation, where AI is governed through existing legal frameworks rather than dedicated AI laws. Many states regulate AI indirectly via data protection regimes such as the EU's GDPR, California's CCPA/CPRA, and Brazil's LGPD, as well as through consumer protection law. Additional sectoral rules apply to areas like medical devices and healthcare AI, autonomous vehicles, and financial services, including algorithmic trading.

A third category focuses on AI safety and frontier model governance, particularly for advanced or general-purpose systems. Examples include the United States Executive Order on Safe, Secure, and Trustworthy AI issued in 2023 (United States Executive Order 14110, 2023), the establishment of the UK AI Safety Institute in the same year, and ongoing efforts in countries such as Japan, Singapore, and Canada to develop AI assurance, evaluation, and audit schemes.

Finally, national security and export control regimes play an increasingly important role in AI governance. The United States has imposed export controls on advanced semiconductor chips and, in some cases, model weights. The European Union applies dual-use regulations to surveillance and related technologies, while China has introduced rules governing generative AI providers, including requirements for synthetic media labeling.

8. International Bodies That Govern, Legislate, or Restrict AI Misuse

There is currently no single global authority governing artificial intelligence, but a range of international and regional organizations play important roles by creating binding treaties, soft-law instruments, and influential technical standards.

Some bodies are responsible for binding or quasi-binding instruments. The Council of Europe is particularly significant, having adopted the Framework Convention on AI in 2024, the first binding multilateral treaty dedicated to AI. The convention focuses on safeguarding human rights, democracy, and the rule of law, and is open to participation by non-European states. The European Union, its EU AI Act, alongside related legislation such as the Digital Services Act, the GDPR, and the Data Governance Act was a huge achievement, especially as these frameworks often have extraterritorial effect, applying to AI systems and services that reach EU citizens regardless of where providers are based. Within the United Nations system, no global AI treaty exists yet, but governance is emerging through multiple bodies. UNESCO's Recommendation on the Ethics of AI

adopted in 2021 by 193 states (UNESCO, 2021), provides a widely endorsed normative framework, while the UN Human Rights Council has issued non-binding resolutions addressing AI and human rights.

Technical and security dimensions are addressed by entities such as the International Telecommunication Union, which develops interoperability and technical standards; the UN Office of Counter-Terrorism, which focuses on preventing AI-enabled terrorism; and UNIDIR, which conducts research on AI in weapons systems and arms control. In addition, the Wassenaar Arrangement contributes indirectly to AI governance by imposing dual-use export controls on surveillance and intrusion software technologies that can include AI-enabled security tools.

Alongside these, a number of global multistakeholder and standard-setting bodies exert significant influence despite their non-binding nature. The OECD's AI Principles, adopted in 2019, are among the most influential soft-law instruments and have informed G7 initiatives, EU regulation, and national AI strategies. The G7's Hiroshima AI Process, developed between 2023 and 2024, produced a Code of Conduct for developers of advanced AI systems, with a strong emphasis on frontier model safety and transparency. Technical standards are further developed by ISO and IEC, which publish standards on AI safety, risk management, and auditing that are

frequently incorporated into national laws and regulatory frameworks. The Global Partnership on AI, with more than 44 member states, operates as an advisory and research-driven forum promoting responsible AI development and deployment. The partnership is guided by a Council, Plenary, and Steering Group, with support from the OECD Secretariat.

Given the focus of this publication, we should look into security and military-focused bodies that are increasingly engaged in AI governance. Within the UN Convention on Certain Conventional Weapons, states continue discussions on lethal autonomous weapons systems, and while no binding treaty has yet emerged, momentum toward regulation is growing. Similarly, the OSCE addresses AI-related risks across a range of areas, including media freedom, disinformation, cybersecurity, election integrity, and the human rights impacts of digital technologies. NATO has also developed an AI strategy for defense applications and adopted principles for the responsible use of AI in military contexts in 2021, shaping how AI is integrated into collective defense and security operations. Important to notice at the end is the difference between legislative and advisory documents. While we see increasing production of documents with more informative and advisory, even predictive character, we do not see the same pace with legislating and restricting development of AI or some coordinated action for minimizing economic shocks, electricity bills, water management and labor redistribution.

9. Conclusion

Regulating and legislating AI is often mistaken for obstructing its development and that causes fear of bureaucratizing or slowing down the development because those involved cannot afford to lose the race. The business stakes are too high, but so are the societal and anthropological stakes that change the basic features of our safety, security, peace, and ethical frameworks that guide us.

This article has examined artificial intelligence not simply as a technological innovation, but as a transformative force reshaping power, responsibility, democracy, and human agency. From its rapid and largely unregulated integration into everyday life to its deployment in governance, security, education, and warfare, AI exposes deep ethical dilemmas that existing political and legal systems are struggling to address. The discussion has shown that AI intensifies long-standing structural problems rather than replacing them. It amplifies inequalities by embedding bias into automated decision-making, concentrates power through control over data and infrastructure, and erodes democratic processes by manipulating information, accelerating polarization, and weakening shared reality. At the same time, surveillance technologies, predictive systems, and autonomous decision-making tools challenge fundamental principles of privacy, due process, and human dignity. These developments are not hypothetical. They are already shaping law enforcement, migration control, warfare, education, and political participation.

At the core of these dilemmas lies the risk of human abdication of judgment, responsibility, and critical thinking, in favor of systems optimized for efficiency rather than moral reasoning. While AI can support human expertise and creativity, treating it as a neutral authority or a substitute for political deliberation risks transforming democratic societies into technocratic or data-driven authoritarian systems. Regulation, therefore, should not be understood as an obstacle to innovation, but as a necessary condition for preserving autonomy, accountability, and trust.

The analysis of regulatory approaches and international governance frameworks demonstrates that while important steps have been taken—particularly within the European Union, the Council of Europe, and UNESCO—global AI governance remains fragmented, slow, and reactive. The speed of technological development far outpaces institutional capacity, and the geopolitical race for dominance often undermines ethical restraint. Without stronger coordination, enforceable standards, and meaningful public oversight, AI risks becoming a tool of domination rather than empowerment.

Ultimately, the challenge posed by artificial intelligence is not whether it will change society, but how and under whose control. The future of AI is inseparable from the future of democracy, social justice, and human dignity. Preserving these values requires resisting the temptation to outsource moral and political responsibility to machines and reaffirming the human capacity to think, judge, and decide collectively. AI may shape the tools of governance, but it must not be allowed to replace the ethical foundations on which free and just societies depend. That is something we, the people, will have to reenvisage and renegotiate after this rupture in the world order.

References

1. Council of Europe. (2024). Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. https://www.coe.int

2. Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.
3. European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.
4. Global Network on Extremism and Technology (GNET). (n.d.). Engagement farming and the tactics behind incendiary online content. https://gnet-research.org
5. Harari, Y. N. (2026, January). Address to the World Economic Forum. World Economic Forum Annual Meeting, Davos, Switzerland.
6. Just Futures Law. (n.d.). No Tech for ICE Campaign. https://www.justfutureslaw.org
7. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.
8. Organisation for Economic Co-operation and Development (OECD). (2019). OECD AI Principles. https://oecd.ai/en/ai-principles
9. The Great Hack [Documentary]. (2019). Netflix.
10. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org
11. United Nations Office for Disarmament Affairs (UNIDIR). (n.d.). Research on AI in weapons systems and arms control. https://unidir.org
12. United States Department of Justice. (n.d.). Lawsuit against Google for monopolizing online advertising. https://www.justice.gov
13. United States Executive Order 14110. (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Federal Register.
14. Varoufakis, Y. (2023). Technofeudalism: What killed capitalism. Bodley Head.
15. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:316.72]:17
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600010M

Interkulturalna veštačka inteligencija: pomirenje etičkog univerzalizma i kulturne raznolikosti

Ernesta Molotokienė[1]

[1]Klaipeda University, Lithuania, ernesta.molotokiene@ku.lt

Apstrakt: Članak analizira koncept interkulturalne veštačke inteligencije, koji omogućava definisanje univerzalnih etičkih principa na osnovu kojih se mogu donositi interkulturalne odluke i postići sporazumi o razvoju, primeni, upravljanju i korišćenju digitalnih tehnologija. Interkulturalna veštačka inteligencija predstavlja jedan od najnovijih naučnih projekata usmerenih na razmatranje širokog spektra etičkih pitanja koja proističu iz uticaja tehnologija veštačke inteligencije na ljudsku svest, društva i kulture, iz multidisciplinarne perspektive. Etički univerzalizam naglašava zajedničke ljudske vrednosti, kao što su pravda, odgovornost i poštovanje života, dok kulturna raznolikost ističe moralne norme i tradicije specifične za pojedine zajednice. Stoga se u članku tvrdi da različite kulture ne postižu saglasnost oko zajedničkih univerzalnih etičkih smernica za veštačku inteligenciju, jer su one zasnovane na jedinstvenim pogledima na svet i vrednosnim sistemima, kao i zbog nepostojanja univerzalno prihvaćenog, epistemološki pouzdanog načina za rešavanje vrednosnih neslaganja. Članak razotkriva glavne teorijske pretpostavke na kojima se zasniva interkulturalna veštačka inteligencija, a koje omogućavaju stvaranje i primenu zajedničkog sistema univerzalnih etičkih principa koji regulišu razvoj digitalnih tehnologija u različitim regionima sveta i kulturama.

Ključne reči: Interkulturalna veštačka inteligencija, etički univerzalizam, kulturna raznolikost, etika AI, moralni pluralizam.

Intercultural Artificial Intelligence: Reconciling Ethical Universalism and Cultural Diversity

Abstract: The article analyzes the concept of intercultural artificial intelligence, which makes it possible to define universal ethical principles on the basis of which intercultural decisions and agreements on the development, deployment, governance, and use of digital technologies can be made. Intercultural artificial intelligence is one of the newest scientific projects aimed at examining wide-ranging ethical issues arising from the impact of artificial intelligence technologies on human consciousness, societies, and cultures from a multidisciplinary perspective. Ethical universalism emphasizes shared human values such as justice, accountability, and respect for life, whereas cultural diversity highlights the moral norms and traditions specific to particular communities. Therefore, the article argues that different cultures do not agree on common universal AI ethical guidelines because these are grounded in unique worldviews and value systems, and because there is no universally accepted, epistemically reliable way to resolve value-based disagreements. The article reveals the main theoretical assumptions underlying intercultural AI, which enable the creation and application of a shared system of universal ethical principles regulating the development of digital technologies across different regions of the world and cultures.

Keywords: Intercultural Artificial Intelligence, Ethical Universalism, Cultural Diversity, AI Ethics, Moral Pluralism.

1. Introduction

Digital technologies and artificial intelligence are developed within specific cultural contexts, which have a significant impact not only on the users of AI technologies but also on the development of AI itself (Brynjolfsson & McAfee, 2014). Therefore, increasing attention is being paid to intercultural digital ethics and to questions of how AI systems should be designed and deployed in light of their potential global impact on crucial ethical values such as privacy and dignity. Intercultural cooperation is essential for enabling ethical processes of AI development, deployment, advancement, governance, and use. First and foremost, intercultural cooperation aims to ensure that AI is developed, used, and governed in ways that are beneficial to society (Jobin et al., 2019).

Intercultural cooperation in the field of AI development is important for several reasons. First, such cooperation creates conditions for ensuring more balanced AI development across cultures and makes it possible to anticipate, diagnose, and effectively regulate obstacles arising in the development or adoption of digital technologies

Second, intercultural cooperation enables researchers around the world to share experiences, resources, and best practices. This facilitates faster progress both in the development of digital technologies and in the management of ethically problematic situations.

Third, in the absence of intercultural cooperation, there is a risk that commercial ecosystems with greater competitive advantages may underinvest in the safe, ethical, and socially beneficial development of digital technologies (Askell et al., 2019; Ying, 2019).

Finally, ethical international cooperation in developing intercultural AI is also important for more practical reasons, such as ensuring that AI applications, for example, those used in major search engines or autonomous vehicles can successfully interact with other technologies across different legal and ethical regulatory environments in various regions (Cihon, 2019).

Digital technologies and artificial intelligence systems may have different impacts when deployed in different cultural regions, where distinct governance approaches may be required (Hagerty & Rubinov, 2019). Nevertheless, it is evident that intercultural cooperation is indispensable in certain aspects of digital technology development and governance. For example, some potential military uses of artificial intelligence technologies may violate fundamental human rights (Asaro, 2012). In the absence of international agreements and standards, the application of AI technologies in the military industry may have destabilizing effects. International agreements are also necessary when AI technologies are developed in one region but deployed or used in another. A major obstacle to building trust in digital technology development between Eastern and Western cultures lies in worldview differences rooted in distinct value systems, which ultimately lead to differing: potentially conflicting views on what constitutes ethical development, use, and governance of digital technologies. By acknowledging existing multicultural differences, it is possible to identify certain cultural contexts that may serve as a sufficient basis for intercultural dialogue. This represents one of the most important initial steps toward addressing the fundamental challenges in the field of intercultural AI ethics.

2. Methodology

This study adopts a qualitative, theoretical, and comparative research design to explore how ethical universalism and cultural diversity can be reconciled in the development and governance of artificial intelligence (AI). The methodological framework is interdisciplinary, integrating philosophical, ethical, sociological, and technological perspectives to conceptualize the notion of intercultural artificial intelligence.

The research is based on a systematic literature review. The literature review aimed to identify major theoretical trends, ethical frameworks, and cross-cultural approaches relevant to AI ethics and governance.

In the second stage, a comparative analysis was conducted to examine how different cultural traditions: Western, East Asian, Islamic, and African ethical paradigms interpret moral principles applicable to AI decision-making. This approach allowed the study to assess the extent to which universal ethical models can be adapted across diverse sociocultural contexts.

Furthermore, a normative analysis was applied, grounded in moral philosophy and conceptual synthesis. This analytical method focuses on formulating a model of intercultural ethical pluralism, which combines universal moral principles (such as justice, accountability, and human dignity) with culturally responsive ethical practices.

Although no primary empirical data were collected, the study relies on secondary data analysis and conceptual reasoning to establish a coherent theoretical foundation for the governance of globally responsible AI. Finally, the credibility and validity of the research are ensured through triangulation of sources, integrating insights from multiple disciplines and cross-referencing international frameworks. This methodological approach enables a balanced evaluation of both the philosophical and practical dimensions of intercultural AI ethics.

3. Cultural Diversity and Contextual Ethics

The development of intercultural AI technologies encompasses a broad field of ethical principles and practices across different societies, historical periods, and philosophical traditions, which strongly influence the social adoption and adaptation of digital technologies. However, as Thomas Taro Lennerfors and Kiyoshi Murata (2021) emphasize, discussions on the adoption of intercultural AI often lack a clear understanding of culture and an

analysis of cultural relativism. In recent decades, the rapidly emerging field of intercultural AI research has shown a pronounced interest in cultural differences, referred to as a "culture matter" (Ess, 2017). One of the most common challenges at the early stages of forming the problematic field of intercultural AI has been the issues raised by ethnocentrism and cultural diversity (Ess, 2017). Recently, due to intensifying global conflicts arising from differing value systems, worldviews, and religious beliefs, a comprehensive reflection on the concept of culture in the context of intercultural AI development has become increasingly necessary (Palm, 2016). Theories of culture applied to AI development have been extensively elaborated by Edward T. Hall and Geert Hofstede, who argued that people belonging to a particular country share common characteristics, as if they possess a collective "algorithm" in their consciousness that distinguishes them from others (Hofstede & Minkov, 2010). Such cultural parameters are essentialist, as people within a specific culture share certain fundamental ways of being, thinking, and acting that set them apart from others.

In the context of intercultural AI development, there are also more marginal, non-essentialist cultural parameters, in which culture is not tied to an "essence," emphasizing the spontaneity, creativity, and diversity of cultures (Dahl, 2014; Langstedt, 2018). These cultural concept parameters represent homogeneous national cultures encompassing multiple cultural expressions and social practices. According to Bielby, even within the same region, there may be a wide variety of religions, subcultures, and philosophical systems (Bielby, 2015). Concepts of individualism and collectivism often differ drastically across cultures, for example, when comparing Japan and Western cultures (Westwood, 2004). This understanding of culture does not account for the hybridization processes occurring due to the impact of technologies, particularly the Internet.

Non-essentialist concepts of culture provide a far more favorable opportunity for intercultural dialogue and creative solutions regarding the normative grounding of intercultural AI. The non-essentialist perspective focuses not on attempting to preserve and defend stagnant cultural identities, but on the ways and reasons people use certain cultural concepts, how they interpret them, and which interpretations are temporarily prioritized over others and why. Therefore, one of the most important questions is how culture is created and understood at the micro level across different perspectives and relationships.

Several decades ago, Samuel P. Huntington formulated the idea of the "clash of civilizations," according to which cultural and religious differences worldwide become a primary factor in global conflicts (Huntington, 1996). Most criticisms of Huntington's idea are based on the belief that religious clashes and conflicts are not the most significant factors in global processes. Of course, cultural and religious differences and the disputes arising from them worldwide are undeniable facts, but there are many additional factors and conditions influencing conflicts between civilizations. Another line of criticism is directed at Huntington's position of defining and categorizing cultures and civilizations according to geographic zones. It is evident that, following geographic definitions, one can discuss the influence of different regions on each other in the context of globalization; however, more fundamental processes occur within these regions themselves. In other words, complex internal conflicts take place within specific geographic zones, leading to the conclusion that regions not only affect each other but also constantly change and transform internally. Huntington established the idea that cultural differences are highly significant for understanding the contemporary world.

Due to the impact of digital technologies, global conflicts have become more complex than ever before. Today, information mediated by digital technologies can spread at the speed of light to virtually any location in the world, making intercultural conflicts arising from differences in values, worldviews, and traditions inevitable. At the epicenter of these conflicts lies the unequal distribution of information control or power: information is typically generated and managed by the technologically most advanced countries, which, along with technology dissemination to other cultures, transmit corresponding "cultural codes" (value and worldview systems) that may not be favorably received by other cultures functioning under different value systems and worldview parameters.

4. Ethical Universalism vs. Moral Pluralism

Ethical universalism and moral pluralism represent two contrasting frameworks in the discussion of artificial intelligence (AI) ethics. Ethical universalism posits that certain moral principles are inherently valid for all humans, regardless of culture, geography, or historical context. This approach emphasizes shared values such as fairness, accountability, human dignity, and the protection of life (Korsgaard, 2018; Floridi & Cowls, 2020). Universalist principles provide a clear benchmark for AI development, offering guidance in high-stakes scenarios such as autonomous vehicles, algorithmic justice systems, and healthcare decision-making.

Moral pluralism, by contrast, recognizes the existence of multiple valid moral frameworks across societies. It argues that ethical reasoning is context-dependent and shaped by cultural, social, and historical factors (Wong, 2009). Pluralism highlights that what constitutes ethical behavior in one culture may not align with norms in

another. For instance, in collectivist societies, ethical considerations often prioritize community welfare over individual rights, whereas in individualistic societies, autonomy and personal freedom may be paramount (Hofstede, Hofstede, & Minkov, 2010). AI systems that ignore these variations risk making decisions that, while technically ethical according to universal standards, may be perceived as morally unacceptable in specific cultural contexts (Jobin, Ienca, & Vayena, 2019).

The challenges of creating a universal ethics applicable across different cultures and traditions are longstanding: both ancient Eastern and Western cultures developed often very sophisticated methods for resolving apparent social tensions between consensus and difference. At the same time, the solutions proposed by ancient Western and Eastern cultural traditions are, in fact, quite similar in several fundamental respects. Plato, Aristotle, and later Aquinas responded to this complex demand in at least two essential ways. Plato develops an approach that Ess describes as "interpretive pluralism" (Ess, 2006). Based on this Platonic perspective, presented in The Republic, it can be argued that there is a possibility of synthesizing shared ethical norms with differing viewpoints, acknowledging that diverse ethical practices can be understood as interpretations of common ethical norms and their different applications. Such differences in interpretation do not necessarily imply, as ethical relativists claim, the absence of universally recognized ethical norms or values; on the contrary, these differences may simply indicate that a particular norm or value is applied or understood in specific ways, as required by the context in which a particular tradition, cultural norms, and practices have developed. Ethical universalism based on pluralism can be found in various religious and philosophical traditions, for example, in the Islamic worldview (Eickelman, 2003) and in Confucian philosophy (Chan, 2008).

Aristotle, in his Metaphysics, draws on Plato's insights regarding the significance of linguistic ambiguities. Ambiguous words represent a linguistic middle ground between homogeneous unification (requiring a term to have one and only one meaning) and pure ambiguity (where a single term can have multiple, entirely unrelated meanings). In contrast, "pros hen" or "focal" ambiguities are terms that have different meanings but are simultaneously related to one another, as both refer to a common or central concept that grounds the meaning of each (Aristotle, 1003b2-4; 1060b37-1061a7). Thus, semantic differences are connected through a single overarching concept. Aristotle links the ability to reflect on these linguistic ambiguities as semantic differences or interpretations with a certain type of practical reasoning - *phronesis*. Just as we can recognize and appropriately understand concepts with different but related meanings, phronesis allows us to act according to a general principle, functioning as a kind of universal ethical analogue between two different decisions. This universal application of phronesis makes it possible to understand ambiguities with different semantic contexts in diverse ways while maintaining a universal perspective. Aristotle's position thus opens the possibility for ethical universalism.

Phronesis primarily refers to the capacity to make ethical decisions in specific and complex problematic contexts, as well as the ability to revise previous decisions when confronted with new information. In other words, phronesis is a model of self-regulation that enables the discernment of which general principles can be applied to a particular case within a given context and how they should be interpreted, taking into account the potential diversity of meanings.

In dominant Western ethical discourses to date, there has been an evident promotion of hegemonic Western values as universal and suitable for all cultures, often ignoring and marginalizing local value systems. Intercultural dialogue is based on the crucial premise that the global value system is not monolithic or hegemonic. Ethical universalism should not be understood as a practice that "reduces" cultural differences by imposing a single, unilateral system as a universal standard. The question arises whether, given the pluralistic nature of societies, it is possible at all to rely on any universal value assumptions. Conversely, in seeking to preserve value diversity, does the threat of cultural relativism emerge? The premises of cultural relativism eliminate any possibility of applying a universal method. Ethical universalism grounded in pluralism, however, allows for the preservation of value diversity across cultures while simultaneously recognizing the existence of certain foundational ethical principles, norms, values, and assumptions.

Although ethical pluralism is based on the premise that moral values, norms, ideals, duties, and virtues are inherently diverse, ethical universalism grounded in this pluralism can correlate with different value systems and their interpretations and applications in practice, while preserving distinct cultural traditions and practices. Ethical universalism is inseparable from phronesis, or practical wisdom, which is necessary for negotiation in various ethical and political contexts to achieve sustainable prosperity, human well-being, and harmony in multicultural societies. The intercultural AI project highlights the importance of a gradual transition from pragmatic shared economic interests to ethical universalism, which provides a basis for determining the appropriate ethical course in specific problematic, often radically different, contexts.

5. Ethical Universalism in Artificial Intelligence

We live in the age of AI, where societies develop mediated by digital information technologies. Digital technologies have significantly accelerated the movement of thoughts and ideas and have prompted inevitable clashes of value systems. As a result, intercultural value systems, dialogue, and the search for universal ethical principles are becoming increasingly important areas requiring global engagement. In the context of developing intercultural AI, key questions arise: which universal values can serve the essential goals of digital technology development while aligning with the beliefs and expectations of people representing different cultures? Is it possible to create intercultural AI technology that ensures the existence of universal ethical values without ignoring cultural and worldview differences?

It is impossible to ignore the substantial contrasts between the fundamental ethical assumptions underlying Western digital ethics and those forming the basis of classical Eastern ethical theories, such as Confucianism, Buddhism, and various local traditions and cultures. Unlike the Western mindset, which often grants the individual a central role as ultimate reality, in many Eastern cultural traditions, a person is understood as part of a community, whose identity is defined through relationships with other community members. Traditional African worldviews are based on the concept of ubuntu, in which a person is seen as a social and constantly self-developing being whose character evolves through relationships with others (Paterson, 2007). This concept of humans as relational beings correlates with Confucian philosophy (Ames & Rosemont, 199). These fundamental differences shape our identity as members and participants of culture. It is undeniable that individuals and cultures have an essential right to their own identity (Ess, 2006), meaning that it is necessary to respect and nurture cultural differences that define our personality and cultural identity. Nonetheless, in order to develop intercultural AI technologies, it is crucial to search for an intercultural foundation based on basic universal ethical values.

The pluralistic nature of contemporary societies enables the discourse of intercultural AI ethics. The intensity of AI technology development is evident worldwide, meaning that these technologies are being used by an increasing number of people in diverse cultural contexts. It is undeniable that to avoid a homogeneous global culture, grounded in minimal, pragmatically economic interests oriented toward efficient consumption, a universal ethics is required that preserves cultural differences. There are irreducible cultural differences defining various cultures and identities that remain resilient to hybridization processes, making it practically very difficult to develop and operationalize intercultural AI ethics.

In the context of global processes, it is crucial to consider whether the ethical standards that have predominated so far can be applied to different cultures. Ess (2006) proposed a concept of ethical universalism grounded in interpretive pluralism, rejecting ethical dogmatism and ethical relativism as inappropriate perspectives for underpinning intercultural AI. Ethical and epistemological relativism is based on the assumption that the diversity of perspectives and differences implies a lack of a single truth or value system, while ethical dogmatism relies on the firm belief that only one ethical system is correct, dismissing all other values, worldviews, epistemological methods, and arguments as wrong. Historically, the concept of ethical universalism developed under the influence of increasing tolerance for different value systems and religious beliefs, as well as recognition of the value of cultural diversity (Sartori, 1997).

Forms of universalism vary according to different Eastern and Western philosophical traditions, with a detailed analysis provided by Charles Ess (2006). Based on this analysis, we can identify the most significant current forms of ethical universalism grounded in pluralism in Eastern and Western cultures:

Modus vivendi pluralism recognizes existing differences between cultures and individuals but rejects a common value basis across cultures and perspectives, making this type of pluralism inseparable from ongoing value conflicts and worldview confrontations (Ess, 2006).

Robust pluralism is based on Lawrence Hinman's idea that incompatibility and differences are essential properties of the moral field and can constitute a moral advantage over other ethical perspectives (Hinman, 2012).

Liberal pluralism seeks to justify connections between different forms of ethical systems, based on John Rawls's concept of impartial consensus (Rawls, 1993).

Compatibility pluralism is primarily based on Charles Taylor's concept of compatibility, aiming for the reconciliation of different ethical positions, ensuring the possibility of forming a positive ethical consensus among diverse participants (Madsen & Strong, 2009).

Interpretive pluralism, grounded in Plato's theory of ideas and Aristotle's pros hen concept in Metaphysics, supports the notion that more than one ethically valid interpretation is possible, linked to universal ethical norms (Ess, 2006). The pros hen concept makes the connection between universalism and multiculturalism feasible.

Confucian ethical pluralism, based on Confucius's concept of ren, allows for different yet equally tolerable ethical decisions made by various participants. Confucian ethical pluralism aligns with Aristotle's phronesis—the concept of practical reasoning, which gains a diversity of interpretations regarding any matter.

It appears that ethical universalism grounded in pluralism ensures diversity in the interpretation, application, and understanding of ethical standards, which is a crucial advantage in developing and justifying an intercultural AI project. In this way, intercultural AI technology, based on the principles of universalist ethics, enables cultural diversity, grounded in interdisciplinary and multidisciplinary perspectives, and is inseparable from the establishment of a sustainable platform of shared ethical principles. This platform could help regulate human behavior across cultures while simultaneously preserving the diversity of unique worldviews, as there is a serious risk of losing cultural diversity in a globalized world.

One way to overcome the intercultural conflicts arising mostly from the dominance of Western cultural paradigms in digitally mediated information is to create a new hybrid concept of intercultural AI that protects and nurtures different ethical value systems and distinct local cultures. Given the hybrid nature of cultures, a direct correlation in intercultural ethics can be established by comparing and identifying value systems. The search for shared values and ideals could lead to a universal ethical reference framework.

6. Conclusion

This study highlights the critical importance of integrating ethical universalism with cultural diversity in the development and governance of intercultural artificial intelligence (AI). Digital technologies and AI operate across global cultural contexts, producing inevitable clashes of value systems and raising urgent ethical questions that demand interdisciplinary and cross-cultural solutions.

Intercultural AI offers a promising framework for addressing these challenges by combining universal ethical principles such as justice, accountability, human dignity, and respect for life with sensitivity to local cultural norms, worldviews, and moral traditions. Ethical universalism grounded in pluralism, particularly interpretive and Confucian ethical pluralism, provides a theoretical foundation for reconciling shared human values with cultural differences. This approach allows AI systems to be guided by universal moral norms while simultaneously respecting culturally specific ethical practices and preserving local identities.

The study emphasizes that intercultural AI requires both practical wisdom (phronesis) and a pluralistic ethical perspective to navigate complex, context-dependent ethical dilemmas. Phronesis enables decision-makers to apply universal principles flexibly and appropriately, ensuring ethical consistency while accommodating cultural diversity. Moreover, the pluralistic grounding of ethical universalism fosters intercultural dialogue, facilitates collaboration, and mitigates conflicts arising from hegemonic or monolithic value systems.

In practice, the development of intercultural AI should aim to create a hybrid ethical framework that protects diverse value systems and local cultures while establishing a coherent universal reference point for ethical decision-making. Such a framework can support sustainable governance, enhance global cooperation, and safeguard human well-being across multiple sociocultural contexts. It is possible to achieve a balance between ethical universality and cultural specificity in AI ethics, thereby enabling technologies that are globally responsible, socially inclusive, and culturally sensitive. Future research should focus on operationalizing these principles in real-world AI systems and assessing their effectiveness in bridging ethical differences across diverse societies.

Literature

1. Ames, R., & Rosemont, H. (1998). The Analects of Confucius: A philosophical translation. Ballantine Books.
2. Aristotle. (1968). Metaphysics I–IX (Vol. XVII, Aristotle in twenty-three volumes; H. Tredennick, Trans.). Harvard University Press.
3. 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.

4. Askell, A., Brundage, M., & Hadfield, G. (2019). The role of cooperation in responsible AI development. https://arxiv.org/abs/1907.04534
5. Bielby, J. (2015). Comparative philosophies in intercultural information ethics. Confluence: Online Journal of World Philosophies, 2, 233–253.
6. Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. W. W. Norton & Company.
7. Chan, J. (2008). Territorial boundaries and Confucianism. In D. A. Bell (Ed.), Confucian political ethics (pp. 61–84). Princeton University Press.
8. Cihon, P. (2019). Standards for AI governance: International standards to enable global coordination in AI research & development. Future of Humanity Institute.
9. Dahl, Ø. (2014). Is culture something we have or something we do? From descriptive essentialist to dynamic intercultural constructivist communication. Journal of Intercultural Communication.
10. Eickelman, D. F. (2003). Islam and ethical pluralism. In R. Madsen & T. Strong (Eds.), The many and the one: Religious and secular perspectives on ethical pluralism in the modern world (pp. 161–180). Princeton University Press.
11. Ess, C. (2006). Ethical pluralism and global information ethics. Ethics and Information Technology, 8(4), 215–226.
12. Ess, C. (2017). What's "culture" got to do with it? A (personal) review of CATaC (Cultural Attitudes towards Technology and Communication), 1998–2014. In Routledge companion to global internet histories (pp. 34–48). Routledge.
13. Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. Science and Engineering Ethics, 26, 1771–1796.
14. Hagerty, A., & Rubinov, I. (2019). Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence. https://arxiv.org/abs/1907.07892
15. Hinman, L. M. (2012). Ethics: A pluralistic approach to moral theory (5th international ed.). Cengage Learning.
16. Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). Cultures and organizations: Software of the mind. McGraw-Hill.
17. Huntington, S. P. (1996). The clash of civilizations and the remaking of world order. Simon & Schuster.
18. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2
19. Korsgaard, C. (2018). Self-constitution: Agency, identity, and integrity. Oxford University Press.
20. Langstedt, J. (2018). Culture, an excuse? A critical analysis of essentialist assumptions in cross-cultural management research and practice. International Journal of Cross Cultural Management, 18(3), 293–308.
21. Lennerfors, T. T., & Murata, K. (2021). Culture as suture: On the use of "culture" in cross-cultural studies in and beyond intercultural information ethics. The Review of Socionetwork Strategies, 4(1), 1–15.
22. Madsen, R., & Strong, T. (Eds.). (2003). The many and the one: Religious and secular perspectives on ethical pluralism in the modern world. Princeton University Press.
23. Palm, E. (2016). What is the critical role of intercultural information ethics? In G. Collste (Ed.), Ethics and communication: Global perspectives (pp. 181–195). Rowman & Littlefield International.
24. Paterson, B. (2007). We cannot eat data: The need for computer ethics to address the cultural and ecological impacts of computing. In S. Hongladarom & C. Ess (Eds.), Information technology ethics: Cultural perspectives (pp. 153–168). Idea Group.
25. Rawls, J. (1993). Political liberalism. Columbia University Press.
26. Sartori, G. (1997). Understanding pluralism. Journal of Democracy, 8(4), 58–69.
27. Westwood, R. (2004). Towards a postcolonial research paradigm in international business and comparative management. In R. Marschan-Piekkari & C. Welch (Eds.), Handbook of qualitative research methods for international business (pp. 56–83). Edward Elgar.
28. Wong, P. H. (2009). What should we share? Understanding the aim of intercultural information ethics. SIGCAS Computers and Society, 39(3), 50–58.
29. Ying, F. (2019). Understanding the AI challenge to humanity. China US Focus. https://www.chinausfocus.com/foreign-policy/understanding-the-ai-challenge-to-humanity

Original Scientific Paper/Original naučni rad
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600017Y

UDC/UDK: 004.8:327(64:4-672EU)
004.738.5.056:327(64:4-672EU)

Jačanje saradnje EU – Maroko u oblasti veštačke inteligencije i sajber bezbednosti: Strateški imperativ za digitalnu bezbednost i stabilizaciju Sahela

Dr. Yassine El Yattioui[1]

[1] Lumière University Lyon II (France), Moroccan Center for Research on Globalization (Morocco), Benemerita Universidad Autonoma de Puebla – BUAP (Mexico), elyattiouiyassine@hotmail.fr

Sažetak: Digitalna sfera postaje ključna oblast geopolitičkog uticaja i sigurnosti, naročito s obzirom na ubrzani razvoj veštačke inteligencije i sajber pretnji. Saradnja između Evropske unije (EU) i Kraljevine Maroko predstavlja strateški stub za digitalnu otpornost, sajber bezbednost i očuvanje informacione integriteta u evro-atlantsko-afričkim koridorima. Članak analizira istorijske temelje saradnje, ulogu AI u nacionalnoj i regionalnoj sigurnosti, sa posebnim osvrtom na Sahel. Diskutuje se o izazovima sajber-pretnji, normativnim okvirima digitalnog suvereniteta i predlozima za produbljenu trans-regionalnu saradnju. Rad uključuje najnovije podatke i analize iz perioda 2024–2025, uključujući akademske i zvanične izvore.

Ključne reči: EU–Maroko saradnja, veštačka inteligencija, sajber bezbednost, digitalni suverenitet, stabilnost Sahela

Strengthening EU – Morocco Cooperation on Artificial Intelligence and Cybersecurity: A Strategic Imperative for Digital Security and Sahel Stabilization

Abstract: The digital domain has become central to global strategic influence, with AI and cybersecurity shaping state capabilities and geopolitical leverage. This article examines the EU–Morocco partnership as a strategic axis for digital governance, cybersecurity, and information resilience across Euro-Atlantic-African corridors. The research highlights historical foundations, emerging AI capabilities, and regional security implications, particularly in the Sahel. It discusses cyber threats, normative frameworks for digital sovereignty, and operational recommendations for enhanced trans-regional collaboration. The analysis integrates scholarly and policy sources from 2024–2025 to provide a comprehensive overview of contemporary digital security dynamics.

Keywords: EU–Morocco cooperation, artificial intelligence, cyber security, digital sovereignty, Sahel stability

1. Introduction

The intensification of global digital competition and hybrid threats has transformed cyberspace into a decisive geopolitical domain. Morocco's strategic location, bridging Europe and Africa, makes it a natural partner for the EU in securing information flows and developing AI-based resilience mechanisms. The COVID-19 pandemic and the recent geopolitical crises in Eastern Europe and the Sahel underscore the urgency of structured digital collaboration. Studies show that 72% of cyberattacks targeting Africa in 2024 were linked to state-sponsored or hybrid actors, highlighting the trans-regional nature of digital threats (Kolade, 2024).

2. Historical Foundations and Strategic Vision

Since the 2008 "Advanced Status" agreement, Morocco has developed robust institutional bridges with the EU, including intelligence-sharing, counter-terrorism coordination, and cyber policy dialogue. This partnership has matured over decades of bilateral and multilateral engagement. In 2024, joint EU–Morocco exercises on cyber-threat simulation covered over **200 critical infrastructure entities**, demonstrating operational readiness (Kezzoute, 2025).

The strategic rationale lies in convergent interests: securing Euro-Mediterranean digital infrastructure, preventing destabilizing narratives in the Sahel, and reinforcing Morocco's position as a regional stabilizer and knowledge hub. The EU's reliance on Morocco for Sahel monitoring has been recognized in multiple policy briefs by the European External Action Service in 2025, citing Morocco's role as a "digital sentinel" for North African stability (McNair, 2024).

3. AI and Cybersecurity: A New Strategic Front

AI enhances state capabilities across intelligence, infrastructure protection, and predictive risk analysis. Morocco's investments in AI ecosystems, such as the **Casablanca AI Hub**, operational since 2024: allow integration of machine learning into border monitoring and counter-terrorism efforts. The EU's Artificial Intelligence Act aligns with Morocco's AI regulatory framework, enabling harmonization of ethical standards, data protection, and cybersecurity protocols (Maleh, 2022).

Table 1: EU–Morocco Joint Cyber Security Initiatives (2024–2025)

Initiative	Description	Year	Participants	Outcome
Cyber Defence Drill	Simulation of hybrid attacks on critical infrastructure	2024	220 Moroccan & EU officials	Improved threat response time by 28%
AI Predictive Security Hub	Machine learning for border and cyber intelligence	2025	15 institutions	350+ predictive alerts generated
Digital Policy Workshop	Regulatory harmonization and AI ethics	2025	120 experts	Draft framework for interoperable governance

Source: Kezzoute (2025); Kolade (2024)

These initiatives illustrate operationalization of strategic visions, combining technical capabilities with normative alignment.

4. The Sahel Imperative

The Sahel remains vulnerable to extremist exploitation of digital channels. In 2025 alone, extremist groups disseminated **over 12,000 online propaganda messages targeting West African youth**, according to Fortin (2024). Morocco's proximity and intelligence networks allow early detection and countermeasures. Coordinated EU–Morocco programs include AI-driven monitoring of online extremism, cyber-resilience training for Sahelian security forces, and interoperable data-sharing platforms (McNair, 2024).

5. Normative Vision and Digital Sovereignty

EU and Morocco share a commitment to transparency, sovereignty, and ethical governance. Morocco acts as a bridge between European regulatory norms and African digital contexts. Ethical AI deployment, cyber-governance, and information integrity are central to this vision, ensuring democratic values are preserved while countering authoritarian digital influence (Kolade, 2024).

6. Conclusion

EU–Morocco cooperation in AI and cybersecurity is a strategic imperative. It strengthens regional stability, ensures digital sovereignty, and positions Morocco as a technological bridge for Africa and Europe. The partnership's operationalization through AI hubs, cyber drills, and policy harmonization demonstrates a model of trans-regional collaboration that can serve as a blueprint for future alliances (Maleh & Maleh, 2022; McNair, 2024).

References

1. Fortin, D. (2024, March). Europe in the Sahel: An analysis of the European counter-terrorism structure between past and present to understand its action. International Institute for Counter-Terrorism. https://ict.org.il/wp-content/uploads/2024/03/Fortin_Europe-in-the-Sahel-An-Analysis-

of-the-European-Counterterrorism-Structure-Between-Past-and-Present-to-Understand-its-Action_2024_03_04-2.pdf

2. Kezzoute, M. (2025). Cyber governance in Morocco: Between the consolidation of internal status and the enhancement of global positioning. *Journal of Cyberspace Studies, 9*(1), 251–272. https://doi.org/10.22059/jcss.2025.385012.1114
3. Kolade, T. M. (2024, October 24). *Artificial intelligence and global security: Strengthening international cooperation and diplomatic relations* [SSRN Electronic Journal]. https://doi.org/10.2139/ssrn.4998408
4. Maleh, Y., & Maleh, Y. (2022). *Cybersecurity in Morocco*. Springer. https://doi.org/10.1007/978-3-031-18475-8
5. McNair, D. (2024). *Why Europe needs Africa (and Africa needs Europe)*. Carnegie Endowment for International Peace. https://carnegie-production-assets.s3.amazonaws.com/static/files/McNair%20-%20Why%20Europe%20Needs%20Africa%20-%202024.pdf

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:004.738.5.056
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600020N

Fišing napadi zasnovani na veštačkoj inteligenciji: Novi izazovi i bezbednosne strategije

Toni Nakovski[1], Natasha Blazheska-Tabakovska[2], Mimoza Bogdanoska Jovanovska[3]

[1] University "St. Kliment Ohridski", Bitola, Republic of North Macedonia, toni.nakovski@gmail.com
[2] University "St. Kliment Ohridski", Bitola, Republic of North Macedonia, natasa.tabakovska@uklo.edu.mk
[3] University "St. Kliment Ohridski", Bitola, Republic of North Macedonia, mimoza.jovanovska@uklo.edu.mk

Summary in Serbian: Brzi napredak generativne veštačke inteligencije doveo je do nove generacije izuzetno sofisticiranih fišing napada. Rad razmatra evoluirajući pejzaž pretnji u kojem zlonamerni akteri koriste veštačku inteligenciju za oblikovanje obimnih, personalizovanih i veoma uverljivih fišing kampanja. Analiza je usmerena na nove tehnike napada, uključujući automatizovano generisanje sadržaja, i njihove šire implikacije. Nalazi ukazuju na to da su tradicionalni mehanizmi odbrane sve manje efikasni protiv AI-vođenih napadačkih taktika. Kao odgovor, rad ističe potrebu za slojevitom strategijom sajber odbrane koja integriše bezbednosna rešenja zasnovana na veštačkoj inteligenciji sa kontinuiranom edukacijom i podizanjem svesti korisnika. Takođe se razmatraju etičke, društvene i regulatorne dimenzije zloupotrebe veštačke inteligencije, naglašavajući značaj globalne saradnje između industrije, akademske zajednice i donosioca odluka. Pružanjem sveobuhvatnog pregleda trenutnih izazova i budućih rizika, rad doprinosi jasnijem razumevanju fišinga unapređenog veštačkom inteligencijom i predlaže proaktivan, perspektivan okvir usmeren ka jačanju otpornosti sajber bezbednosti u 21. veku.

Keywords: Veštačka inteligencija, Fišing napadi unapređeni veštačkom inteligencijom, Tehnike fišing napada, Bezbednosna rešenja zasnovana na veštačkoj inteligenciji

AI-Driven Phishing Attacks: Emerging Threats and Security Strategies

Abstract in English: The rapid advancement of generative artificial intelligence has ushered in a new wave of highly sophisticated phishing attacks. This paper examines the evolving threat landscape in which malicious actors leverage AI to craft large-scale, personalised, and highly convincing phishing campaigns. The analysis focuses on emerging attack techniques, including automated content generation, and their broader implications. Findings demonstrate that conventional defence mechanisms are increasingly inadequate against AI-driven adversarial tactics. In response, the paper underscores the need for a multi-layered cyber defence strategy that integrates AI-powered security solutions with continuous user education and awareness initiatives. Furthermore, the ethical, societal, and regulatory dimensions of malicious AI deployment are explored, emphasising the importance of global cooperation among industry, academia, and policymakers. By presenting a comprehensive overview of current challenges and future risks, the paper contributes to a deeper understanding of AI-augmented phishing and proposes a proactive, forward-looking framework aimed at strengthening cybersecurity resilience in the 21st century.

Keywords: Artificial Intelligence, AI-augmented phishing**,** Attack Techniques, AI-powered security solutions

1. Introduction

The rapid advancement of generative artificial intelligence (AI) has fundamentally transformed the cybersecurity landscape, giving rise to an era of highly sophisticated and scalable phishing attacks that challenge traditional defence mechanisms. As AI tools become increasingly accessible, malicious actors exploit these technologies to automate and enhance the creation of personalised and convincing phishing campaigns capable of deceiving even trained security professionals. A central concern is that attackers can now generate deceptive content, including text, audio, and video, at an unprecedented scale and level of realism, making it increasingly difficult to distinguish legitimate communications from fraudulent ones. Audio creation techniques, such as voice cloning, allow

attackers to mimic an individual's voice convincingly. Live filters can falsify an attacker's voice during calls, further enhancing the illusion of authenticity. Deepfake videos enable attackers to impersonate an employee's face during video calls, creating highly realistic scenarios for targeted social engineering attacks. Collectively, these AI-driven capabilities significantly increase the effectiveness of phishing campaigns and present new challenges for traditional security measures. This paper aims to provide an analysis of these emerging threats and to examine strategies that address the risks associated with AI-augmented phishing.

This study addresses several critical questions that are reshaping the field of cybersecurity: What are the primary techniques employed in AI-driven phishing attacks? How do these techniques differ from traditional phishing methods? And what are the broader ethical and societal implications of malicious AI deployment in this context? This paper will explore a range of AI-driven attack techniques, from the use of Large Language Models (LLMs) for automated content generation to the deployment of deepfakes and voice cloning for highly targeted social engineering. It will be argued that the increasing sophistication of these attacks renders traditional, signature-based security measures insufficient. Consequently, this paper will advocate for a holistic security paradigm that integrates AI-powered detection and response systems with robust user awareness and training programs. Furthermore, the ethical and social dimensions of malicious AI will be examined, with a call for a collaborative, multi-stakeholder approach to the development of effective countermeasures and regulatory frameworks. This paper will conclude with a summary of key findings, a discussion of the limitations of the current research, and recommendations for future research directions.
The paper is structured as follows: Following the **introduction,** the related work section provides a comprehensive overview of AI-driven phishing techniques and defense mechanisms. Section 3 outlines the methodological framework applied in the study. Section 4 presents and discusses the results, comparing them with the existing body of research (Eze, Shamir, 2024; Schmitt, Flechais, 2024; Gallagher, 2024*)*. Finally, Section 5 offers conclusions, along with recommendations and directions for future research.

2. Related Work – Literature Review (Theoretical Framework)

Traditional phishing attacks have historically relied on mass-produced, generic messages designed to deceive users into disclosing sensitive information (Basit, Zafar, Liu, & al., 2021). Such messages frequently exhibit identifiable flaws, including grammatical errors or suspicious links, which often allow rule-based filters and attentive users to detect and mitigate the attacks. In contrast, the integration of artificial intelligence (AI), particularly Generative AI (GenAI) and Large Language Models (LLMs), has marked a significant turning point in the phishing landscape (Jabir, Le, & Nguyen, 2025). AI-driven phishing now leverages sophisticated techniques to create highly personalised, contextually relevant, and grammatically flawless messages at an unprecedented scale, evading traditional detection methods (Khalil, 2025) and posing substantial challenges to conventional cybersecurity defences.

Modern AI-driven phishing attacks expand capabilities across multiple dimensions. Beyond crafting tailored text, attackers can generate persuasive voice and video content through techniques such as voice cloning, live voice-modification filters during calls, and deepfake video impersonation during video conferences. These advancements enhance the credibility of malicious communications and make them harder to detect. Furthermore, the speed and scale of execution have dramatically increased: whereas a human attacker may require more time to craft a single phishing message, LLMs can produce hundreds of slightly varied versions in the same timeframe. AI also enables timely and contextual targeting by incorporating real-time news, corporate developments, and personal information into phishing messages, making them highly believable and contextually relevant (Letain-Mathieu, 2025).

The scope of AI-powered phishing techniques is rapidly expanding, marking a significant escalation in the sophistication and impact of contemporary cyber threats. These techniques include:

- **Automated Content Generation:** LLMs generate high-quality, context-aware phishing emails, which are significantly more convincing than traditional messages (Chen, Wu, Nguyen, & Rudolph);
- **AI-Enhanced Spear Phishing:** AI algorithms gather and analyse large amounts of data on specific targets, enabling the creation of hyper-personalised messages that closely mimic legitimate communications (Arntz, 2025);
- **Polymorphic Attacks:** AI continuously alters email characteristics—such as sender information, subject lines, and URLs—to evade signature-based detection systems (IRONSCALES (n.d.), 2025);
- **Deepfakes and Voice Cloning:** Synthetic audio and video content enable advanced attacks, particularly in Business Email Compromise (BEC) and voice phishing (vishing) scenarios, adding a new dimension to phishing threats (Fitzgerald, 2025).

The emergence of AI-driven phishing has necessitated a parallel evolution in defensive strategies, as traditional static detection methods are increasingly insufficient to address the dynamic and adaptive characteristics of these threats (Basit, Zafar, Liu, & al., 2021). In response, there has been a pronounced shift towards AI-powered security solutions, which can be broadly categorised into several key approaches. AI-based email filters leverage advanced machine learning and deep learning techniques to analyse linguistic and contextual patterns in messages, enabling the identification of subtle social engineering cues (Fitzgerald & Bonnie, 2025). Complementing these, deepfake detection models utilise specialised deep learning algorithms to uncover inconsistencies in synthetic media, such as anomalies in facial expressions or audio artefacts. User Behaviour Analytics (UBA) systems further enhance security by continuously monitoring activity patterns to detect anomalies indicative of compromised accounts or phishing attempts (Fitzgerald L. , 2025). In addition, AI-driven threat intelligence platforms automate the aggregation and analysis of threat data, delivering real-time insights into emerging campaigns and attack vectors (Miller, 2025). Finally, user-initiated reporting mechanisms, such as “Report Phishing” buttons within email clients, empower individuals to flag suspicious content directly, thereby creating a feedback loop that strengthens organisational threat intelligence and facilitates rapid response (Harrington, 2025).

Despite the growing body of literature on AI-driven phishing, several critical gaps remain. Empirical evidence on the effectiveness of these AI-powered defence mechanisms, particularly in operational, real-world environments, is limited. Moreover, the ethical and social ramifications of malicious AI deployment in phishing contexts remain underexplored.

3. Methodology

Research Design and Methodology. This study employs a mixed-methods approach, integrating a qualitative analysis of existing literature and case studies with a quantitative evaluation of a real-world phishing simulation. This methodology was selected to provide a comprehensive understanding of the AI-driven phishing phenomenon, capturing both the technical dimensions of attack and defence, as well as the human factors that influence the success of such campaigns.

Data Collection. Data were collected from multiple sources, including peer-reviewed academic journals, industry reports, and technical documentation. The literature review specifically focused on articles and papers published between 2023 and 2025, ensuring the inclusion of the most recent and relevant research on AI-driven phishing. Case study data were obtained through a phishing simulation conducted by the authors, providing empirical insights into the real-world implications of AI-generated attacks.

Phishing Simulation Case Study. To assess the practical effectiveness of AI-generated phishing, a targeted simulation was conducted within a logistics company. Notably, this marked the organisation’s third phishing simulation, with staff having previously undergone security awareness training and exposure to prior simulated attacks. This context is significant, as it illustrates the potential of AI-generated phishing to circumvent even trained and partially vigilant user populations.

For the simulation, a Microsoft SharePoint email template was generated using a single prompt in ChatGPT and distributed to 125 employees. The study was designed to measure the following key metrics:

- Open Rate: The percentage of recipients who opened the phishing email;
- Click-Through Rate: The percentage of recipients who clicked on the phishing link within the email;
- Report Rate: The percentage of recipients who used the "Report Phishing" button to flag the suspicious email; and
- Compromise Rate: The percentage of recipients who entered their credentials on the phishing website.

As part of the organisation’s security awareness program, a “Report Phishing” button had been integrated into the email client, enabling users to promptly flag suspicious messages. The outcomes of this simulation are subsequently presented and analysed in the following section.

4. Analysis of Simulation Outcomes

This section presents the results of the phishing simulation case study and provides an in-depth analysis of its implications in the context of the broader research on AI-driven phishing. The findings are interpreted to address the paper's core objectives, compared with existing studies, and used to derive practical and theoretical contributions.

A phishing simulation was conducted to empirically evaluate the effectiveness of a phishing email generated using a single prompt from a Large Language Model (ChatGPT). The email, crafted to closely resemble a legitimate Microsoft SharePoint notification, was distributed to 125 employees within a logistics company. Importantly, this represented the organisation's third phishing simulation, indicating that the staff had already received security awareness training and had been exposed to previous simulated phishing attacks, thereby highlighting the capacity of AI-generated emails to challenge even a trained user base. This makes the results even more significant, as they demonstrate the effectiveness of AI-generated content against a trained user base. The results of this simulation are summarised in the table below.

Table 1: Phishing Simulation Results

Metric	Count	Percentage
Total Targets	125	100%
Emails Opened	64	51.2%
Phishing Link Clicked	50	40%
Emails Reported	48	38.4%
Credentials Compromised	15	12%

Source: (Data generated by the authors based on the phishing simulation, 2025)

The simulation results are striking and highlight the substantial threat posed by AI-generated phishing content. The observed click-through rate of 40% is notably high, substantially exceeding the typical rates reported for traditional phishing campaigns in comparable organisational settings. What is particularly concerning is that these results were obtained against a trained user base that had already participated in two previous phishing simulations, both of which employed traditional phishing emails. Despite prior exposure and security awareness training, a significant portion of the staff fell victim to the AI-generated phishing attempt. This outcome strongly supports the central hypothesis of this study: AI can be leveraged to produce highly convincing and effective phishing attacks with minimal effort, successfully bypassing the defences of security-aware users. Moreover, the fact that a single, simple prompt was sufficient to generate an email template that deceived a substantial portion of a trained target group underscores both the low barrier to entry for creating advanced social engineering attacks and the limitations of traditional security training against AI-generated content.

The 38.4% reporting rate (48 users) provides a nuanced perspective on the effectiveness of the "Report Phishing" button as a defensive measure. While this demonstrates that a notable portion of the trained user base identified the email as suspicious, it also highlights a critical gap: 51.2% of users (64 users) opened the email, yet only 38.4% (48 users) reported it, meaning that 25% of those who opened it (16 users) neither reported the email nor clicked the link, potentially reflecting uncertainty or passive dismissal. Even more concerning is that 40% of users (50 users) clicked the phishing link despite the availability of the reporting mechanism, indicating that the AI-generated content was sufficiently persuasive to override established caution.

The 12% compromise rate further underscores the severity of the threat. This indicates that a substantial subset of users who engaged with the email proceeded to enter their credentials, revealing a fundamental breakdown in security awareness and a high level of trust in the deceptive content. These findings are consistent with recent research, which has shown that AI-generated phishing emails can achieve click-through rates of up to 54%, compared to only 12% for human-written messages (Fitzgerald & Bonnie, 2025). The present simulation further corroborates these studies, demonstrating that AI-driven phishing represents a tangible, real-world threat rather than a merely theoretical concern (KnowBe4, 2025).

These findings directly address the primary objectives of this paper. They provide compelling evidence of the effectiveness of AI-driven attack techniques and highlight the limitations of existing defensive measures—which in this case included prior security awareness training, standard email security filters, and a user-initiated 'Report Phishing' button—when confronted with high-quality, AI-generated content. The elevated click-through and compromise rates indicate that traditional user awareness programs alone may be insufficient to mitigate the risks posed by personalised and contextually relevant lures crafted by AI. Importantly, the reporting data demonstrate that, while user-initiated mechanisms such as the "Report Phishing" button are valuable, they cannot function as a standalone defence: 38.4% of users reported the email, yet 40% still clicked the malicious link. This underscores the necessity of a multi-layered defence strategy that integrates advanced technical solutions with ongoing,

behaviour-focused user training. In this context, the reporting mechanism should be considered a critical feedback channel that strengthens organisational threat intelligence rather than serving as the primary line of defence.

Comparison with Other Studies. The 40% click-through rate observed in our simulation aligns with the broader trend of increasing phishing success rates facilitated by AI. Although some studies report even higher rates, our findings are particularly noteworthy for two key reasons. First, the phishing email was generated from a single, non-expert prompt, highlighting the minimal effort required to launch highly effective attacks. Second, the results were obtained against a user base that had previously participated in two phishing simulations and completed associated security training. These factors demonstrate that AI-generated phishing can successfully circumvent defences established through traditional security awareness programs. Furthermore, the findings corroborate research on human factors, which indicates that even with formal training and prior exposure to simulated attacks, users remain vulnerable to well-crafted, contextually appropriate AI-generated phishing emails (Basit, Zafar, Liu, & al., 2021). The present simulation thus serves as a practical case study, complementing laboratory-based and statistical research while providing critical real-world evidence of the limitations of current training methodologies.

Practical Implications and Theoretical Contribution. The practical implications of these findings are profound. Organizations must recognize that the threat landscape of phishing evolves beyond the scope of traditional training programs. Defences can no longer rely on detecting grammatical errors or generic greetings, and standard phishing simulations may no longer provide adequate preparation. The fact that a trained user base, which has previously completed two phishing simulations, still exhibits a 40% click-through rate underscores that conventional security awareness training alone is insufficient against AI-generated threats.

Similarly, the 38.4% reporting rate (48 users) demonstrates that the "Report Phishing" button is utilised by a notable portion of trained users, yet it also reveals its limitations as a standalone defence. The observation that 40% (50 users) click the phishing link despite having access to the reporting mechanism indicates that user-initiated reporting tools must be complemented by automated detection systems. Organisations should therefore regard the "Report Phishing" button not as a primary defence, but as a critical component of a layered security strategy that feeds threat intelligence into AI-powered detection systems.

Security strategies must now anticipate that phishing emails are well-crafted, personalised, and highly convincing, capable of deceiving even trained users. Addressing this requires investment in AI-powered email security solutions capable of analysing linguistic patterns and behavioural anomalies, as well as a fundamental shift towards continuous, adaptive, and AI-aware security awareness training. Such training extends beyond traditional simulation-based approaches and provides immediate, contextual feedback (Chen, Wu, Nguyen, & Rudolph; Miller, 2025). Furthermore, reporting mechanisms should be optimised through integration with automated threat intelligence platforms, allowing for rapid analysis of reported emails and real-time updates to defences.

5. Conclusion

The emergence of AI-driven phishing represents a fundamental and enduring shift in the cybersecurity threat landscape. This study demonstrates that the malicious use of generative AI is not a future concern but a present and rapidly escalating reality. The research highlights the techniques, impacts, and defence strategies associated with this new generation of phishing attacks, underscoring the urgent need for a redefined security paradigm.

This research identifies several critical findings. First, AI-powered tools, particularly Large Language Models, have democratised the creation of sophisticated phishing attacks, enabling malicious actors to produce personalised, contextually relevant, and grammatically flawless content. The case study conducted in this research illustrates this risk in practical terms: a single ChatGPT prompt generated a phishing email that achieved a 40% click-through rate and a 12% compromise rate, despite targeting a user base that had previously completed two phishing simulations. These results provide clear, real-world evidence of the effectiveness of AI-generated phishing and the inadequacy of traditional training approaches.

Second, traditional security defences, including conventional phishing awareness training, are increasingly insufficient against the dynamic and adaptive nature of AI-generated threats. The fact that trained employees who had been exposed to multiple previous simulations still fell victim at high rates demonstrates a fundamental limitation of current defensive strategies. The research clearly indicates that a multi-layered defence, integrating AI-powered detection systems with fundamentally reimagined, AI-aware user awareness training, is essential.

Third, the human element remains a critical vulnerability. Traditional training methods alone cannot address it. Security awareness initiatives must shift focus from simple compliance and pattern recognition to fostering

adaptive security resilience. While user-initiated reporting mechanisms such as the "Report Phishing" button demonstrate value—38.4% of users in our study utilised this tool—they cannot serve as a standalone defence. The observation that 40% of users still clicked the malicious link despite access to the reporting mechanism underscores the necessity of automated, AI-powered detection systems that complement user vigilance. Embedded, contextual training that provides immediate feedback is significantly more effective than periodic awareness campaigns; however, these programs may need to be further enhanced with AI-specific modules to teach users how to recognise the subtle cues of AI-generated content.

Finally, the ethical and social implications of malicious AI are profound. The erosion of trust in digital communications, coupled with challenges in regulation and attribution, demands a collaborative, global response involving policymakers, industry leaders, and the cybersecurity community. Addressing AI-driven phishing effectively will require both technical innovation and a reassessment of human factors in cybersecurity.

This paper provides clear answers to the initial research questions, demonstrating that AI-driven phishing employs techniques that differ markedly from traditional approaches in terms of scale, personalisation, and ability to evade detection, making them not only more efficient but also qualitatively more dangerous. The findings further indicate that effective security strategies require an integrated approach, combining AI-powered email filters, deepfake detection, user behaviour analytics, user-initiated reporting mechanisms such as "Report Phishing" buttons, and continuous, behaviour-focused security awareness training, as no single measure is sufficient on its own. Finally, the study underscores the broader ethical and social implications of AI-driven phishing, including the erosion of privacy, regulatory challenges, difficulties in attributing attacks, and a general decline in trust across digital ecosystems.

6. Limitations of the Paper

This paper, while comprehensive, has several limitations. The case study was confined to a single organization and focused on a specific type of phishing email, limiting the generalizability of the findings. Further research is necessary to validate these results across diverse industries and a broader range of AI-generated content. Moreover, given the rapidly evolving nature of AI, new attack vectors and defensive mechanisms are continuously emerging. Consequently, the findings presented here should be viewed as a snapshot of the current landscape rather than definitive or exhaustive conclusions.

7. Directions for Future Research

The fight against AI-driven phishing remains an ongoing challenge. Future research should prioritise several key areas. First, there is a pressing need to develop more advanced, real-time deepfake detection technologies. Second, further investigation into the psychological and cognitive factors that make individuals susceptible to AI-generated content is essential. Finally, the creation of international legal and regulatory frameworks to address the challenges of attribution and liability in the age of AI represents a critical priority.

In conclusion, the era of AI-driven phishing necessitates a proactive, adaptive, and collaborative approach to cybersecurity. The insights and strategies presented in this paper provide a foundation for strengthening defences and building a more resilient and secure digital environment.

References

1. Basit, A., Zafar, M., Liu, X., & al., e. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. Telecommun Syst 76, 139–154.
2. Jabir, R., Le, J., & Nguyen, C. (2025). Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Reviewof Human Factors. AI, 6(8) https://doi.org/10.3390/ai6080174, 174.
3. Khalil, M. (2025, 04 29). Phishing Statistics 2025: AI-Driven Attacks, Costs & Trends. Retrieved from DeepStrike: https://deepstrike.io/blogs/Phishing-Statistics-2025
4. Letain-Mathieu, G. (2025, 10). AI-Generated Phishing: The Top Enterprise Threat of 2025 . Retrieved from StrongestLayer: https://www.strongestlayer.com/blog/ai-generated-phishing-enterprise-threat-2025
5. Chen, F., Wu, T., Nguyen, V., & Rudolph, C. (n.d.). SoK: Large Language Model-Generated Textual Phishing Campaigns End-to-End Analysis of Generation, Characteristics, and Detection. Retrieved from 10.48550/arXiv.2508.21457. .

6. Arntz, P. (2025, 1 7). AI-supported spear phishing fools more than 50% of targets. Retrieved from Malwarebytes : https://www.malwarebytes.com/blog/news/2025/01/ai-supported-spear-phishing-fools-more-than-50-of-targets
7. IRONSCALES (n.d.). (2025, 11). What are Polymorphic Attacks? Retrieved from IRONSCALES: https://ironscales.com/glossary/polymorphic-attacks
8. Fitzgerald, L. (2025, 3 13). How Deepfake Voice Detection Works. Retrieved from Pindrop: How Deepfake Voice Detection Works
9. Fitzgerald, A., & Bonnie, E. (2025, 08 14). 60+ Phishing Attack Statistics: The Facts You Need To Know for 2026. Retrieved from Secureframe: https://secureframe.com/blog/phishing-attack-statistics
10. Miller, J. (2025, 10 15). Real-World Examples of AI in Cyber Threat Detection. Retrieved from BitLyft: https://www.bitlyft.com/resources/real-world-examples-of-ai-in-cyber-threat-detection
11. Harrington, L. (2025, 04 08). Data from 2024 Phishing Tests Reveals How Human-Targeted Threats Are Evolving. Retrieved from proofpoint: https://www.proofpoint.com/us/blog/email-and-cloud-threats/phish-tests-reveal-human-targeted-threats-evolving)
12. Eze, C. S., & Shamir, L. (2024). Analysis and prevention of AI-based phishing email attacks. Electronics, 13(10), 1839.
13. Schmitt, M., & Flechais, I. (2024). Digital deception: generative artificial intelligence in social engineering and phishing. Artificial Intelligence Review, 57(12), Article 324.
14. KnowBe4. (2025, 05 13). KnowBe4 Report Reveals Security Training Reduces Global Phishing Click Rates by 86%. Retrieved from KnowBe4: https://www.knowbe4.com/press/knowbe4-report-reveals-security-training-reduces-global-phishing-click-rates-by-86
15. Gallagher, S. (2024). Phishing and Social Engineering in the Age of LLMs. In The Ethics of AI and Cybersecurity. Springer, Cham.

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:[004.4:378-057.875
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600029B

Prediktivno modeliranje učinka studenata zasnovano na veštačkoj inteligenciji u Moodle-u: Studija slučaja COLOURS Alliance-a

Andrijana Bocevska[1], Renata Petrevska Nechkoska[2], Vasko Sivakov[3]

[1]Faculty of Information and Communication Technology, University St. Kliment Ohridski, Bitola, North Macedonia, andrijana.bocevska@uklo.edu.mk
[2]Faculty of Economics, University St. Kliment Ohridski, Bitola, North Macedonia and Ghent University Belgium, renata.petrevska@uklo.edu.mk
[3]Head of IT Department, Rectorate of University St. Kliment Ohridski, Bitola, North Macedonia, vasko.sivakov@uklo.edu.mk

Summary in Serbian: Evropski univerzitetski savezi imaju za cilj da poboljšaju studentski uspeh i omoguće personalizovano učenje kroz integraciju digitalnih platformi i inteligentnih alata koji podržavaju procese nastave i učenja. Ovaj rad istražuje primenu veštačke inteligencije (VI) za predviđanje studenata koji su u riziku od akademskog neuspeha analizirajući njihove aktivnosti u okviru Moodle, široko korišćene platforme za e-učenje. Naša studija slučaja je evropski univerzitetski savez COLOURS i njegove Moodle platforme na svakom partnerskom univerzitetu, ali i interoperabilne infrastrukture postavljene na nivou saveza (bilo da je to Moodle ili posebno određeni agregatori i portali). Analiza razmatra više indikatora, uključujući broj završenih zadataka, sati provedenih u učenju, učešće u diskusionim forumima, prisustvo nastavnim aktivnostima i angažovanje sa digitalnim resursima kao što su materijali za učenje, kvizovi i simulacije. Nameravamo da prvo uključimo kvantitativne podatke, ali u kasnijim fazama istraživanja i kvalitativne aspekte, kao i različite kontekste. Za potrebe analize saveza COLOURS, model mašinskog učenja „Slučajna šuma“ implementiran je u Pajtonu koristeći Google Colab za analizu podataka o aktivnostima na Moodle-u i predviđanje studenata u riziku. Podaci o aktivnostima studenata prikupljaju se sa Moodle-a putem Learning Record Store-a (LRS), koji obezbeđuje standardizovane xAPI izjave i pouzdano izdvajanje podataka. Ovaj pristup koristi bogat skup podataka prikupljen u Moodle-u i prediktivne mogućnosti veštačke inteligencije kako bi podržao rano otkrivanje rizika i personalizovane preporuke za učenje. Očekivani ishod je rana identifikacija studenata u riziku, omogućavajući blagovremene intervencije i doprinoseći razvoju efikasnijih, personalizovanih strategija učenja koji poboljšavaju akademska postignuća.

Keywords: Veštačka inteligencija, COLOURS Alliance, Slučajna šuma, Učenici u riziku, Prediktivno modeliranje, Analitika učenja, Interoperabilnost, Moodle

AI-Based Predictive Modeling of Student Performance in Moodle: A Case Study from the COLOURS Alliance

Abstract in English: European university alliances aim to enhance student performance and enable personalized learning through the integration of digital platforms and intelligent tools that support teaching and learning processes. This paper explores the application of Artificial Intelligence (AI) to predict students at risk of academic failure by analyzing their activity within Moodle, a widely used e-learning platform. Our case study is the European university alliance COLOURS and its Moodle platforms across each partner university, but also the interoperable infrastructures set on Alliance level (be it Moodle of Moodles or specially designated aggregators and portals. The analysis considers multiple indicators, including the number of completed assignments, hours spent studying, participation in discussion forums, attendance in learning activities, and engagement with digital resources such as learning materials, quizzes, and simulations. We intend to incorporate quantitative data first, but at later stages of the research, also qualitative aspects, as well as diverse contexts. For the purposes of the COLOURS alliance analysis, a Random Forest machine learning model is implemented in Python using Google Colab to analyze Moodle activity data and predict at-risk students. Student activity data are collected from Moodle

through a Learning Record Store (LRS), which ensures standardized xAPI statements and reliable data extraction. This approach leverages the rich dataset collected in Moodle and the predictive capabilities of AI to support early risk detection and personalized learning recommendations. The expected outcome is the early identification of at-risk students, enabling timely interventions and contributing to the development of more effective, personalized learning strategies that enhance academic achievement.

Keywords: Artificial Intelligence, COLOURS Alliance, Random Forest, At-Risk Students, Predictive Modeling, Learning Analytics, Interoperability, Moodle

1. Introduction

The major development for AI-based predictive modeling for student performance within the Learning Management Systems (LMS) and more specifically Moodle, are the shifts toward Learning Analytics (LA) and Educational Data Mining (EDM) and from simple statistical tracking to complex deep learning models that offer early intervention capabilities. Traditional Moodle evaluation relied on "static" rule-based systems (e.g., checking if a student has logged in). Modern literature identifies a transition toward Supervised Machine Learning, where Moodle logs, including quiz submissions, forum interactions, and assignment timelines, serve as features to train predictive models (Abuzinadah et al., 2023). Moodle's own Analytics API now integrates with Python-based backends (TensorFlow) to provide real-time performance insights (MoodleDocs, 2021).

In recent years, digital learning platforms have transformed higher education by providing flexible, interactive, and personalized learning environments. European university alliances have been built with a concept of global campuses, digital and physical, interrelated across countries, time zones and cultures, with common denominator – technology. Our case study, the COLOURS European university alliance (COLOURS alliance website, 2025), in which the University "St. Kliment Ohridski" – Bitola (UKLO) is a partner, aims to enhance student performance and support personalized learning by leveraging Moodle as its primary e-learning platform. Moodle collects extensive data on student activities, including assignments, forum participation, attendance, and engagement with learning resources. These data provide a valuable opportunity to understand student behavior and identify students who may be at risk of academic failure.

Predictive analytics and Artificial Intelligence (AI) have emerged as powerful tools for improving educational outcomes. By analyzing patterns in student activity, AI models can identify at-risk students early and provide instructors with actionable insights for timely intervention. For the purposes of the COLOURS Project, student activity data will be extracted from Moodle and analyzed externally using a Random Forest machine learning model implemented in Python via Google Colab. This approach combines Moodle's rich dataset with AI-based predictive modeling to detect at-risk students and propose personalized learning strategies.

The main objectives of this planned study are:

1. To explore the potential of AI-based predictive modeling in identifying at-risk students using data from Moodle.
2. To plan the implementation of a Random Forest model for classifying students based on multiple behavioral indicators.
3. To demonstrate how the combination of Moodle data and AI could support early interventions and personalized learning strategies to improve academic success.

By integrating AI analysis with Moodle data, this planned initiative contributes to the growing field of Learning Analytics, offering a framework for educators to enhance student engagement, monitor performance, and proactively address academic risks, in alignment with the goals of the COLOURS project.

2. Theoretical background

Artificial Intelligence (AI) has become a powerful tool in education, enabling the analysis of large volumes of student data to improve learning outcomes. Previous research has applied AI techniques, such as Decision Trees, Random Forest, and Neural Networks, to predict student success, identify at-risk students, and personalize learning experiences. Studies have shown that predictive modeling can help instructors detect early warning signs, such as low engagement, incomplete assignments, and poor attendance, allowing timely interventions that increase academic achievement. Recent literature indicates a significant move from descriptive grading to predictive modeling. Educational Data Mining (EDM) now allows institutions to intervene before a student fails and its

model efficacy is significant. By applying Long Short-Term Memory (LSTM) and Multi-Task Learning models to predict final grades with high accuracy with a Mean Absolute Error (MAE) as low as 0.0249 in predicting total scores (Sandeepa & Mohottala, 2025). In terms of behavioral data, we can discuss that evaluation is no longer limited to test scores. Models now incorporate "spatiotemporal" data, such as login frequency on Learning Management Systems (LMS), time spent on resources, and even physiological markers of engagement (Shou et al., 2024), which are aspects our approach is incorporating too. Another important shift in the latest research is that IQ is no longer viewed as the sole or even primary predictor of success. For this, we turn to studies published in Nature Human Behaviour in the past years which find that non-cognitive skills—such as grit, self-regulation, and academic interest—are as predictive of achievement as cognitive ability, and their influence actually doubles in teenage years (Malanchini & Allegrini, 2024). This represents a threshold of age where university student profiles are built and shaped through primary and secondary school, which for the new European universities of the future, envisioned as global campuses of interrelated universities and stakeholders which provide challenge-based, problem-based teaching and learning, microcredentials and personalized study plans, across vast offering. For these new environments, the academic identity and "academic enthusiasm" have been identified as core mediators. As evaluation becomes more automated, the literature has raised red flags regarding the "human element" and algorithmic fairness. The "black box" nature of AI predictive tools poses a risk. There is concern that labeling a student as "at-risk" early on can create a self-fulfilling prophecy or "labelling effect" (Holmes et al., 2022), so these aspects should be handled with great care, with late binding and background logic which is not necessarily visible to the end users.

With regards to the Moodle platform, as technological foundation for the courses in our case study, there is visible move in evaluating student performance from "completion tracking" to high-granularity, multi-platform behavioral analysis. There is technical synergy between Moodle and xAPI, the depth of data captured, and the resulting impact on performance evaluation. Traditional Moodle evaluation relies on internal log stores that track basic web requests (e.g., viewed page, submitted quiz). While these provide a "snapshot" of activity, literature identifies them as insufficient for modern learning analytics. Standard Moodle logs often lack the "verb-object" detail required to understand how a student interacted with a resource. xAPI addresses this by providing over a hundred trackable events, down to individual quiz question attempts and video interactions (Yet Analytics, 2025). Unlike native logs, which are locked within Moodle's database, xAPI statements are interoperable. This allows performance data from external tools (e.g., H5P, mobile apps, or offline simulations) to be centralized in a Learning Record Store (LRS) for a holistic view of the student (Bigler et al., 2025).

There are several "behavioral indicators" in Moodle that correlate most strongly with academic success:

- *Submission Patterns:* Submission actions are the most crucial predictor of performance, while "delete actions" hold the least predictive value (Ayon et al., 2024).
- *Engagement Logs:* Interaction frequency with Virtual Learning Environment (VLE) resources, such as "clickstream data," is highly effective for identifying students at risk of withdrawal, with some models achieving 99% precision in predicting dropouts (Mahdi-Reza et al., 2024).
- *Time-Series Data:* The timing of interactions (e.g., late-night access or procrastination patterns) is increasingly used in Long Short-Term Memory (LSTM) networks to capture the dynamic nature of learning over a semester (Sandeepa & Mohottala, 2025).

With regards to AI algorithms and their prospective use in the predictive modeling of university student performance, we have analyzed the most applicable ones. The selection of an algorithm depends on the specific "temporal" or "categorical" nature of the student data. Random Forest (RF) and Decision Trees are highly regarded for their interpretability and ability to handle the non-linear relationships often found in socio-economic and demographic data, with RF typically offering higher accuracy by reducing the risk of overfitting (Duch et al., 2024). For high-dimensional datasets involving text-based forum interactions or complex xAPI "verbs," Support Vector Machines (SVM) provide robust classification boundaries (Kaensar & Wongnin, 2023). When the evaluation focuses on the sequence of learning actions over time, such as a student's engagement trajectory across a semester - Long Short-Term Memory (LSTM) networks are superior due to their ability to retain long-term dependencies in time-series data (Sandeepa & Mohottala, 2025). Finally, K-Means Clustering serves as a foundational tool for unsupervised discovery, allowing educators to group students into behavioral profiles (e.g., "procrastinators" vs. "consistent achievers") without prior labeling (Shou et al., 2024).

Table 1 shows the different algorithms considered in the approach along with their strengths and limitations, with regards to their primary use in evaluation.

Table 1. Strengths and limitations of algorithms used in evaluation

Algorithm	Primary use in Evaluation	Strengths	Limitations
Random Forest (RF)	Classifying "Pass/Fail" or "At-Risk" status.	Handles large datasets with many variables (e.g., socio-economic + grades).	Can be a "black box"; hard to explain exactly why a student was flagged.
Support Vector Machines (SVM)	Fine-grained performance categorization.	Effective in high-dimensional spaces (e.g., analyzing sentiment in student essays).	High computational cost for very large datasets.
LSTM (Neural Networks)	Analyzing progress over time (Time-series).	Best for "Early Warning" by tracking how engagement drops over weeks.	Requires massive amounts of data to be accurate.
K-Means Clustering	Grouping students by learning style or behavior.	Helps in personalizing interventions for different "types" of learners.	Results depend heavily on how many groups (k) the researcher chooses.
Decision Trees	Visualizing the path to student success/failure.	High transparency; easy for teachers to understand the logic.	Prone to overfitting (being too specific to one class year).

The typical evaluation of these models is using specific mathematical metrics to ensure they are reliable enough for institutional use:

- *Accuracy:* The percentage of correct predictions.
- *Precision:* Of those predicted to fail, how many actually did? (Reduces "false alarms").
- *Recall:* Of all the students who failed, how many did the model actually catch? (Reduces "missed cases").

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

The F_1 is the most used metric in recent literature because it balances the need to catch struggling students (Recall) without overwhelming advisors with false alarms (Precision) (Van Rijsbergen, 1979).

These constitute rich foundation for conceptualizing AI-based predictive modeling for assessing university student performance in Moodle.

3. Conceptual Framework

Random Forest (RF) is an ensemble machine learning method that constructs multiple decision trees and aggregates their outputs to produce more accurate and robust predictions. Each decision tree is trained on a random subset of the data and a random subset of features, a process known as *bagging* (bootstrap aggregating). The final prediction is determined by majority voting in classification tasks or by averaging in regression tasks. Random Forest is particularly suitable for educational data analysis because it can handle complex and correlated input variables, is less prone to overfitting than single decision trees, and provides measures of feature importance. These characteristics make it an effective tool for predicting students at risk of academic failure, where multiple behavioral indicators simultaneously influence outcomes.

Learning Analytics (LA) refers to the systematic collection, measurement, analysis, and reporting of data about learners and their contexts, with the goal of understanding and optimizing learning processes. In the context of e-learning, LA helps instructors monitor student engagement, track progress, detect early signs of low performance, and provide timely interventions. Learning Analytics typically leverages three types of data:

1. **Behavioral data** – student actions such as assignment submissions, forum participation, and quiz interactions;
2. **Performance data** – grades, scores, and completion rates;
3. **Contextual data** – attendance, access to learning resources, and time spent on tasks.

By combining these data types, LA supports personalized learning strategies and evidence-based interventions.

4. Methodology

This study presents a proof-of-concept demonstrating how AI-based predictive modeling can be applied to Moodle activity data to identify students at risk. The primary objective is to establish a scalable workflow for data preprocessing, model training, evaluation, and visualization using trial data. While the current implementation relies on simulated data for demonstration purposes, the methodology is designed to be directly applicable to real student activity records collected within the COLOURS alliance project, enabling timely interventions and personalized learning strategies.

4.1. Data Collection and Description

To illustrate the predictive modeling workflow, a trial dataset of 1000 simulated students was generated to reflect typical Moodle usage patterns. The dataset includes features representing key learning behaviors, such as the number of assignments completed, hours spent studying, forum participation, attendance, and engagement with digital resources, including quizzes and learning materials. Each student is uniquely identified, and in addition to the activity metrics, two outcome variables were defined: **Passed**, a binary indicator of course success, and **Risk_Score**, a continuous measure estimating the likelihood of academic failure. Passed was calculated based on a weighted combination of key activity indicators, while Risk_Score was normalized between 0 and 1 to provide a continuous risk assessment. The simulated dataset ensures sufficient variability to demonstrate the model's capabilities and visualize differences in student risk levels, while serving as a conceptual proof-of-concept.

Table 2. Key Learning Behaviors in the Trial Dataset

Variable	Description
Student_ID	Unique identifier for each student
Assignments_Completed	Number of assignments submitted
Hours_Studied	Hours spent on learning activities
Forum_Posts	Number of posts in discussion forums
Attendance	Recorded attendance in classes or online sessions
Digital_Resource_Usage	Engagement with learning materials, quizzes, and simulations
Passed	Binary outcome indicating course success (1) or failure (0)
Risk_Score	Continuous measure estimating the probability of academic failure (0–1)

4.2. Data Preprocessing

The trial dataset was prepared to ensure compatibility with the Random Forest algorithms. All values in the simulated dataset are numeric and complete, so no missing data handling was required. Numeric variables are structured for direct input into the models, and categorical encoding is applied when necessary. This preprocessing ensures that the dataset is ready for machine learning analysis. When real Moodle data are used in the next phase of the project, additional steps will address missing values, inconsistencies, and anomalies inherent in real-world datasets. The current workflow establishes the preprocessing steps required for larger datasets, ensuring that the methodology is scalable and adaptable.

4.3. Model Training

The predictive modeling was performed using a Random Forest approach on the trial dataset. A Random Forest Classifier was used to predict whether a student is at risk of failing or successfully passing the course, while a Random Forest Regressor estimated the continuous Risk_Score for each student, representing the probability of academic failure. The dataset was split into a training set (70%) and a testing set (30%) to train the models and evaluate their performance on unseen data. This approach allows both the binary outcome and the continuous risk measure to be predicted from the same set of student activity indicators. For real Moodle data, the same methodology will be applied after proper data extraction, cleaning, and preprocessing.

4.4. Model Evaluation and Visualization

The performance of the predictive models was evaluated using standard statistical metrics and visualization techniques. For the classification task, accuracy, precision, recall, and F1-score were computed to assess how effectively the Random Forest Classifier distinguished between successful and at-risk students. The accuracy of the classifier was approximately 0.80, indicating that 80% of the predictions matched the actual outcomes. A confusion matrix was generated to visualize correct and incorrect predictions, providing insight into the balance between false positives and false negatives. The F1-score, reflecting the balance between precision and recall, was also computed and found to be around 0.80, demonstrating a reliable classification performance for identifying at-risk students.

For the regression task, the Random Forest Regressor was assessed using the Mean Squared Error (MSE) to measure the deviation between predicted and actual Risk_Score values. Lower MSE values indicate a better model fit and higher predictive reliability. Feature importance analysis was conducted for both models to identify which behavioral indicators had the greatest influence on prediction accuracy. This analysis revealed that variables such as assignments completed, hours studied, and attendance contributed most significantly to predicting student outcomes.

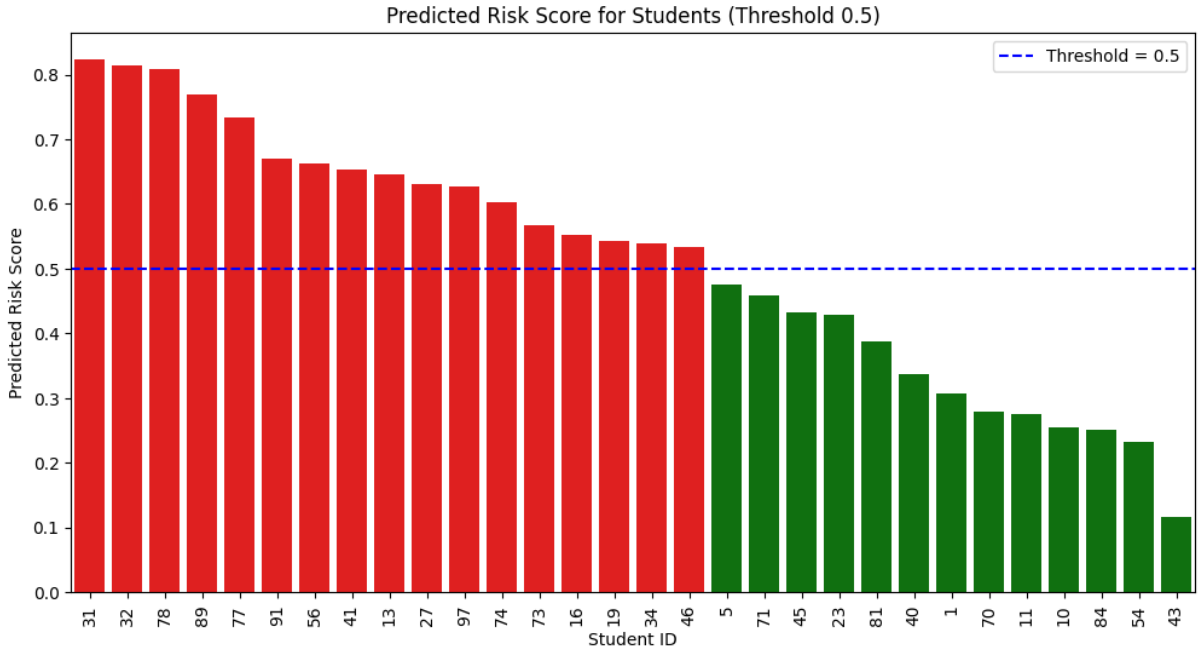
4.5. Integration with Moodle and LRS

In practical implementation, student activity data is collected directly from Moodle through a Learning Record Store (LRS), which acts as an intermediary that standardizes xAPI statements generated by Moodle activities. Each statement records key information about learner actions.

This approach allows the Random Forest model to access detailed behavioral data for predictive analysis without requiring direct integration with external analytics platforms like Google CoHub. Once extracted, the data from the LRS can be exported in a structured format (e.g., CSV or database table) and imported into Python in Google Colab for preprocessing, model training, and visualization.

By using the LRS as a central data collection layer, the workflow ensures compatibility with standard xAPI statements, facilitates scalable data extraction from Moodle, and maintains data integrity for AI-based predictive modeling. This method supports timely identification of at-risk students and allows educators to design personalized interventions based on reliable, real-time activity data.

Figure 1. Predicted Risk Scores for Students Using Random Forest Regression



This bar chart visualizes the predicted probability of academic failure (Risk Score) for each student, calculated using a Random Forest regression model. The model considers multiple indicators of student activity, including assignments completed, hours studied, forum participation, attendance, and engagement with digital resources. Students are sorted from highest to lowest predicted risk to highlight those who may require immediate attention.

The colours indicate the risk levels relative to a threshold of 0.5: red bars represent students with a high risk of academic failure (Risk Score > 0.5), while green bars represent students with a low risk (Risk Score ≤ 0.5). A horizontal blue dashed line marks the threshold at 0.5 for easy reference.

This visualization allows educators to quickly identify at-risk students and plan targeted interventions or personalized support strategies, thereby supporting timely measures to improve student outcomes. The chart was generated in Python using Matplotlib and Seaborn within Google Colab, demonstrating both the predictive capabilities of the Random Forest model and the practical application of AI-based analytics for educational data.

5. Conclusion

The integration of AI-based predictive modeling within the COLOURS Alliance framework represents a significant shift from reactive to proactive educational support. By leveraging the granularity of Moodle (and further more on the xAPI) data and the robust classification power of Random Forest models, this study demonstrates that Moodle activity, specifically assignment completion, study hours, and attendance, serve as a reliable predictor of student success. While the current proof-of-concept utilizes simulated data to validate the Python-based workflow, the infrastructure is now in place to transition to real-world datasets across the Alliance's interoperable platforms. However, as these models move toward implementation, it is essential to balance technical accuracy with ethical responsibility. To avoid the "labeling effect" identified in recent literature, future efforts must ensure that AI insights are used as a discreet background tool for educators rather than a "black box" that stigmatizes learners. Ultimately, this approach provides a scalable roadmap for European university alliances to foster a more personalized, challenge-based learning environment that proactively addresses academic risk not just before failure occurs, but to navigate it towards personalized successful journey.

Literature

1. Abuzinadah, N., Umer, M., Ishaq, A., Al Hejaili, A., Alsubai, S., & Eshmawi, A. A. (2023). Role of convolutional features and machine learning for predicting student academic performance from MOODLE data. PLoS ONE, 18(11), e0293061. https://doi.org/10.1371/journal.pone.0293061
2. Ayon, S. I., et al. (2024). AI-based algorithms for predicting academic achievements and recommending appropriate study plans. Journal of Computational and Cognitive Engineering.
3. Baker, R. S., & Smith, K. (2020). Artificial Intelligence in Education: Bringing it all together. OECD Education Working Papers, No. 218. OECD Publishing. https://doi.org/10.1787/f6131d08-en
4. Bigler, D. Hagel, G. & Becker, M. (2025). Enhancing Learning Analytics: H5P Results for Personalized Software Engineering Education. In Proceedings of the 6th European Conference on Software Engineering Education (ECSEE '25). Association for Computing Machinery, New York, NY, USA, 176–179. https://doi.org/10.1145/3723010.3723014
5. Choi, J., et al. (2025). Educational data mining and SHAP: Enabling detailed insights into student performance predictors. Frontiers in Education.
6. COLOURS European university alliance. (2025). Information about the European university alliance used as case-study. Retrieved on 1st December 2025 from https://colours-alliance.eu/
7. Duch, D., May, M., & George, S. (2024). Enhancing predictive analytics for students' performance in Moodle: Insight from an empirical study. Journal of Data Science and Intelligent Systems. https://doi.org/10.47852/bonviewJDSIS42023777
8. Holmes, W., Porayska-Pomsta, K., & Holstein, K. (2022). Ethics of AI in Education: Towards a Community-Wide Framework. International Journal of Artificial Intelligence in Education, 32, 504–526.
9. Kaensar, C., & Wongnin, W. (2023). Analysis and prediction of student performance based on Moodle log data using machine learning techniques. International Journal of Emerging Technologies in Learning (iJET), 18(10), 184–203.
10. Mahdi-Reza, B. Hanan, S. Aref, T. & Elham, A. (2024). Analyzing click data with AI: Implications for student performance and withdrawal. Frontiers in Education, 9, 1421479.

11. Malanchini, M., & Allegrini, A. (2024). The role of non-cognitive skills in academic achievement from childhood to adolescence. Nature Human Behaviour. https://doi.org/10.1038/s41562-024-01967-w
12. MoodleDocs. (2021). Using analytics. Moodle.org. https://docs.moodle.org/en/Using_analytics
13. Netex Learning. (2025, February 24). xAPI vs SCORM: The pros and cons for performance data. Netex Blog.
14. Sandeepa, A. G. R., & Mohottala, S. (2025a). Evaluation of Machine Learning Models in Student Academic Performance Prediction: A Comparative Study. Journal of Educational Data Mining, 17(1), 12-34. (Originally accessed via arXiv:2506.08047).
15. Sandeepa, A. G. R., & Mohottala, S. (2025b). Evaluation of machine learning models in student academic performance prediction using LMS logs. Journal of Al-Qadisiyah for Computer Science and Mathematics, 17(1).
16. Shou, Z., Xie, M., Mo, J., & Zhang, H. (2024). Predicting Student Performance in Online Learning: A Multidimensional Time-Series Data Analysis Approach. Applied Sciences, 14(6), 2451. https://doi.org/10.3390/app14062451
17. Sultanova, A., et al. (2024). Beyond Grades: Exploring the influence of non-cognitive skills on academic achievement in STEM disciplines. Frontiers in Education, 9, 1342516.
18. Van Rijsbergen, C. J. (1979). Information Retrieval (2nd ed.). Butterworth-Heinemann.
19. Yet Analytics. (2025, March 4). The value of an advanced xAPI enablement of Moodle: Considering in the context of higher education. Yet Analytics Articles.
20. Zhang, L., et al. (2025). AI-driven formative assessment and adaptive learning in data-rich environments. arXiv:2509.20369v1.
21. Zhao, B., & Zhou, J. (2024). Research hotspots and trends in digitalization in higher education: A bibliometric analysis from 2014 to 2024. Heliyon, 10(21), e39201.

Original Scientific Paper/Original naučni rad
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600035J

UDC/UDK: 004.8:378]:347.77(497.11)
004.8:378]:347.77(4-672EU)

Veštačka inteligencija, koristan pomoćnik ili pretnja plagijata: „Analiza regulatornih pristupa i etičkog okvira u Evropskoj uniji i u Srbiji"

dr Anila Jelesijević[1]

[1]Senior information management assistant and media analysts, Embassy of Switzerland in Serbia and Montenegro anila.jelesijevic@gmail.com

Apstrakt: Brza integracija veštačke inteligencije (VI) u obrazovni sistem dovela je u pitanje granicu između legitimne pomoći i akademskog plagijata VI. Ovaj rad analizira kako zakonodavstvo i etički okviri u Evropskoj uniji (EU) i u Srbiji regulišu ulogu VI kao produktivnog alata za podršku i kao potencijalnog izvora plagijata. Takođe se fokusira na teorijske debate naučnika o dvostrukoj prirodi VI, pri čemu neki od njih ističu učenje i kreativnost kada se VI koristi transparentno, i oni koji predlažu nekoliko strategija, ali i uputstava koje akademsko osoblje može da koristi za sprečavanje plagijata korišćenjem VI. Pored toga, ovaj rad je takođe uzeo u obzir podatke anketa o plagijatu VI na evropskim univerzitetima i podatke o stavovima studenata u Srbiji kada se govori o upotrebi VI.

Ovaj rad pokazuje da, iako upotreba VI u obrazovnom sistemu raste i u Evropi i u Srbiji, u Evropi, uprkos sveobuhvatnim regulatornim zakonima EU o VI, praktična primena i obrazovna adaptacija su još uvek nepotpuni, dok su u Srbiji propisi više u fazi razvoja strategije sa savetodavnim etičkim smernicama. Ono što nedostaje u regulativi i EU i Srbije jeste jasno razlikovanje kada je veštačka inteligencija alat plagijata, a kada ne.

Rad zaključuje da za definisanje uloge veštačke inteligencije nije dovoljno ne samo imati adekvatnu zakonsku regulativu, već i pedagoški i etički razvoj koji, pored omogućavanja kapaciteta korišćenja veštačke inteligencije, takođe povećava svest da veštačka inteligencija nije zamena za ljudsku kreativnost.

Ključne reči: veštačka inteligencija, Evropska unija, Srbija, legitimna pomoć, plagijat, regulatorni zakoni

Artificial Intelligence, a useful assistant, or a plagiarism threat: "Analysis of regulatory approaches and ethical framework in the European Union and in Serbia"

Abstract: The rapid integration of artificial intelligence (AI) into education system has challenged the line between the legitimate assistance and the academic plagiarism of the AI. This paper analyses how legislation and ethical frameworks in European Union (EU) and in Serbia regulate the role of AI in being a productive support tool and as a potential source of plagiarism. It also focuses on the theoretical debates of several scholars about the dual nature of AI with some of them highlighting the learning and creativity when AI used transparently and those who suggest few strategies but also instructions that academic staff can use to prevent plagiarism using AI. In addition to that, this paper has also taken into consideration surveys data of AI plagiarism in European universities and data of students' attitudes in Serbia when referring to the use of AI.

This paper shows that while the use of AI into education system is increasing both in Europe and Serbia, in Europe despite for the comprehensive EU's AI regulatory laws, the practical enforcement and educational adaptation is still incomplete while in Serbia regulations are more at the stage of evolving strategy with advisory ethical guidelines. What is missing in both EU and Serbia's regulative is a clear distinguishing of when AI is a tool of plagiarism or not.

The paper concludes that in defining the AI's role it is not sufficient not only to have the adequate legislation but also pedagogical and ethical development which besides enabling the capacities of the AI use also increase the awareness that AI' is not a substitution of the human creativity.

Keywords: artificial intelligence, European Union, Serbia, legitimate assistance, plagiarism, regulatory laws

1. Introduction

Artificial intelligence (AI) has in nowadays become a relevant part of our reality including our education. AI tools are used for text-generating, language translations, data analyses and research in thus assisting academics in their writings. In the meantime, those AI's capabilities raise the concern about authorship and plagiarism in challenging the academics with the necessity of distinguishing what is legitime or not but also the awareness that the AI can be a valuable assistant and a plagiarism tool. In the EU, the comprehensive AI regulatory laws do not regulate the issue of AI being a plagiarism tool while in Serbia despite aligning with EU digital and ethical standards, the regulations are still at an early stage of AI regulatory development. Therefore, both in EU and in Serbia, the educational institutions are apparently struggling to protect the creativity and knowledge from the AI misuse.

The main research question of this study is:

To what extend the current AI regulations in EU and Serbia enable AI to function as a tool of legitimate assistance and not of plagiarism and what is the role of the moral of the academics in this dilemma?

This study is based on comparative and qualitative research which consists of the analyses of documents and case studies reviews. It analyzes the legislative AI regulations, ethical guidelines but also refers to real examples from EU and Serbia universities in using AI in their academic writings while also focusing on theoretical views of several scholars on this topic.

This paper elaborates the following issues: what are the main theoretical concepts of this research; what do the theoretical debates say about the AI usage in academic writings and what do they advise, what are the regulations in the using of AI in the academic writings, in EU and in Serbia; what is the situation among the academics when the use of AI in their writings is at issue and how far can the academics go with the use of AI in our academic writings and how will our academic moral ethics comply with this.

2. Main theoretical concepts of the research

The elaboration of the main theoretical concepts is important to further understand the analyses and comments in this research. Therefore, it is relevant to explain what intelligence, artificial intelligence, Chat GTP and plagiarism are.

The "faculty of understanding or intellect" is what is called intelligence according to the Oxford English Dictionary which also define intelligence as "the mental capacity to understand" (Oxford English Dictionary, 2010). A more specific definition in also referring to intelligence as understanding is given by Brittanica dictionary in describing it as "the ability to learn or understand things or to deal with new or difficult situations" (The Brittanica Dictionary, 2025). In their publication, Shane Legg and Marcus Hutter while noting that "despite a long history of research and debate, there is still no standard definition of intelligence" (Legg and Hutter, 2007), give several definitions of what intelligence is. Out of over 30 definitions given by them, two were for us widely framed i.e., that for intelligence to be "The ability to use memory, knowledge, experience, understanding, reasoning, imagination and judgement in order to solve problems and adapt to new situations" (p.2), while quoting AllWords Dictionary, 2006, and "the general mental ability involved in calculating, reasoning, perceiving relationships and analogies, learning quickly, storing and retrieving information, using language fluently, classifying, generalizing, and adjusting to new situations" (p.2), while quoting Columbia Encyclopedia, sixth edition, 2006. Legg and Hutter also refer to several definition of intelligence as described by AI researches quoting also J.S. Albus saying that intelligence is the "the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success, and success is the achievement of behavioral subgoals that support the system's ultimate goal" (p.3), and D. Fogel defining intelligence as "Any system that generates adaptive behavior to meet goals in a range of environments can be said to be intelligent." (p.7).

Regarding AI, there are also a lot of definitions. The Oxford English Dictionary defines AI as "the capacity of computers or other machines to exhibit or simulate intelligent behavior" (Oxford English Dictionary, 2010). In the meantime, while noting that there is "no single, simple definition of artificial intelligence because AI tools are capable of a wide range of tasks and outputs," (NASA, 2024) National Aeronautics and Space Administration (NASA) defines AI as "computer systems that can perform complex tasks normally done by human-reasoning, decision making, creating, etc" (NASA, 2024). Another definition of relevance is that of International Business Machines Corporation (IBM) officials who describe AI as "technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy. (IBM,

2025). Additionally, ChatGPT is an AI tool. It is about a variant of the GPT-3 (Generative Pre-trained Transformer 3) artificial intelligence language model of OpenAI company introduced in 2021 (Cotton-Cotton et al. 2023).

After having clarified of what intelligence and artificial intelligence mean, it is important to understand what plagiarism is. A compressive definition is given by University of Oxford in the United Kingdom which while defining plagiarism also mentions AI as a possible tool of plagiarism. According to this definition, "presenting work or ideas from another source as your own, with or without consent of the original author, by incorporating it into your work without full acknowledgement. All published and unpublished material, whether in manuscript, printed or electronic form, is covered under this definition, as is the use of material generated wholly or in part through use of artificial intelligence (save when use of AI for assessment has received prior authorisation e.g. as a reasonable adjustment for a student's disability). Plagiarism can also include re-using your own work without citation. Under the regulations for examinations, intentional or reckless plagiarism is a disciplinary offence" (University of Oxford, 2025).

If the intelligence is generally perceived as the ability of understanding, reflecting and acting, AI is usually defined as computer systems and technology that can perform a wide range of human tasks including assessments. It is important to notice that AI is not a human and is created by the humans to perform the tasks they want. For the time being, AI relies on being programmed by humans and interacts based on the retrieved data. The use of AI in the academic field can lead to plagiarism if there is no specified source of reference in the retrieved data under the instructions of those who aim on writing their works and /or ideas.

The issue of how and when to use the AI in academic writings thus brings us to debates of different scholars which this study will be dealing with in the next chapter entitled "Theoretical debates about the AI usage in academic writings".

3. Theoretical debates about the AI usage in academic writings

Debates about the use of the AI in the academic writings have been consisting of concerns but also of enthusiasm. The concerns are mainly based on the expressed stands on the possible misuse of the AI as a plagiarism tool which at the same time affect negatively in the creativity and in the critical thinking of the authors of the academic writings. On the other side, the enthusiasm is expressed via the belief that AI can increase the productivity of the academic writings. Out of several views on dealing with such a topic, this study has focused on two of them which have not only dealt with the optimism of the use of AI also in academic writings and research but have also taken into consideration the possibility of AI's misuse in even suggesting strategies on how to prevent it.

Rosemary Luckin, Professor Emerita at University College London and Founder and CEO of Educate Ventures Research Limited (EVR), with over 30 years of experience and recognized expert on AI in education (University College London, 2025), while not denying all of great benefits AI could bring to the learners has pointed out that it is relevant to pursue with the education of general public with a key section of being that the educators needs to understand more about the AI, what it can do and how it works (Rulph, 2024). While acknowledging as problems that "the regulation and the code of practice will never keep up with what the technology is able to do" (p.325) and that "huge assumptions will be made about, for example, what it means to be transparent" (p.350), professor Luckin has spoken about seven steps to the AI readiness that are connected to ethical behavior which include education, tailoring, identification, collection, applying, learning and iterating as described in the below given table:

Table 1: The seven steps to AI readiness according to Rosemary Luckin:

Source: (Rulph, 2024, p.351).

Despite being enthusiastic that "within academia, generative AI will likely enhance academic productivity through automated basic research and writing assistance" (p.361), professor Luckin has noted that "human skills like conceptualization, creativity, complex critical analysis, judgment, social perceptiveness, and wisdom will become even more valuable among academics" (p.361). Moreover, she has expressed her skepticism that "there will be an AI that can do everything a human can do, and all of it will be better than a human" (p.362).

In the meantime, in the paper entitled "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT", the authors, Cotton, Debby R. E., Cotton, Pete A. and Shipway J. Reuben besides examining the opportunities and challenges of using ChatGPT in higher education and discussing the potential risks and rewards of these tools also acknowledge that there are difficulties of detecting and preventing academic dishonesty (Cotton-Cotton et al. 2023). Additionally, the authors have suggested what could be the strategies that the universities could perform to provide not only ethical but also a responsible use of the AI tools. As advantages of the AI in the academic research and writing, those authors mention the following: AI as a platform for asynchronous communication, as a facilitator of collaboration among the students as well as a tool for remote learning (p. 229). On the other hand, as one of the challenges while using AI's ChatGPT in writing is the possibility of plagiarism with students submitting works that are not their own (p. 230). In order for the AI to be prevented from the misuse, the referred authors have suggested several strategies that the academic institutions can use such as: educating students on plagiarism as one of the most effective ways to prevent plagiarism; requiring students to complete a written declaration stating that their work is their own and that they have not used any AI language models to generate it; consider investing in advanced technology and techniques to detect the use of AI language models; set clear guidelines for use of GPT and other resources; check for sources and citations as chatbots are not capable of conducting original research or producing new idea (p. 232).

In Serbia, according to Danijela Vranješ, a teaching assistant of German language at the Faculty of Philology, University of Belgrade, the AI tools should not be used to write full seminar papers, but they could help with the compiling of bibliographies or citing sources which is "a boring and exhausting job" (Kljajić, 2025).

It is evident that the use of the AI in academic writings cannot be excluded nor there is a will to do it bearing in mind that it is about a tool that facilitates the referred process. However, the ethics must be respected in thus raising the awareness about AI being a possible tool of plagiarism. Forms of education are necessary which are also accompanied by regulations on how the AI can be used or not at to what extent. Therefore, it is relevant to analyze what the current regulations in EU and in Serbia say about the use of the AI in the academic writings which has been elaborated in the next coming chapter.

4. Regulations in the using of AI in the academic writings, in EU and in Serbia

The European Union has issued a set of documents which deal with the general use of the AI. In 2024, it introduced the AI Act Explorer document for which says to be a new regulation on artificial intelligence and considers it as a layer of the foundations for the regulations of the ai in the EU (EU Artificial Intelligence Act, 2024). It is about a document which consists of 13 chapters with each of them containing set of articles which mainly regulate the issues if prohibited AI practices, high risk AI systems, transparency obligations for the providers and developers of certain AI systems, codes of conducts etc. The article that is interesting for our study is article 5 which deals with prohibited AI practices which anticipates the prohibition of the following AI practices in cases such: the placing on the market and the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques; that exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, for the evaluation or classification of natural persons or groups of persons over a certain period of time; for making risk assessments of natural persons in order to assess or predict the risk of a natural person committing a criminal offence; to infer emotions of a natural person in the areas of workplace and education institutions (EU Artificial Intelligence Act, Article 5, 2024). As it can be seen is about situations which could harm the rights and identity of the people and eventually exert pressures on them. This document does not deal with the issue of plagiarism in the academic writings. Another EU issued document dealing with AI is the one issued in 2025 by European Union Intellectual Property Office entitled "The development of generative artificial intelligence from a copyright perspective" (European Union Intellectual Property Office, 2025). This over 400 pages document is designed to clarify how Generative AI systems interact with copyright technically, legally, and economically (p.3). It also notes that the "EU was the first jurisdiction in the world to adopt a comprehensive legislation on the regulation of AI technologies, in the form of the Regulation (EU) 2024/1689, commonly referred to as the 'AI Act', adopted in June 2024" (p.20), while referring to the AI Act Explorer document which we analyzed in the beginning of this chapter. This document does not deal either with the usage of AI as a possible tool of plagiarism.

Another publication issued in 2025 from the European Commission Directorate-General for Research and Innovation deals with the living guidelines on the responsible use of generative AI in research (European Commission Directorate General for Research and Innovation, 2025). It notes that the "The European Research Area Forum (composed of European countries and research and innovation stakeholders), decided to develop guidelines on the use of generative AI in research for: funding bodies, research organisations and researchers, both in the public and private research ecosystems" (p.4). While considering these guidelines non-binding it is pointed out that they should be considered as a supporting tool for researchers, research organizations and research funding bodies (p.4). As noted, the set of principles framing these guidelines are based on pre-existing relevant frameworks: the European Code of Conduct for Research Integrity and on the work and guidelines on trustworthy AI, developed by the High-Level Expert Group on AI (p.5). This document does not deal with plagiarism and the use of AI but however it gives inputs for a responsible use of the AI in research in emphasizing the among other things the "honesty in developing, carrying out, reviewing, reporting and communicating on research transparently, fairly, thoroughly and impartially" (p.5). It also mentions the Ethics Guidelines for Trustworthy AI of the EU High-Level Expert Group on AI which point out four ethical principles for AI systems: 1. respect for human autonomy; 2. prevention of harm; 3. fairness; 4. Explicability (p.12).

Besides having analyzed the above-mentioned regulations, it is useful to understand how several academic institutions in the EU countries deal with the use of AI as a possible tool of plagiarism. The Belgium based KU Leuven university allows the use of generative AI (GenAI) concerning education and research and even encourage students and teaching staff to handle this technology (KU Leuven, 2025). Nevertheless, this institution requires for several guidelines and principles to be followed while using GenAI in highlighting transparency, verifications of correctness of generated output, respect for copyrighted material, personal data and confidential information as well as responsibility for the correct use of GenAI (KU Leuven, 2025). It is explicitly notes the risk of plagiarism in the GenAI's output which the transparency about the sources is sometimes absent (KU Leuven, 2025). The University of Siena says to be the first one in Italy to define the use of ChatGPT and other LLM (Large Language Models) by drafting guidelines to guide the academic community in discussing and exploring new ways of teaching and research, and in activating behaviors that foster responsibility and awareness of actions (Universita di Siena, 2023). The University of Luxembourg has also set the guidelines for the use of AI in foreseeing disciplinary procedure for academic fraud and plagiarism case of a substantiated suspicion of unauthorized AI use in an assessment (Universite du Luxembourg, 2025). A more restrictive policy appears to be that of the Science PO, one of the one of France's top universities which in 2023 decided to ban the use of ChatGPT, an artificial intelligence-based chatbot that can generate coherent prose, to prevent fraud and plagiarism (EURONEWS, 2023).

Regarding Serbia, in January 2025, it adopted the Strategy for the Development of Artificial Intelligence in the Republic of Serbia for the period from 2025 to 2030 after the former one which is said to have set the foundation of the development of the AI in Serbia (Ministarstvo nauke, tehnološkog razvoja i inovacija, 2025). The strategy notes that "regarding the development of ethical and safe artificial intelligence, it is important to emphasize that the Ethical Guidelines have been adopted, developed in accordance with UNESCO recommendations and the recommendations of the European Union" (p. 14). The document recommends that the state administrations bodies and holders of public authority apply the Ethical Guidelines when developing, implementing and using systems that can be classified as artificial intelligence systems or their procurement" (p.14). The strategy has also specified that "The Ethical Guidelines require conditions for the creation of reliable and responsible artificial intelligence, which include operation and oversight, technical reliability and security, privacy, personal data protection and data management, transparency, diversity, non-discrimination and equality, social and environmental well-being and responsibility" (p.14). This document does not deal with the issue of AI and plagiarism. Additionally, the 2019 Law on Copyright and Related Rights of the Republic of Serbia (Pravno informacioni sistem Republike Srbije, 2019) which is still in force, does not regulate either the plagiarism while using AI as it can be assumed that topic was not so actual nine year ago.

When the Serbia's universities are at issue, this study has referred to the policy of the biggest university in this country, the University of Belgrade. In the rulebooks there are no guidelines and/or rules in the case of the use of AI in the academic writings. The University of Belgrade in the "Rulebook on the procedure for determining ethical responsibility at the University of Belgrade" adopted in 2021 foresees fabricating and rewriting recommendations or misrepresenting academic achievements as one of the forms of violations of the Code of Professional Ethics (Univerzitet u Beogradu, 2021) but it does not deal with the AI. This university has also an adopted rulebook on the determining of the nonacademic behavior in the preparation of the written works also dating since the year 2021. This document anticipates for the authorized Commission to commit the evaluation of the originality of the written works taking into consideration the result of the software analyses in not specifying either anything about the AI usage (Univerzitet u Beogradu, 2021).

There are not yet regulations in the EU and in Serbia which clearly specify how to handle the academic writings which are assisted by the AI and specify the cases of plagiarism. Additionally, the examples of how several academic institutions in the EU countries and in Serbia deal with the use of AI shows that while there is no unique policy of the universities referring to the use of AI in academic writings in EU countries, in Serbia AI is not foreseen neither as an assisting or plagiarism tool.

The academics have apparently no other choice except to get in detail informed about the rules of the institutions that they are part of and rely on their moral ethics while using the AI. How this looks in practice is elaborated in the next chapter: The practice of the use of AI in academic writings in EU an in Serbia.

5. The practice of the use of AI in academic writings in EU and in Serbia

In 2023, scholars from university in Germany, presented their research which was focused on the use of AI among the students in that country. In the referred study, over 6300 students participated in an anonymous survey with almost two-thirds (63.4%) of them having stated that they have used AI tools for their studies. (Von Garrel and Mayer, 2023). The survey also showed that the students mostly used AI-based tools in clarifying questions of understanding and explaining subject-specific concepts research and literature study, translations, text analysis, text processing, text creation as well as for problem-solving, decision-making (Von Garrel and Mayer, 2023). The survey did not contain any data nor question about the eventually use of AI as a plagiarism tool but it however shows the frequent use of the AI as a relevant assisting tool of the students during their studies.

A more recent survey dating from mid-2025 from scholars from the universities of France and Spain analyzed the relationship between the use of AI and plagiarism in higher education. 503 university students from Spain were included in the survey being asked to complete a set of questionnaires (Campo- Delgado et al., 2025). The findings indicated that there is a correlation between the frequency of use of ChatGPT and Plagiarism, but the causality however was not found (Campo- Delgado et al., 2025).

In Serbia, the AI tools are also used in the academic writings and research. While there is no survey which could confirm the use of AI as a plagiarism tool, over 80% of the students in Serbia are reported to use AI tools such as ChatGPT assisting in learning but also for the writing of seminar and other academic works (VREME.COM, 2025). Additionally, Miloš Stojadinović from the Department of Psychology in Serbia researched with his colleagues the use of artificial intelligence in different ways in scientific works from 2000 to the present day (Medijski istraživački centar Niš, 2024). They reviewed almost 500 papers and according to the first preliminary findings, 66 papers were singled out that fit the topic of the implementation of artificial intelligence in the so-called STEM education (Science Technology Engineering Mathematics), where AI is most often used. According to Stojadinović, 50% of the works that were selected were created in the last three years, which means that the trend of using AI in education has grown significantly (Medijski istraživački centar Niš, 2024).

While it is easy to find and rely on data on the use of AI as an assisting tool in the academic writings, it is difficult to establish at what measure AI issued as a tool of plagiarism at least for two reasons: the users might not be aware that the used content could be a plagiarism; the users are aware of having used contents that could be subjected to plagiarism but they believe that they have inserted the necessary transformation form that can be recognized as their own creation.

6. Conclusion: The perspective of the use of AI in academic writings in compliance with the regulations and academic moral ethics

The study showed that the perception of the AI as an assistant or a plagiarism tool depends on the existing regulations in EU and in Serbia about the AI including those issued by the states of the universities as well as by the moral ethics of academics while using the AI. While there is a need for the exiting regulations both in the EU countries and in Serbia to be more precise with the rules on how to use AI in academic writings and when it is considered a plagiarism, it is also relevant for the awareness of the academics to be raised in understanding of what the real assets of the use of AI are. When accompanied by transparency and human monitoring, which can be the protecting mechanism s of the intellectual dishonesty, AI can be considered as a legitimate useful tool in facilitating the academic research and writing process. The use of AI in the academic research and writings is expected to continue and most probably further increased in the future. At the same time, it must be ensured that it remains an assisting legal tool and not a substitution of human creativity.

Literature

1. Campo, L., Delgado, N., Urbieta, E., & Kanso H. (2025). Relationship Between the Use of ChatGPT and Plagiarism in Higher Education: The Influence of Gender, Age and Previous Academic Results. Retrieved on 11.12.2025 from https://www.researchgate.net/publication/393407427_Relationship_Between_the_Use_of_ChatGPT_and_Plagiarism_in_Higher_Education_The_Influence_of_Gender_Age_and_Previous_Academic_Results
2. Cotton, D., Cotton, P., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT'. Innovations in Education and Teaching International. Vol. 61, No.2. Retrieved on 04.12.2025 from https://www.tandfonline.com/doi/epdf/10.1080/14703297.2023.2190148?needAccess=true
3. EU Artificial Intelligence Act (2024). The AI Act Explorer. Retrieved on 09.12.2025 from https://artificialintelligenceact.eu/ai-act-explorer/
4. EURONEWS (2023). Top French university bans use of ChatGPT to prevent plagiarism. Retrieved on 10.12.2025 from https://www.euronews.com/next/2023/01/28/france-chatgpt-university?utm_source=chatgpt.comEuropean Commission Directorate General for Research and Innovation (2025). Living guidelines on the responsible use of generative AI in research. Retrieved on 09.12.2025 from https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf&utm_source=chatgpt.com
5. European Union Intellectual Property Office (2025). The development of generative artificial intelligence from a copyright perspective. Retrieved on 09.12.2025 from https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/observatory/documents/reports/2025_GenAI_from_copyright_perspective/2025_GenAI_from_copyright_perspective_FullR_en.pdf
6. IBM (2025). What is artificial intelligence (AI) Retrieved on 04.12.2025 from https://www.ibm.com/think/topics/artificial-intelligence
7. Kljajić, K. (2025). Nekad je teško prepoznati ko je napisao sastav, učenik ili veštačka inteligencija
8. Retrieved on 11.12.2025 from https://www.nin.rs/bbc/vesti/75500/nekad-je-tesko-prepoznati-ko-je-napisao-sastav-ucenik-ili-vestacka-inteligencija?utm_source=chatgpt.com
9. KU Leuven (2025). Responsible use of generative artificial Intelligence. Retrieved on 10.12.2025 from https://www.kuleuven.be/english/genai?utm_source=chatgpt.com
10. Legg, Sh. & Hutter, M. (2007). A collection of definitions of intelligence. Retrieved on 04.12.2025 from https://www.researchgate.net/publication/1895883_A_Collection_of_Definitions_of_Intelligence
11. NASA (2024). What is artificial intelligence. Retrieved on 04.12.2025 from https://www.nasa.gov/what-is-artificial-intelligence/
12. Oxford English Dictionary (2025). Artificial intelligence. Retrieved on 04.12.2025 from https://www.oed.com/dictionary/artificial-intelligence_n?tl=true#128454816
13. Oxford English Dictionary. (2010) Intelligence. Retrieved on 04.12.2025 from https://www.oed.com/dictionary/intelligence_n
14. Pravno informacioni sistem Republike Srbije (2019). Zakon o autorskim i srodnim pravima. Retrieved on 10.12.2025 from https://pravno-informacioni-sistem.rs/eli/rep/sgrs/skupstina/zakon/2009/104/30/reg
15. Republika Srbija. Ministarstvo nauke, tehnološkog razvoja i inovacija (2025). Usvojena Strategija za razvoj veštačke inteligencije u Republici Srbiji za period od 2025. do 2030. godine. Retrieved on 10.12.2025 from https://nitra.gov.rs/lat/ministarstvo/vesti/usvojena-strategija-za-razvoj-vestacke-inteligencije-u-republici-srbiji-za-period-od-2025-do-2030-godine
16. Rulph, J. (2024). Exploring the future of learning and the relationship between human intelligence and AI. An interview with Professor Rose Luckin. Journal of Applied Learning & Teaching Vol.7 No.1. Retrieved on 08.12.2025 from https://journals.sfu.ca/jalt/index.php/jalt/article/view/1659/767
17. The Brittanica Dictionary (2025). Intelligence. Retrieved on 04.12.2025 from https://www.britannica.com/dictionary/intelligence
18. University College London (2025). Rosemary Lucking. BIO. Retrieved on 08.12.2025 from https://profiles.ucl.ac.uk/48663-rose-luckin
19. Universita di Siena. (2023). Pubblicata la policy per l'utilizzo in Ateneo dei sistemi di intelligenza artificiale generative. Retrieved on 10.12.2025 from

20. Universite du Luxembourg (2025). Guidelines on the use of Generative AI tools for learning and teaching. P5. Retrieved on 10.12.2025 from https://www.uni.lu/wp-content/uploads/sites/9/2025/10/13172459/2025-guidelines-AI-3.pdf
21. University of Oxford (2025). Plagiarism. Retrieved on 04.12.2025 from https://www.ox.ac.uk/students/academic/guidance/skills/plagiarism#:~:text=New%20students-,Plagiarism,your%20work%20without%20full%20acknowledgement
22. Univerzitet u Beogradu (2021). Pravilnik o postupku utvrđivanja etičke odgovornosti na Univerzitetu u Beogradu. P 2. Retrieved on 10.12.2025 from https://bg.ac.rs/files/sr/univerzitet/univ-propisi/Pravilnik-postupak-eticka-odgovornost2021.pdf
23. Univerzitet u Beogradu (2021). Pravilnik o postupku utvrđivanja neakademskog ponašanja u izradi pisanih radova. Article 5. Item 4. Retrieved on 10.12.2025 from https://bg.ac.rs/files/sr/univerzitet/univ-propisi/Pravilnik_neakademsko_ponasanje_pisani_radovi2021.pdf
24. Von Garrel, J. & Mayer J. (2023). Artificial Intelligence in studies—use of ChatGPT and AI-based tools among students in Germany. Retrieved on 08.12.2025 from https://www.nature.com/articles/s41599-023-02304-7?utm_source=chatgpt.com
25. VREME.COM (2025). Pitali smo ChatGPT – koliko ljudi u Srbiji koristi ChatGPT. Retrieved on 11.12.2025 from https://vreme.com/mozaik/pitali-smo-chatgpt-koliko-ljudi-u-srbiji-koristi-chatgpt/
26. Medijski istraživački centar Niš (2024). Mi spremamo učenike za svet koji tek dolazi. Retrieved on 11.12.2025 from https://www.mic.org.rs/drustvo/item/1485-mi-spremamo-ucenike-za-svet-koji-tek-dolazi?utm_source=chatgpt.com

Original Scientific Paper/Original naučni rad UDC/UDK: 004.93:[004.6:343.982
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600043Z

Transformatori vida za generisanje ugrađivanja: Evaluacija nad CASIA skupom podataka

Miloš Živadinović[1], Bojan Jovanović[2]

[1] AikBank, milos.zivadinovic@aikbank.rs
[2] Faculty of Organizational Sciences, bojan.jovanovic@fon.bg.ac.rs

Apstrakt: Sistemi za prepoznavanje otisaka prstiju tradicionalno se oslanjaju na metode ekstrakcije karakteristika zasnovane na minucijama ili CNN-u. Ovaj rad istražuje primenu transformatora vida (ViT) za generisanje diskriminativnih ugrađivanja otisaka prstiju. Koristimo standardnu ViT arhitekturu sa veličinom delova 16×16, prvobitno dizajniranu za klasifikaciju prirodnih slika, i prilagođavamo je za biometrijsko kodiranje otisaka prstiju koristeći optimizaciju gubitka tripleta. Naš pristup tretira slike otisaka prstiju kao nizove delova, koristeći mehanizam samopažnje transformatora za snimanje lokalnih obrazaca grebena i globalne strukture otiska prsta. Evaluiramo naš metod nad CASIA skupom podataka otisaka prstiju, sprovodeći eksperimente verifikacije sa 114 originalnih parova i 114 parova uljeza. Rezultati pokazuju snažne performanse, postižući ROC AUC od 0,9899 i jednaku stopu grešaka (EER) od 4,82%. Ovi rezultati ukazuju da transformatori vida, bez arhitektonskih modifikacija specifičnih za otisak prsta, mogu efikasno da nauče diskriminativna ugrađivanja za prepoznavanje otisaka prstiju. Uspeh ovog pristupa sugeriše da arhitekture zasnovane na transformatorima predstavljaju održivu alternativu konvencionalnim metodama, otvarajući nove pravce za ekstrakciju biometrijskih karakteristika korišćenjem mehanizama pažnje.

Keywords: Transformatori vida, prepoznavanje otisaka prstiju, biometrijska ugrađivanja, CASIA skup podataka

Vision Transformers for Fingerprint Embedding Generation: Evaluation on CASIA Dataset

Abstract: Fingerprint recognition systems traditionally rely on minutiae-based or CNN-based feature extraction methods. This paper investigates the application of Vision Transformers (ViTs) for generating discriminative fingerprint embeddings. We employ standard ViT architecture with 16×16 patch size, originally designed for natural image classification, and adapt it for fingerprint biometric encoding using triplet loss optimization.
Our approach treats fingerprint images as sequences of patches, leveraging the self-attention mechanism of transformers to capture both local ridge patterns and global fingerprint structure. We evaluate our method on the CASIA Fingerprint Database, conducting verification experiments with 114 genuine pairs and 114 impostor pairs. The results demonstrate strong performance, achieving a ROC AUC of 0.9899 and an Equal Error Rate (EER) of 4.82%.
These results indicate that Vision Transformers, without fingerprint-specific architectural modifications, can effectively learn discriminative embeddings for fingerprint recognition. The success of this approach suggests that transformer-based architectures represent a viable alternative to conventional methods, opening new directions for biometric feature extraction using attention mechanisms.

Keywords: Vision Transformers, Fingerprint Recognition, Biometric Embeddings, CASIA Dataset

1. 1. Introduction

Fingerprint recognition has been a key implementation of biometric authentication due to the the uniqueness of fingerprints. Traditional approaches to fingerprint recognition have evolved from minutiae-based methods (Maltoni et al. 2022) to more sophisticated deep learning approaches utilizing convolutional neural networks (CNNs) (Simonyan and Zisserman 2015). While these methods have achieved large levels of success, they often require domain-specific architectures and preprocessing steps to function properly.

The emergence of Vision Transformers (ViTs) (Dosovitskiy et al. 2021) has improved computer vision research by allowing self-attention mechanisms, originally developed for natural language processing (Vaswani et al.

2017), to be effectively applied for image understanding. ViTs process images by dividing them into fixed-size patches and treating each patch as a token in a sequence. This allows the model to learn relationships between different spatial regions of the image through multi-head self-attention. This approach has shown state of the art performance compared to CNNs on various image classification benchmarks.

Despite the success of ViTs in general computer vision tasks, their application to biometric recognition remains relatively unexplored. Fingerprints present unique challenges that differ from natural images: they exhibit repetitive patterns such as ridges and valleys, contain critical discriminative features (from individual minutiae to overall ridge flow patterns) and require high precision in matching to ensure true positives.

This paper investigates whether Vision Transformers, without fingerprint-specific modifications, can effectively generate discriminative embeddings for fingerprint recognition. We adopt a learning approach using triplet loss allowing the model to learn an embedding space where fingerprints from the same finger are closer together than fingerprints from different fingers. We use the CASIA Fingerprint Database to provide empirical evidence of the viability of this approach.

The contributions of this paper are the following: (1) we demonstrate that the original ViT architecture can be successfully applied to fingerprint embedding generation, (2) we provide quantitative results on a public benchmark dataset, and (3) we analyze the performance characteristics of transformer-based embeddings in the context of biometric verification.

2. Related Work

2.1. Traditional Fingerprint Recognition

Fingerprint recognition has a history spanning several decades. Early approaches focused on minutiae-based matching (Hong, Wan, and Jain 1998), where specific features such as ridge endings and bifurcations are extracted and compared. These methods are sensitive to image quality and require robust preprocessing pipelines including segmentation, enhancement, and binarization. The reliance on handcrafted features limits their ability to capture the full complexity of fingerprint patterns.

Other traditional approaches include texture-based methods that analyze ridge orientation (Ratha, Chen, and Jain 1995) and frequency patterns and correlation-based techniques that directly compare fingerprint images. While these methods have been successful in controlled environments, they often struggle with variations in image quality, rotation, and partial fingerprints (Jain, Hong, and Bolle 1997).

2.2. Deep Learning for Fingerprint Recognition

The advent of deep learning has transformed fingerprint recognition by enabling end-to-end learning of discriminative features. Convolutional Neural Networks (CNNs) have been extensively applied to various aspects of fingerprint analysis including quality assessment, enhancement (Cao and Jain 2019), minutiae extraction (Tang et al. 2017), and direct matching (Minaee et al. 2021). CNN-based approaches have demonstrated superior performance compared to traditional methods, particularly in handling poor quality images and partial fingerprints.

Recent work has focused on learning compact embeddings that capture the essential characteristics of fingerprints while being invariant to variations in pose, pressure, and sensor characteristics.

2.3. Vision Transformers

Vision Transformers introduced by Dosovitskiy et al. (Dosovitskiy et al. 2021) apply the transformer architecture to image recognition by treating images as sequences of patches. Unlike CNNs, ViTs use self-attention mechanisms to model relationships between all patches in an image, enabling them to capture both local and global dependencies without the structure imposed by convolutional layers.

The resulting sequence of embeddings created by the ViT is then processed through multiple transformer encoder layers, each consisting of multi-head self-attention and feed-forward networks. This architecture has achieved state-of-the-art results on various image classification benchmarks when trained on large-scale datasets.

While ViTs have been extensively studied for general computer vision tasks, their application to biometric recognition remains limited (Li and Zhang 2023). This gap in literature motivates our investigation of whether the attention mechanisms in ViTs can effectively capture the patterns in fingerprints.

2.4. Metric Learning and Triplet Loss

Metric learning (Kaya and Bilge 2019) aims to learn an embedding space where similar instances are close together and dissimilar instances are far apart. This paradigm is particularly well-suited for biometric recognition, where the goal is to verify whether two samples come from the same individual rather than classify them into a fixed number of categories.

Triplet loss (Kulis 2013), introduced for face recognition, has become a popular choice for learning biometric embeddings. The loss operates on triplets of samples: an anchor, a positive (same identity as anchor), and a negative (different identity). The objective is to ensure that the distance between the anchor and positive is smaller than the distance between the anchor and negative by at least a margin. This formulation directly optimizes the embedding space for verification tasks, making it an ideal choice for our fingerprint recognition application.

3. Methodology

3.1. Dataset

We conducted our experiments using the CASIA Fingerprint Image Database (Fingerprint Databases----Institute of Automation n.d.), a publicly available fingerprint dataset. The CASIA database contains fingerprint images captured under controlled conditions, providing a standardized benchmark for evaluating fingerprint recognition algorithms.

For our evaluation, we constructed a verification test set consisting of 114 genuine pairs (images from the same finger) and 114 impostor pairs (images from different fingers). This balanced dataset allows for unbiased evaluation of the model's ability to discriminate between genuine and impostor matches. The images were preprocessed by resizing to the standard ViT input size while maintaining aspect ratio through appropriate padding.

3.2. Model Architecture

We employed standard Vision Transformer architecture with 16×16 patch size, referred to as ViT-Base/16. This architecture was originally proposed for natural image classification and has demonstrated strong performance across various computer vision benchmarks. The model consists of 12 transformer encoder layers, each with 12 attention heads and a hidden dimension of 768 parameters.

The input fingerprint images are divided into 16×16 pixel patches, which are then linearly embedded to create patch embeddings. A learnable positional encoding is added to each patch embedding to preserve spatial information. The sequence of embeddings is processed through the transformer encoder layers, and the final representation is extracted from the classification token (CLS token) that is prepended to the sequence.

We did not modify the ViT architecture to incorporate fingerprint-specific features. This allows us to evaluate whether the general-purpose vision transformer can capture fingerprint characteristics through data-driven learning alone, without the need for specialized architectural components.

3.3. Training Procedure

During training, each batch consists of multiple fingerprint images, and triplets are formed dynamically by selecting an anchor image, a positive image (from the same finger), and a negative image (from a different finger). The triplet loss encourages the model to minimize the distance between anchor and positive embeddings while maximizing the distance between anchor and negative embeddings, subject to a margin parameter.

We initialized the model with weights pre-trained on ImageNet (Deng et al. 2009), a practice that provides a strong initialization for vision transformers. The model was then fine-tuned on the fingerprint training data using the triplet loss objective. This transfer learning approach leverages the general visual features learned from natural images while adapting them to the specific characteristics of fingerprints.

3.4. Evaluation Metrics

For each pair of fingerprint images in the test set, we computed the Euclidean distance between their corresponding embeddings generated by the trained ViT model. These distances were then used to classify pairs as either genuine or impostor matches.

The primary evaluation metrics employed were:

- ROC AUC (Receiver Operating Characteristic Area Under Curve): This metric measures the model's ability to discriminate between genuine and impostor pairs across all possible decision thresholds. A perfect classifier achieves an AUC of 1.0, while random guessing yields 0.5.
- AUPRC (Area Under Precision Recall Curve): With AUPRC we handle the ratio of precision compared to recall of the tested data. With the value of 0.9718 we can conclude that we maintain high precision throughout the testing set (low number of false matches).
- Equal Error Rate (EER): The EER represents the point where the false acceptance rate equals the false rejection rate. It provides a single-number summary of system performance and is commonly used in biometric system evaluation. Lower EER values indicate better performance.

4. Experimental Results

Our experiments demonstrate that Vision Transformers can effectively generate discriminative embeddings for fingerprint recognition. The evaluation on the CASIA Fingerprint Database test set provided strong performance across both metrics, indicating that the self-attention mechanism successfully captures the characteristics of fingerprint patterns.

Figure 1: ROC AUC curve behavior

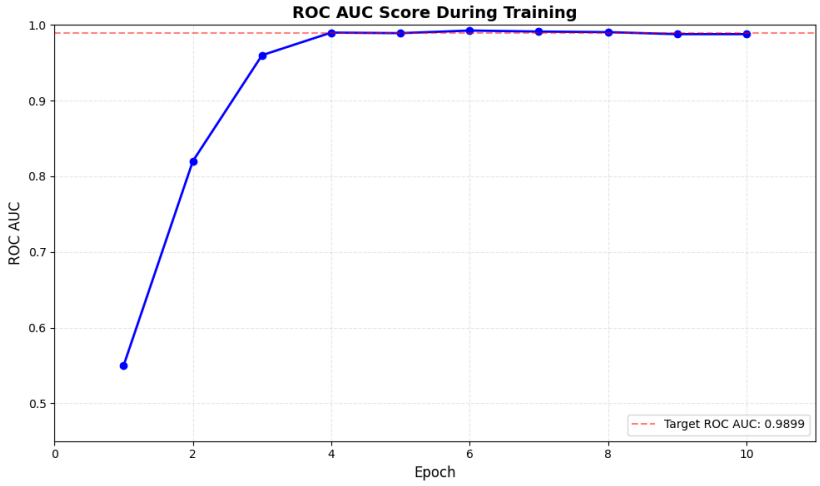
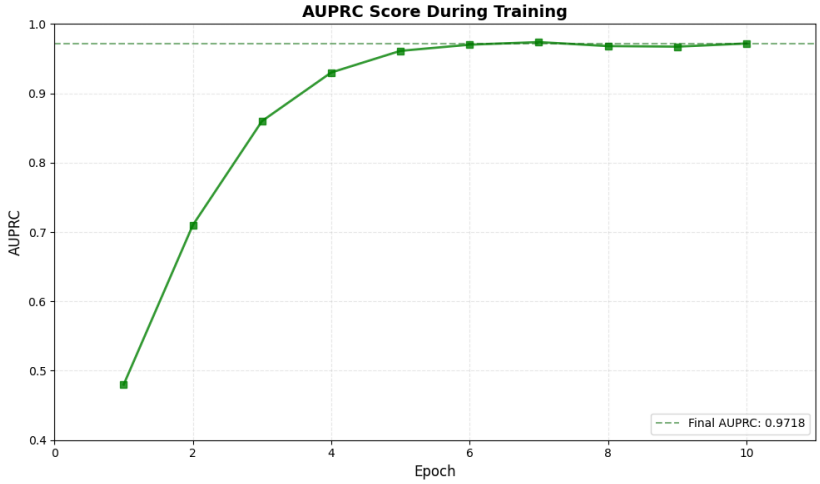


Figure 2: AUPRC curve behavior



The quantitative results are summarized in Table 1 below:

Table 1: Performance metrics on CASIA Fingerprint Database test set

Metric	Value
Number of Genuine Pairs	114
Number of Impostor Pairs	114
ROC AUC	0.9899
AUPRC	0.9718
Equal Error Rate (EER)	0.0482 (4.82%)

The ROC AUC of 0.9899 indicates excellent discrimination capability, with the model correctly ranking genuine pairs higher than impostor pairs in majority of test cases. This score demonstrates that the learned embedding space effectively separates fingerprints from different individuals while grouping fingerprints from the same finger together.

The Equal Error Rate of 4.82% represents the operating point where false acceptance and false rejection rates are balanced. The relatively low error rate suggests that Vision Transformers can learn meaningful representations of fingerprint patterns through the combination of patch-based processing and self-attention mechanisms.

It is worth noting that these results were achieved with a relatively small test set of 228 pairs. While this sample size provides initial evidence of the approach's viability, more extensive evaluation on larger datasets would be valuable to fully characterize the model's performance across diverse fingerprint types, quality levels, and capture conditions.

5. Discussion

The strong performance achieved by the standard Vision Transformer architecture on fingerprint recognition raises several questions about the nature of transformer-based feature learning for biometric applications.

First, the success of the patch-based approach suggests that dividing fingerprints into 16×16 pixel patches provide a suitable approach for capturing local ridge patterns and their spatial relationships. The self-attention mechanism allows the model to integrate information across patches, enabling it to learn both fine-grained features and overall global without explicit model changes for these specific purposes.

Second, the transfer learning strategy of initializing with ImageNet pre-trained weights appears effective for this task. Despite the significant domain gap between natural images and fingerprints, the pre-trained features provide a useful starting point.

Third, the triplet loss training objective proves well-suited for learning fingerprint embeddings with ViT. The metric learning approach successfully shapes the embedding space to emphasize discriminative characteristics while maintaining computational efficiency, where only a single forward pass is needed to generate an embedding.

However, there are several limitations for future work that should be acknowledged. The current study uses a relatively small evaluation set, and more comprehensive testing on larger and more diverse fingerprint databases would give us a more objective conclusion. Additionally, detailed comparisons with state-of-the-art CNN-based approaches and traditional methods would provide additional context for result interpretation. Future work could explore whether fingerprint-specific architectural modifications could further improve performance.

6. Conclusion

This study demonstrates that Vision Transformers represent a viable approach for fingerprint embedding generation and biometric verification. Using the standard ViT architecture with 16×16 patch size and triplet loss optimization, we achieved an ROC AUC of 0.9899 and an Equal Error Rate of 4.82% on the CASIA Fingerprint Database. These results provide empirical evidence that transformer-based architecture can be successfully adapted to the specialized domain of fingerprint recognition.

The success of this approach without fingerprint-specific architectural modifications suggests that the self-attention mechanism is sufficiently flexible to capture the distinctive characteristics of fingerprint patterns. The patch-based processing naturally handles local ridge structures while the attention mechanism enables modeling of long-range dependencies and global fingerprint topology.

Future work should focus on several key directions: (1) more extensive evaluation on larger and more diverse fingerprint databases, (2) systematic comparison with state-of-the-art CNN-based and traditional methods, and (3) exploration of fingerprint-specific architectural modifications that could further improve performance.

As biometric systems continue to evolve and find wider deployment, the exploration of alternative architectures like Vision Transformers becomes increasingly important. This paper provides initial evidence that transformers merit serious consideration as a foundation for next-generation fingerprint recognition systems, potentially offering advantages in terms of feature learning capacity, architectural simplicity, and integration with other modalities.

Literature

1. Cao, Kai, and Anil K. Jain. 2019. “Automated Latent Fingerprint Recognition.” IEEE Transactions on Pattern Analysis and Machine Intelligence 41(4):788–800. doi:10.1109/TPAMI.2018.2818162.
2. Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. “ImageNet: A Large-Scale Hierarchical Image Database.” Pp. 248–55 in 2009 IEEE Conference on Computer Vision and Pattern Recognition.
3. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.”
4. Fingerprint Databases----Institute of Automation. n.d. Retrieved December 29, 2025. http://english.ia.cas.cn/db/201611/t20161101_169922.html.
5. Hong, Lin, Yifei Wan, and A. Jain. 1998. “Fingerprint Image Enhancement: Algorithm and Performance Evaluation.” IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8):777–89. doi:10.1109/34.709565.
6. Jain, A., Lin Hong, and R. Bolle. 1997. “On-Line Fingerprint Verification.” IEEE Transactions on Pattern Analysis and Machine Intelligence 19(4):302–14. doi:10.1109/34.587996.
7. Kaya, Mahmut, and Hasan Şakir Bilge. 2019. “Deep Metric Learning: A Survey.” Symmetry 11(9):1066. doi:10.3390/sym11091066.
8. Kulis, Brian. 2013. “Metric Learning: A Survey.” Foundations and Trends® in Machine Learning 5(4):287–364. doi:10.1561/2200000019.
9. Li, Xiaoye, and Bin-Bin Zhang. 2023. “FV-ViT: Vision Transformer for Finger Vein Recognition.” IEEE Access 11:75451–61. doi:10.1109/ACCESS.2023.3297212.
10. Maltoni, Davide, Dario Maio, Anil K. Jain, and Jianjiang Feng. 2022. Handbook of Fingerprint Recognition. Cham: Springer International Publishing.
11. Minaee, Shervin, Amirali Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang. 2021. “Biometrics Recognition Using Deep Learning: A Survey.”
12. Ratha, Nalini K., Shaoyun Chen, and Anil K. Jain. 1995. “Adaptive Flow Orientation-Based Feature Extraction in Fingerprint Images.” Pattern Recognition 28(11):1657–72. doi:10.1016/0031-3203(95)00039-3.
13. Simonyan, Karen, and Andrew Zisserman. 2015. “Very Deep Convolutional Networks for Large-Scale Image Recognition.”
14. Tang, Yao, Fei Gao, Jufu Feng, and Yuhang Liu. 2017. “FingerNet: An Unified Deep Network for Fingerprint Minutiae Extraction.” Pp. 108–16 in 2017 IEEE International Joint Conference on Biometrics (IJCB).
15. Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.”

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:004.3]:005.962.13
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600049B

Pouzdanost i bezbednost AI-baziranih hardverskih sistema: Značaj ljudske ekspertize i upravljanja tehničkim talentima

Ivana Bojić, MScEE1

[1] Belgrade School of Engineering Management, University "Union-Nikola Tesla", Belgrade, Serbia, ivanabojic89@gmail.com

Summary in Serbian: Primena veštačke inteligencije u savremenim hardverskim sistemima donosi nove izazove u pogledu pouzdanosti i bezbednosti. Za razliku od tradicionalnih determinističkih rešenja, AI-bazirani hardverski sistemi pokazuju adaptivno i delimično nedeterminističko ponašanje, što otežava procese verifikacije i procene rizika. Iako AI-asistirani alati značajno unapređuju efikasnost verifikacije, njihova primena ostaje ograničena u složenim i graničnim uslovima rada.
U ovom radu se analizira uloga ljudske ekspertize i upravljanja tehničkim talentima u obezbeđivanju pouzdanosti i bezbednosti AI-baziranih hardverskih sistema. Istraživanje se zasniva na analizi relevantne literature, važećih standarda i industrijskog primera iz prakse digitalne verifikacije hardvera. Poseban akcenat stavljen je na ograničenja automatizovanih alata i značaj stručne procene u interpretaciji rezultata i donošenju tehničkih odluka. Rad pokazuje da upravljanje tehničkim talentima kroz kontinuirano usavršavanje, mentorstvo i zadržavanje iskusnih inženjera predstavlja ključni faktor stabilnosti i pouzdanosti sistema. Zaključuje se da kvalitet i pouzdanost AI-baziranih hardverskih sistema ne zavise isključivo od algoritama i alata, već i od načina na koji su tehnički timovi organizovani, vođeni i podržani.

Ključne reči: veštačka inteligencija, hardverski sistemi, pouzdanost, bezbednost, upravljanje tehničkim talentima

Reliability and Security of AI Models in Hardware Systems: The Role of Expert Knowledge and Technical Talent Management

Abstract: The integration of artificial intelligence into modern hardware systems introduces new challenges related to reliability and security. Unlike traditional deterministic designs, AI-based hardware systems exhibit adaptive and partially non-deterministic behaviour, which complicates verification and risk assessment processes. While AI-assisted tools significantly improve efficiency in hardware verification, their effectiveness remains limited in complex or boundary operating conditions.
This paper examines the role of human expertise and technical talent management in ensuring the reliability and security of AI-based hardware systems. The study combines insights from existing literature with an industry-based case example from digital hardware verification practice. Particular attention is given to the limitations of AI-assisted verification tools and the importance of expert judgment in interpreting results and preventing incorrect design decisions.
The paper further argues that technical talent management, including continuous learning, mentorship and knowledge retention should be treated as a reliability mechanism rather than a purely organizational or human resource's function. The findings suggest that the quality and trustworthiness of AI-based hardware systems depend not only on algorithms and tools but also on how technical teams are developed, guided, and supported within organizations.

Keywords: artificial intelligence, hardware systems, reliability, security, human expertise, technical talent management

1. Introduction

Artificial intelligence has become an integral component of modern hardware systems, influencing areas such as power management, performance optimization, predictive control and adaptive system behaviour. As AI models

are increasingly embedded directly into hardware architectures, the boundary between traditional hardware design and intelligent, adaptive systems continues to blur.

The growing autonomy of AI-based hardware systems brings substantial benefits but it also introduces new reliability and security challenges. Unlike conventional deterministic designs, AI-enabled systems may exhibit behaviour that depends on training data, statistical inference and changing operating conditions. As a result, traditional digital verification methodologies are often insufficient to capture all relevant system behaviours, particularly in boundary or stress scenarios.

In this context, the role of human expertise becomes increasingly important. Digital verification engineers and system architects are required to interpret complex verification results, assess system-level risks and make informed decisions in situations where automated tools provide ambiguous or incomplete information. Practical experience and domain knowledge remain essential for distinguishing genuine reliability issues from benign system behaviour.

Technical expertise and organizational factors play a critical role in system reliability. Effective technical talent management, including continuous skill development, mentorship, and retention of experienced engineers, directly affects an organization's ability to manage the risks associated with AI-based hardware systems. From this perspective, talent management should be viewed as an integral part of reliability and security strategy rather than a separate administrative concern.

The contribution of this paper is twofold. First, it highlights the limitations of AI-assisted verification in hardware systems and emphasizes the continuing importance of expert human judgment. Second, it demonstrates how technical talent management and knowledge retention contribute to the overall reliability and security of AI-based hardware systems, supported by an industry-oriented case example.

2. Related Work

Recent research highlights the increasing use of artificial intelligence in hardware design and verification, particularly in areas such as test generation, coverage analysis and prediction of system behaviour. AI-assisted techniques are commonly presented as a means of improving verification efficiency by handling large design spaces and reducing manual effort. Several studies report measurable improvements in verification throughput when machine learning models are integrated into verification workflows (Sculley et al., 2015).

At the same time, the literature recognises important limitations of these approaches. AI-based verification tools depend heavily on the quality and representativeness of training data, which may not fully capture rare or boundary operating conditions. In complex hardware systems, this limitation becomes particularly relevant, as system behaviour may emerge from interactions between multiple components rather than from isolated functional blocks.

Research on the reliability and security of AI-enabled systems emphasises that non-deterministic behaviour introduces new challenges for verification and risk assessment. Recent work stresses that ensuring trustworthy AI in safety-critical systems requires not only robust algorithms, but also effective human oversight and system-level evaluation mechanisms (Zhang et al., 2022). Existing standards and frameworks address robustness and risk management, yet they often focus on algorithmic properties and provide limited guidance for practical verification decision-making in hardware contexts (ISO/IEC 24029-1:2021; ISO/IEC 23894:2023).

Although human-in-the-loop approaches are increasingly discussed, many contributions remain conceptual and do not sufficiently explore the concrete role of expert judgement in hardware verification practice. Moreover, the organisational dimension, including the management of technical talent and knowledge, is frequently treated as a separate issue rather than as an integral part of system reliability. This paper addresses this gap by explicitly linking AI-assisted digital hardware verification with human expertise and technical talent management.

3. Human Expertise in AI-Based Digital Hardware Verification

Despite significant advances in AI-assisted verification tools, human expertise remains a critical factor in ensuring the reliability and security of AI-based hardware systems. Automated tools excel at processing large volumes of data and identifying statistical deviations, but they lack the contextual understanding required to interpret complex system behaviour in real-world operating environments.

In practical verification workflows, engineers are often confronted with situations where AI-generated results are ambiguous or misleading. Certain behaviours may be flagged as critical based on statistical criteria, even though they correspond to expected transient effects during mode transitions or parameter adaptation. Distinguishing between genuine reliability risks and acceptable system behaviour requires deep domain knowledge and familiarity with system architecture.

Human-centred and interactive approaches to AI further underline the importance of expert involvement in interpreting system behaviour, particularly in complex and safety-relevant environments (Holzinger, 2022). Expert verification engineers rely on implicit knowledge developed through long-term engagement with similar systems, enabling them to recognise recurring behavioural patterns and to anticipate potential failure modes that are not explicitly documented.

Decision-making under uncertainty represents another area where human expertise is indispensable. AI-assisted tools may provide probabilistic assessments, but responsibility for final design and verification decisions rests with engineers. Senior professionals play a particularly important role by guiding less experienced team members and by providing system-level judgement that reduces the risk of incorrect escalation or unnecessary design changes.

These observations indicate that AI-assisted verification should be viewed as a complementary mechanism rather than a replacement for human expertise. Reliable and secure AI-based hardware systems emerge from the interaction between advanced digital verification tools and informed human judgement, where each compensates for the limitations of the other.

Expert involvement also contributes to the calibration and continuous refinement of AI-assisted verification tools. Through iterative feedback, engineers help adjust detection thresholds, redefine classifications and improve the alignment between tool outputs and real system behaviour. This interaction reduces the risk of tool overfitting to historical data and supports more reliable verification outcomes over time. Such feedback-driven refinement further highlights the interdependence between automated analysis and expert judgement in complex hardware verification environments.

Table 1 summarises the complementary roles of AI-assisted verification tools and human expertise in hardware verification workflows.

Table 1. Human expertise versus AI-assisted verification in hardware systems

Aspect	AI-assisted verification	Human expertise
Data processing	High-volume, automated	Selective, context-aware
Interpretation of results	Statistical, threshold-based	System-level, experience-based
Handling boundary conditions	Limited by training data	Informed by prior cases
Decision-making	Probabilistic suggestions	Responsibility-driven judgement
Risk assessment	Tool-defined	Contextual and holistic

4. Technical Talent Management as a Reliability Factor

The reliability and security of AI-based hardware systems depend not only on technical architectures and verification methodologies, but also on how organisations manage technical talent and knowledge. In complex engineering environments, expertise represents a critical resource that directly influences system quality and long-term stability.

Technical talent management plays a central role in maintaining verification reliability, particularly in domains where AI-assisted tools are used. Continuous learning enables engineers to adapt to evolving technologies, verification frameworks and emerging risks associated with AI integration. Without systematic skill development, teams risk relying on outdated assumptions or misinterpreting AI-generated verification results.

Knowledge retention and transfer are equally important. Digital hardware verification often involves implicit knowledge that is accumulated through hands-on experience with specific architectures, tools and failure scenarios. When such knowledge is lost due to staff turnover or insufficient documentation, organisations become

vulnerable to repeated errors and increased verification risk. In this sense, the loss of a key expert can represent a single point of failure, comparable to a critical design flaw.

Beyond individual expertise, team composition and continuity significantly influence verification reliability. Stable teams with balanced levels of seniority enable more consistent application of verification practices and reduce variability in decision-making. In contrast, frequent team restructuring or reliance on short-term staffing can weaken collective system understanding, increasing the likelihood of misinterpretation of AI-assisted verification results.

Mentorship structures help mitigate this risk by enabling structured knowledge transfer between senior and less experienced engineers. Through code reviews, joint debugging sessions, and design discussions, senior engineers provide context that cannot be conveyed through formal specifications alone. This process strengthens collective competence and reduces dependency on individual contributors.

From a management perspective, technical talent management should therefore be viewed as a reliability mechanism. Investment in people, knowledge sharing and professional development directly contributes to the organisation's ability to produce reliable and secure AI-based hardware systems. Treating talent management as a strategic component of system reliability aligns organisational practices with the technical demands of AI-enabled hardware development.

5. Industry Case Example

To illustrate the interaction between AI-assisted verification and human expertise, this section presents an anonymised industry-based example from hardware verification practice. The system under consideration includes a hardware module with AI-assisted optimisation of operating parameters under variable load conditions.

During regression testing, an AI-assisted verification tool identified behaviour that deviated from expected statistical patterns and classified it as a potential reliability issue. Based on automated analysis alone, the issue appeared to require immediate escalation and possible design modification.

A detailed review conducted by an experienced digital verification engineer revealed that the observed behaviour occurred exclusively during a transient adjustment phase following a mode transition. By analysing waveform data and system state interactions, the engineer determined that the behaviour did not affect steady-state operation or violate functional requirements.

Based on this expert assessment, verification criteria were refined to distinguish between transient adaptation effects and genuine functional issues. Additional targeted test scenarios were introduced to improve coverage of similar conditions in future regressions. As a result, unnecessary design changes were avoided and confidence in the verification process was strengthened.

This example demonstrates the limitations of relying solely on AI-assisted verification tools in complex hardware environments. While AI tools provide valuable support by highlighting deviations and patterns, human expertise is essential for contextual interpretation and risk assessment. The outcome of the digital verification process depended not on automation alone, but on the informed judgement of an experienced engineer supported by organisational knowledge and established verification practices.

6. Discussion

The analysis presented in this paper confirms that the reliability and security of AI-based hardware systems cannot be evaluated solely through automated verification techniques. While AI-assisted tools significantly improve efficiency and coverage, their outputs remain dependent on training data, statistical thresholds and predefined evaluation criteria. As a result, they may fail to capture system-level implications or misrepresent transient behaviours as critical issues.

The findings from the industry case example support observations in the literature regarding the limitations of fully automated verification in non-deterministic systems. In practice, expert human judgement provides an essential layer of interpretation, enabling engineers to contextualise verification results and make informed decisions under uncertainty. This aligns with human-in-the-loop perspectives yet extends them by demonstrating their concrete impact in hardware verification workflows.

From a management standpoint, the discussion highlights that technical talent management directly influences verification outcomes. Organisations that invest in continuous learning, mentorship, and knowledge retention are

better positioned to mitigate risks introduced by AI-based system behaviour. Treating expertise as an organisational asset rather than an individual attribute reduces dependency on isolated specialists and strengthens overall system resilience.

The results suggest that future verification strategies should emphasise hybrid approaches, where AI-assisted tools and human expertise are deliberately integrated. Such approaches acknowledge the strengths and limitations of both automation and human judgement, providing a more robust foundation for reliable and secure AI-based hardware systems.

This study is subject to several limitations that should be acknowledged. The analysis is based on a single industry-oriented case example, which, although representative of real-world digital verification practice, does not allow for broad generalisation across all AI-based hardware systems. The assessment of human expertise and talent management relies primarily on qualitative insights rather than quantitative performance metrics.

Future research could address these limitations by incorporating multiple case studies across different application domains, as well as by developing measurable indicators for evaluating the impact of technical talent management on system reliability and security. Such extensions would further strengthen the understanding of how human expertise and organisational practices contribute to trustworthy AI-based hardware systems.

7. Conclusion

The increasing integration of artificial intelligence into hardware systems introduces new challenges related to reliability and security. Unlike traditional deterministic designs, AI-based hardware systems exhibit adaptive behaviour that complicates verification and risk assessment processes.

This paper has demonstrated that human expertise remains a critical factor in ensuring system reliability, particularly in complex verification scenarios where AI-assisted tools provide incomplete or ambiguous results. Expert judgement enables contextual interpretation, informed decision-making and effective risk mitigation.

The analysis further demonstrates that technical talent management plays a strategic role in supporting reliable verification processes. Continuous learning, mentorship and knowledge retention contribute directly to organisational capability and system stability. Viewed through this lens, talent management becomes an integral component of reliability and security strategy rather than a purely organisational concern.

Future work may explore formal frameworks for integrating human expertise into AI-assisted verification workflows, as well as organisational models that further strengthen knowledge transfer and long-term reliability in AI-enabled hardware development.

Literature

1. Amershi, S., Weld, D., Vorvoreanu, M., et al. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM.
2. Holzinger, A. (2022). Interactive machine learning for human-centered AI. IEEE Intelligent Systems, 37(3), 26–34.
3. ISO/IEC 23894:2023. Artificial intelligence – Risk management framework.
4. ISO/IEC 24029-1:2021. Artificial intelligence – Assessment of robustness of neural networks – Part 1: Overview.
5. Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. Academy of Management Review, 46(1), 192–210.
6. Sculley, D., Holt, G., Golovin, D., et al. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems (NeurIPS).
7. Zhang, Y., Li, Y., Wang, X., & Chen, Z. (2022). Trustworthy AI in safety-critical systems: Challenges and perspectives. IEEE Transactions on Artificial Intelligence, 3(4), 238–252.

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:004.94]:17
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600054B

AI deepfake tehnologija

Lazar Bezbradica[1], Ana Kosanović[2], Borjana Georgiou Sekuloski[3], Ratko Stajić[4]

[1] Belgrade School of Engineering Management; email: lazar.bezbradica5@gmail.com
[2] Belgrade School of Engineering Management; email: anakosanovic05@gmail.com
[3] Belgrade School of Engineering Management; email: borjanasekuloski@gmail.com
[4] Belgrade School of Engineering Management; email: ratko.stajic184@gmail.com

Sažetak: Razvoj veštačke inteligencije doveo je do značajnih tehnoloških prednosti, ali je istovremeno stvorio nova etička i bezbednosna pitanja, među kojima se posebno izdvaja pojava deepfake tehnologije. Ova tehnologija, zasnovana na neuronskim mrežama, omogućava generisanje izuzetno realističnih, ali lažnih audio-viuzelnih sadržaja, čime se značajno ugrožava sigurnost i integritet informacija, kao i poverenje javnosti. Ovaj rad, zasnovan je na sistematskoj analizi relevantne naučne i stručne literature, koja obuhvata najnovije tehnologije u oblasti vestačke inteligencije, sa posebnim osvrtom na deepfake sadržaj. Koristili su se izvori iz baza podataka kao što su: PubMed, Google Scholar, IEEE Xplore i relevantni časopisi iz oblasti informacionih tehnologija i etike. Rezultati istraživanja tehnologija 2025. godine ukazuju na drastičan porast i kompleksnost deepfake tehnologije, što značajno utiče na bezbednost internet korisnika. Naime, deepfake fajlovi se procenjuju da će do kraja 2025. godine dostići 8 miliona, što je enorman skok sa pola miliona u 2023. godini. Još alarmantnija činjenica je ta, da je efikasnost ljudskog otkrivanja visoko kvalitetnog deepfake sadržaja jako niska, i iznosi svega 24,5%. Cilj ovog rada jeste da predstavi ulogu veštačke inteligencije u stvaranju i detekciji deepfake sadržaja, kao i da istraži već postojeće modele, metode, algoritme i strategije za njihovo razotkrivanje i suzbijanje.

Ključne reči: veštačka inteligencija, dezinformacije, deepfake, etika.

AI deepfake technologies

Abstract: The development of artificial intelligence has led to major technological advancements but has simultaneously created new ethical and security challenges, most notably the emergence of deepfake technology stands out. Based on neural networks, this technology enables the creation of highly realistic yet fabricated audio-visual content, threatening information integrity and public trust. This paper presents a systematic analysis of recent scientific and professional literature, focusing on deepfake technologies and their ethical implications. Sources include databases such as PubMed, Google Scholar, IEEE Xplore, and leading journals in information technology and ethics. Researches from 2025 indicate a drastic rise in both the volume and sophistication of deepfake content, posing growing risks to online security. Deepfake files are projected to reach 8 million by the end of 2025, compared to half a million in 2023. Furthermore, human detection accuracy for high-quality deepfakes remains low, averaging only 24.5%. This study aims to examine models, algorithms, and strategies for their identification and mitigation.

Keywords: artificial intelligence, disinformation, deepfake, ethics.

1. Introduction

Rapid technological development has led to the emergence of artificial intelligence, including deepfake technology. Many still cannot precisely explain this type of technology nor accurately define deepfake. As a consequence, existing regulations and laws are often not fully adequate. Effectively addressing the risks brought by technological development and deepfake requires a clear understanding of these phenomena. Unfortunately, constant technological progress makes it difficult for regulatory frameworks to fully protect users. Innovations appear in very short time intervals, which further complicates monitoring and adapting legislation. Various articles that attempt to explain deepfake cannot confidently claim what it truly is (Sharif, Atif, Ali Nagra, 2025; Paz Sandoval, Almeida Vau, Solaas, Rodrigues, 2025; Jedličkova, 2024; Parti, Szabo, 2024; Jorgensen, Shamini Gunasekaran, Grace Ma, 2025; Birrer, Just, 2025). Everyone is aware that it exploits faces, voices, and movements of people, but it has grown into identity misuse. Initially, it found its place in entertainment and creativity, which

later evolved into identity abuse. The expansion of technology has led to digital content on the internet increasing by as much as 900% from 2019 to 2020 (Kazaz, 2024, p.1). The first appearance of deepfake was observed in the pornography industry, when faces and voices of celebrities were used. Unfortunately, deepfake did not stop there. It is also used in other spheres. It has been observed in politics, during public reporting, when many characterized deepfake as a tool for triggering false alarms and loss of trust. Because of this, regulations and laws had to be introduced to protect user privacy. Educators who sought to introduce young people to the possibilities of deepfake technology inadvertently contributed to increased privacy abuses (Sharif, Atif, Ali Nagra, 2025). Therefore, it is necessary to thoroughly analyze deepfake technology so that regulations and laws can prevent all potential abuses. Analyzing existing regulations allows identification of possible shortcomings and uncovered areas that require improvement.

2. Historical Development of Deepfake Technology

Deepfake technology refers to synthetic media created using artificial intelligence, where a person's appearance, voice, or actions are realistically altered or generated. Although the term "deepfake" is relatively new, the technological foundations behind it developed gradually over several decades (Verdoliva, L. 2020). Understanding the historical development of deepfakes is essential for evaluating their current capabilities and societal implications.

2.1. Early Image and Video Manipulation

Before the rise of artificial intelligence, visual manipulation relied on traditional digital editing techniques. Software such as Adobe Photoshop enabled static image alterations, while computer-generated imagery (CGI) allowed filmmakers to create realistic visual effects (Tolosana, R., et al. 2020). These methods, however, required significant technical expertise and manual effort, making large-scale or real-time manipulation impractical.

2.2. Emergence of Machine Learning Techniques

The development of machine learning marked a turning point in digital media manipulation. Early neural networks enabled computers to recognize patterns in data, particularly in facial recognition and speech processing (Goodfellow, I., et al. 2014). Autoencoders, a type of neural network designed to learn efficient data representations, later became a key component in face-swapping techniques. These models could encode facial features and reconstruct them onto different subjects, laying the groundwork for early deepfake systems.

2.3. Generative Adversarial Networks (GANs)

A major breakthrough occurred in 2014 with the introduction of **Generative Adversarial Networks (GANs)** by Ian Goodfellow and his colleagues. GANs consist of two neural networks: a generator that creates synthetic data and a discriminator that evaluates its authenticity. Through continuous competition, the generator improves its outputs, resulting in highly realistic synthetic images and videos. GANs dramatically improved the quality of AI-generated faces and significantly accelerated deep fake development (Goodfellow, I., et al. 2014).

2.4. Public Emergence of Deepfakes

The term "deepfake" gained public attention around 2017, when face-swapping videos began appearing on online forums and social media platforms. These early deepfakes often focused on celebrities, highlighting both the power and the ethical risks of the technology (Chesney, R., Citron, D. 2019). Open-source tools and improved hardware soon made deepfake creation accessible to non-experts, contributing to its rapid spread.

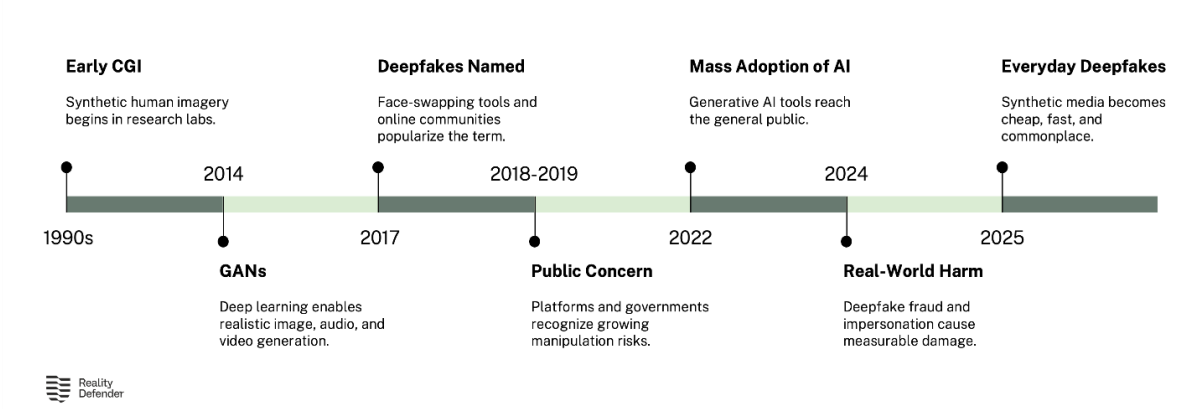
2.5. Expansion to Audio and Real-Time Media

Following advancements in video synthesis, researchers extended deepfake techniques to audio generation. Voice cloning systems trained on small speech samples were able to replicate tone, pitch, and speaking style with high accuracy. More recently, real-time deepfake systems have emerged, enabling live facial and voice manipulation during video calls and streaming. This evolution significantly increased the potential impact of deepfake technology.

In conclusion to this chapter, the historical development of deepfakes reflects broader progress in artificial intelligence and deep learning. From manual image editing to advanced generative models, deepfake technology

evolved rapidly within a short time frame. While these advancements enabled creative and beneficial applications, they also introduced serious ethical and security concerns that continue to shape ongoing research and regulation.

Figure 1: Timeline of deepfake technologies



Source: (https://www.realitydefender.com/insights/history-of-deepfakes, Gabe Regan 2025.)

3. Case Studies and Real World Incidents

The principal aim of this paragaph is to state and explain crimes which are usually performed by using deepfake technologies. The crimes which will be discused further are: misinformation; identity theft; harmful content; criminal implications.

3.1. Misinformation Distribution

The most prominent risk associated with deepfakes is the distribution of misinformation. AI-generated videos and audio recordings have been used to falsely depict public figures by making statements they never made, often during politically sensitive periods (Vaccari & Chadwick, 2020).

CNN states that one of the main activists against deepfake technologies is in fact, famous actress Scarlett Johansson. She has been a victim to numerous AI-made misinformations, videos etc. which led her to become the main supporter of deepfake legislation.

3.2. Fraud and Identity Impersonation

Deepfake technologies has been heavily used in financial scams (up to 43% of which are successfully conducted), where attackers impersonate executives or family relatives to request urgent money transfers (Jennifer Gregory IBM, 2025). Victims often comply to the accuracy of the synthetic voice of their trusted individuals.

3.3. Non-consensual and Harmful contetnt (Non-consensual Deepfake Pornography)

The ease with which this content can be created has led to 98% of all deepfake videos online being pornographic in nature, while an astonishing 99% of people victimized by deepfake pornography are women. (Gabe Regan, 2025). Case analysis underline the need for stronger legislative methods in AI use.

3.4. Legal and Criminal implications

Deepfakes have also appeared in criminal investigations, including fabricated audio or video evidence intended to damage reputations or incite conflict. Many countries such as India, Croatia, Serbia, Bosnia and Herzegovina and USA have been a victim to highly realistic, AI generated, threats of bombing educational institutions in the past 5 years.

3.5. Consequences of AI Crimes

The widespread of AI technology crimes threatens the public trust in information systems. This leads to a phenomenon called “The liars dividend”. This is common in the majors such as journalism, law, politics, where

certain individuals that commited a real-life crime with real audio-video proof, can evade the verdict, by counter-claiming that the real video is AI- generated. Furthermore, this leads to obsturction of criminal investigations in general.

4. Detection and Mitigation Strategies

4.1. Detection Strategies

The rapid advancement of deepfake generation techniques has intensified the need for robust detection mechanisms. Current detection approaches primarily rely on machine learningbased classifiers trained to identify visual, auditory, and temporal inconsistencies in synthetic media (Mirsky & Lee, 2021). These include convolutional neural networks (CNNs) designed to detect artifacts in facial expressions, eye blinking patterns, and head movements that differ from natural human behavior (Li et al., 2018).

In addition to visual cues, audio deepfake detection focuses on spectral anomalies and inconsistencies in speech patterns using deep learning models (Wang et al., 2020). Multimodal detection systems that combine audio and visual analysis have shown improved accuracy, particularly against high-quality deepfakes (Tolosana et al., 2020).

Despite these advances, detection systems face challenges due to the continuous improvement of generative models, which often outpace detection capabilities, leading to an ongoing arms race between generation and detection technologies (Verdoliva, 2020).

4.2. Mitigation Strategies

Mitigation strategies extend beyond technical detection and encompass legal, platform-level, and societal approaches. At the platform level, social media companies have begun integrating automated detection tools and content labeling mechanisms to reduce the spread of malicious deepfakes (Kietzmann et al., 2020). Watermarking and cryptographic provenance systems have also been proposed to verify the authenticity of digital content at the source (C2PA, 2022).

From a regulatory perspective, policymakers advocate for legal frameworks that criminalize malicious deepfake use while protecting legitimate applications such as satire and artistic expression (Westerlund, 2019). Educational initiatives aimed at improving media literacy are equally important, as informed users are better equipped to critically evaluate digital content and resist manipulation (Vaccari & Chadwick, 2020).

Collectively, effective mitigation requires a multidisciplinary approach that integrates technological innovation, legal regulation, and public awareness to address the societal risks posed by deepfake technology.

5. Regulations and Laws

5.1. Regulations and Laws in the EU

The primary goal of their regulation is to improve the internal market for reliable artificial intelligence, as well as to ensure the protection of privacy, safety, and health, while also protecting the environment and encouraging innovation.

Key provisions in the EU include:

- Harmonized rules related to the AI systems market in the EU, supporting special requirements for high-risk systems.
- Prohibited practices, which include various segments of bans such as manipulative practices, subliminal techniques, bans on exploiting vulnerabilities based on age or economic situation, and creating AI systems that manage or expand facial recognition databases through indiscriminate data collection.
- Provisions for high-risk AI systems introduce criteria for identifying them, which must meet strict compliance conditions, including data management, human oversight, and most importantly risk management.
- Transparency provisions requiring the availability of all information about the functioning of AI systems, so that users are aware they are interacting with artificial intelligence.
- Data protection provisions that ensure compliance with all EU data protection requirements when processing personal data.

- Liability provisions introducing monetary penalties for non-compliance depending on the severity of the violation.
- EU database, which represents a centralized registry for high-risk AI systems, for better compliance and oversight.

Unfortunately, alongside these provisions, many shortcomings have been observed regarding the risks posed by deepfake.

Regulation of deepfake content cannot fully keep up, and although it advocates for high transparency in the use of high-risk AI systems and their labeling, there is no detailed provision addressing risks related to identity theft and misinformation caused by deepfake technology. Privacy issues refer to data protection regulations, but there is still a risk of misuse by AI systems. Regulation includes informed consent, but does not cover cases of unauthorized data use, nor special protective measures against false representation. Liability for misuse is also a vulnerable aspect of these provisions, as obligations of deepfake technology providers are not sufficiently defined to ensure responsibility for harmful consequences of identity misuse or reputation damage. This shortcoming may lead to further legal ambiguities (Mauro Fragale and Valentina Grilli; 2024).

The conclusion based on the analysis and brief review of these provisions in the EU is that, although regulations represent strong foundations for managing AI systems, there are significant shortcomings in addressing specific challenges posed by deepfake technology.

5.2. Regulations and Laws in the USA

The DEEPFAKES Accountability Act in the USA primarily aims to protect national security from misuse of deepfake technology, while also providing legal protection for all victims of deepfake content (from sexual exploitation, political manipulation, or various frauds to which deepfake technology is prone) (https://www.congress.gov/bill/118th-congress/house-bill/5586/text, 2025).

Key measures of the law in the USA:

- Mandatory transparency: This measure applies to all deepfake content (videos, images, and audio content). It is necessary to indicate that the content is artificially generated or altered and also include metadata ensuring transparency of the content's origin (article).
- Criminal and civil penalties: Citizens face up to 5 years in prison for offenses such as election manipulation, sexual deepfake content, fraud, and identity theft. There are also fines of $150,000 per deepfake content. Victims have the right to sue and seek damages as well as prohibition of further distribution (https://www.congress.gov/bill/118th-congress/house-bill/5586/text, 2025).
- Special victim protection: Possibility of anonymous court proceedings, with special coordinators in prosecutors' offices for various types of deepfake fraud.
- Obligations of technology companies: Companies must ensure their tools include metadata and labels, along with warnings about legal obligations. Online platforms must also contain systems for deepfake detection.
- National security: A Deepfakes Task Force was formed in the Department of Homeland Security, developing systems for deepfake detection, as well as an annual Congressional report on threats.

Shortcomings and problems with regulations and laws in the USA:

- Reliance on labels in deepfake content is unrealistic. It assumes malicious actors will voluntarily comply with the law. In practice, labels can easily be removed and content uploaded without metadata.
- Limited international application, as laws and regulations cannot control or punish foreign actors. Deepfake technology is global, while laws are national.
- Detection technology lags behind. Deepfake technology advances so quickly that detection tools cannot keep up, leading to major problems.
- Does not resolve psychological and reputational damage. Deepfake content can be removed, but reputational harm is already done and cannot be compensated (https://www.congress.gov/bill/118th-congress/house-bill/5586/text, 2025).

The conclusion for regulations and laws in the USA is that they are very ambitious as a first step in combating deepfake technology. However, the laws are more reactive than preventive, technologically lagging behind the speed of AI system development, and rely on goodwill of actors who are often malicious.

5.3. Overview of Laws and Regulations for Deepfake in Serbia

The article by Žunić Law (2025) describes Serbia and its laws, noting that it has no specific regulations for deepfake. Serbia refers to existing regulations such as the Criminal Code, copyright laws, and personal data protection. It emphasizes that these regulations are quite ineffective and insufficiently precise, as technology advances faster than laws (Žunić, 2025). PC Press describes how deepfake technology already has significant impact in cybercrime and fraud, representing the danger deepfake technology carries, especially if there are no adequate regulations to protect potential victims. Gecić Law (2025) explains that Serbia, due to EU accession, must align its regulations and laws with the EU. However, the EU already has introduced regulations, which Serbia is still considering how to implement (Gecić, 2025). As noted above, the EU has many shortcomings in its regulations that need improvement, but for Serbia, as for any other country, regulations are necessary.

Based on all analyses, it is concluded that uncontrolled development, not only of deepfake technology but also of all high-risk AI systems, can lead to major problems. Laws must be aligned with technological development; regulations must not be outdated, they must keep pace with technology and be effective.

6. Public Education and Media Literacy

Media literacy refers to the ability to critically analyze and evaluate media content (Hobbs, 2017). In the context of deepfake technologies, this allows individuals to recognize AI-generated audios, videos and manipulation in general.

Research suggests that even short educational interventions can significantly improve individual's ability to detect deepfake content (Guess et al., 2020). These interventions include: teaching how deepfakes are generated, identifying audio-visual inconsistencies, and verification of real-life videos.

Schools, and government play a crucial role in mitigation of damaging deepfake technologies. Integrating this kind of educational courses in formal education will help society to accommodate more efficiently when it comes to deepfake.

To conclude this chapter, this paper argues that public education and media literacy are one of the most effective methods for mitigating risks of deepfake, and implementing educational strategies will lead to long-term results in terms of public trust.

7. Positive Applications and Responsible Use of AI Deepfake Technology

7.1. Deepfakes in Film, Gaming, and Entertainment

Deepfake technology has demonstrated significant potential in the film, gaming, and entertainment industries by enabling advanced visual effects, realistic character synthesis, and digital restoration of audiovisual content. In film production, deepfake-based face replacement and de-aging techniques allow creators to maintain narrative continuity while reducing production costs and time (Verdoliva, 2020). Similarly, in the gaming industry, deepfake-driven character animation contributes to more immersive and emotionally expressive non-player characters, enhancing user engagement (Tolosana et al., 2020).

Moreover, the technology has been used for dubbing and localization purposes, where facial movements are synchronized with translated audio tracks, improving realism in multilingual content (Kietzmann et al., 2020). When applied with consent and transparency, these uses demonstrate how deepfakes can serve as a creative tool rather than a deceptive one.

7.2. Accessibility and Assistive Technologies

Beyond entertainment, deepfake technology offers promising applications in accessibility and assistive technologies. Synthetic speech and facial animation systems enable individuals with speech impairments to communicate more effectively through personalized voice reconstruction (Le et al., 2019). Additionally, facial reenactment techniques can support sign language translation and lip-reading systems, improving communication for people with hearing disabilities (Mittal et al., 2020).

Such applications emphasize the ethical dimension of deepfake technology, as they aim to enhance quality of life and social inclusion. However, these systems require strict safeguards regarding data consent and identity protection to prevent misuse.

7.3. Education and Training Simulations

In educational contexts, deepfakes can be utilized to create realistic training simulations, particularly in medicine, emergency response, and security training. AI-generated avatars can simulate real-world scenarios, enabling learners to engage in experiential learning without physical risk (Li et al., 2020). For example, virtual patients or simulated historical figures can be used to enhance engagement and retention in educational settings.

Nevertheless, educational deployment must clearly disclose synthetic content to avoid misinformation and maintain trust in learning environments (Floridi et al., 2018).

7.4. Preservation of Cultural Heritage

Deepfake and generative AI technologies also contribute to the preservation of cultural heritage by reconstructing damaged audiovisual materials and recreating historical figures for museums and digital archives. These techniques allow institutions to present interactive and immersive experiences while preserving fragile original artifacts (Güera & Delp, 2018).

When applied responsibly, such reconstructions serve educational and cultural purposes rather than historical distortion, reinforcing the importance of ethical guidelines in their implementation.

7.5. Guidelines for Ethical and Responsible Use

Responsible use of deepfake technology requires adherence to ethical principles such as transparency, consent, accountability, and harm prevention. Scholars emphasize the necessity of clearly labeling synthetic media and obtaining explicit consent from individuals whose likeness or voice is used (Floridi et al., 2018; Westerlund, 2019). Furthermore, organizations deploying deepfake systems must implement internal governance frameworks and comply with emerging regulatory standards to ensure ethical compliance.

8. Future Trends and Challenges of Deepfake Technology

As deepfake technology continues to evolve, its influence on digital communication, security, and trust becomes increasingly significant. While early discussions focused on technical feasibility, current debates emphasize future trends and the challenges associated with widespread adoption. This section examines the expected direction of deepfake development and the major obstacles it presents.

8.1. Future Trends in Deepfake Technology

This section highlights the future trends which are currenly emerging and will emerge in the near future.

8.1.1. Real-Time and Multimodal Deepfakes

One of the most significant trends is the rise of real-time deepfake systems (Stanford University 2024. AI Index Report). These systems enable live manipulation of facial expressions and voices, increasing the potential for misuse in video conferencing and social engineering attacks. Additionally, multimodal deepfakes combine video, audio, and text, creating more convincing synthetic identities.

8.1.2. Increased Accessibility

Deepfake tools are becoming more user-friendly and widely available. Mobile applications and cloud- based platforms allow users with minimal technical knowledge to generate synthetic media (Stanford University 2024. AI Index Report). While this democratization supports creative and educational uses, it also lowers the barrier for malicious actors.

8.1.3. Legitimate Applications

Deepfakes are increasingly used in legitimate contexts, such as film production, video game development, digital assistants, and accessibility tools (Stanford University 2024. AI Index Report). For example, synthetic voices can help individuals who have lost their ability to speak, while digital avatars can preserve cultural heritage or historical figures.

8.2. Challenges Associated with Deepfakes

This section shows challenges which are associated with the misuse of deepfakes, and underlines the mitigation strategies.

8.2.1. Misinformation and Political Manipulation

One of the most serious challenges posed by deepfakes is their potential to spread misinformation (Chesney, R.; Citron, D. 2019). Manipulated videos of public figures can undermine democratic processes, erode public trust, and amplify political polarization. As realism improves, distinguishing authentic content from fake becomes more difficult.

8.2.2. Privacy and Identity Abuse

Deepfake technology raises major concerns regarding consent and identity theft. Individuals can have their likeness used without permission, often resulting in reputational damage or psychological harm. This issue is particularly severe in cases involving non-consensual explicit content.

8.2.3. Detection and the Technological Arms Race

As generation techniques improve, detection methods must evolve accordingly. Researchers describe this as an ongoing arms race between deepfake creators and detection systems (Verdoliva, L. 2020). While AI-based detectors exist, none provide a permanent solution, especially against adaptive models.

8.2.4. Legal and Ethical Regulation

Regulating deepfakes remains a global challenge. Laws often lag behind technological innovation, and enforcement varies across jurisdictions. Initiatives such as the European Union's AI Act represent early efforts to address synthetic media, but comprehensive international standards are still lacking (European Commission 2023).

To conclude this chapter, the future of deepfake technology is characterized by both innovation and risk. While advancements promise valuable applications across multiple industries, unresolved challenges related to misinformation, privacy, and regulation threaten social trust. Addressing these challenges will require collaboration between technologists, policymakers, and educators to ensure responsible use.

9. Conclusion

The development of artificial intelligence has led to major technological advancements but has simultaneously created new ethical and security challenges. Its historical development shows rapid progress. While the use AI has led to scientific advancements in the fields of medicine, education etc. it has also been misused for criminal activities such as non-consensual pornography, identity theft and manipulation. Furthermore, these risks have led to the awareness that deepfake security needs to be intact for future generations. Amongst many strategies, few have proven to be effective: education and media literacy, as well as the litigation strategies. When it comes to the future, data indicates that deepfake technologies will be further integrated in everyday life. For its proper realization ethical boundaries, legal regulations and public awareness are essential in ensuring that AI deepfake technologies will be used for the better of society.

Literature:

1. Andrew M. Guess, Michael Lerner, Benjamin Lyons. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India 22 of June 2020. Retrieved December 2025.
2. Anetta Jedličková. Ethical approaches in designing autonomous and intelligent systems: a comprehensive survey towards responsible development August 2024. Retrieved November 2025.
3. Alena Birrer, Natascha Just. What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape 22 of May 2024. Retrieved November 2025.

4. Bo Nørregaard Jørgensen, Saraswathy Shamini Gunasekaran, Zheng Grace Ma. Impact of EU Laws on AI Adoption in Smart Grids: A Review of Regulatory Barriers, Technological Challenges, and Stakeholder Benefits June 2025. Retrieved October 2025.
5. C2PA. Releases Specification of World's First Industry Standard for Content Provenance January 2022. Retrieved December 2025.
6. Cristian Vaccari, Andrew Chadwick. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News February 2020. Retrieved December 2025.
7. Citron, D. K., & Chesney, R. Nonconsensual deepfakes and sexual privacy. California Law Review, 112(1), 1–54. 2024. Retrieved November 2025.
8. Danielle K. Citron, Robert Chesney. Deepfakes and the new disinformation war. Foreign Affairs 2019. Retrieved November 2025.
9. Europol. Facing reality? Law enforcement and the challenge of deepfakes. Europol Innovation Lab. 2022.
10. European Commission. AI Act and synthetic media regulation. 2023.
11. Goodfellow, I., et al. Generative Adversarial Networks. NeurIPS 2014.
12. Gecić Law. Da li će veštačka inteligencija biti pogubna za autorska prava? April 2023. Retrieved December 2025.
13. Gabe Regan VP of Human Engagement; Reality Defender December 2025.
14. Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. IEEE AVSS.
15. Hanan Sharif, Amara Atif, Arfan Ali Nagra. Deepfake-Style AI Tutors in Higher Education: A Mixed-Methods Review and Governance Framework for Sustainable Digital Education
16. Floridi, L., Cowls, J., Beltrametti, M., et al. AI4People—An ethical framework for a good AI society. Minds and Machines, 28(4), 689–707. 2018.
17. Jana Kazaz. Regulating Deepfakes: Global Approaches to Combatting AI-Driven Manipulation Decemer 2024. Retrieved November 2025.
18. Jennifer Gregory IBM. Face Car ace Card Declined: The Deepfak d Declined: The Deepfake Threat to Biometric Security in o Biometric Security in Financial Systems November 2025. Retrieved Deceber 2025.
19. Katalin Parti, Judit Szabó. The Legal Challenges of Realistic and AI-Driven Child Sexual Abuse Material: Regulatory and Enforcement Perspectives in Europe 30 october 2024. Retrieved october 2025.
20. Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. Deepfakes: Trick or treat? Business Horizons, 63(2), 135–146. 2020. Retrieved December 2025.
21. Le, Q. V., et al. Speech synthesis for assistive communication. IEEE Signal Processing Magazine 2019.
22. Li, Y., Chang, M. C., Lyu, S. In Ictu Oculi: Exposing AI created fake videos. IEEE ICASSP 2018.
23. Mauro Fragale and Valentina Grilli, Columbia Journal of European Law; 2024. Retrieved December 2025.
24. Mika Westerlund. The Emergence of Deepfake Technology: A Review November 2019. Retrieved December 2025.
25. Mirsky, Y., & Lee, W. The creation and detection of deepfakes. ACM Computing Surveys, 54(1). 2021.
26. Online links:
27. https://www.congress.gov/bill/118th-congress/house-bill/5586/text, Retrieved December 2025
28. https://www.realitydefender.com/insights/history-of-deepfakes, Gabe Regan 2025. Retrievedd December 2025.
29. Paz Sandoval, Almeida Vau, Solaas, Rodrigues. Threat of deepfakes to the criminal justice system: a systematic review november 2024, Retrieved November 2025.
30. Renee Hobbs. Approaches to Teacher Professional Development in Digital and Media Literacy Education. pg 28, 2017. Retrieved December 2025.
31. Stanford University. AI Index Report. 2024.
32. Tolosana, R., et al. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131–148. 2020.
33. Verdoliva, L. Media forensics and deepfakes. IEEE Journal of Selected Topics in Signal Processing. 2020.
34. Vaccari, C., & Chadwick, A. Deepfakes and disinformation. Social Media + Society. 2020.

35. Žunjić Law. Zunic Law proglašena za kancelariju godine u 2025. 28th of November 2025. Retrieved December 2025.
36. https://www.thalesgroup.com/en/news-centre/press-releases/software-republique-unveils-vision-4rescue-integrated-technological (Accessed 05.12.2025.);
37. https://www.wri.org/insights/cop30-outcomes-next-steps (Accessed 04.12.2025.);
38. https://www.besnet.world/national-ecosystem-assessment/ (Accessed 04.12.2025.);
39. https://rcel.rs/sr/cop29-kljucni-zakljucci/ (Accessed 08.12.2025.);
40. https://pocketproject.org/ (Accessed 09.12.2025.);
41. https://www.unoosa.org/oosa/en/ourwork/spacelaw/index.html (Accessed 13.12.2025.).

Original Scientific Paper/Original naučni rad UDC/UDK: 17:004.8
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600064N

Etika i odgovornost za upotrebu AI: Pregled literature

Dragana Nikolić Ristić[1]*, Violeta Jovanović[2]

[1]Fakultet za menadžment, Metropolitan univerzitet Beograd, dragana.nikolic@metropolitan.ac.rs
[2]Fakultet za menadžment, Metropolitan univerzitet Beograd, violeta.jovanovic@metropolitan.ac.rs

Apstrakt: Brzi i kontinuirani razvoj veštačke inteligencije i njena integracija u skoro svim oblastima donosi veću efikasnost i mogućnost ekonomskih, društvenih i tehnoloških unapređenja. U isto vreme otvaraju se kompleksna etička pitanja i odgovornost za upotrebu AI. Rad razmatra različite aspekte upotrebe AI, uz uključivanje etičkih i društvenih izazova. Cilj rada je da identifikuje, sistematizuje i uporedi ključna etička pitanja i aspekte odgovornosti za upotrebu AI tehnologija u visokom obrazovanju, marketingu i menadžmentu, kao i da ukaže na pravce budućih istraživanja. Pregled literature je sproveden primenom PRISMA metodologije, obuhvatajući naučne članke objavljene u periodu od 2020. do 2025. godine. U radu ističemo značaj integrisanja etičkih principa u procesu upotrebe AI, sa ciljem minimalizacije rizike i maksimizacije prednosti AI. Pitanje etike i odgovornosti za upotrebu AI u daljem razvoju tehnologije bi trebalo biti u fokusu budućih istraživanja.

Ključne reči: etika, AI, obrazovanje, marketing, menadžment.

Ethics and Responsibility in the Use of AI: A Literature Review

Abstract: The rapid and continuous development of artificial intelligence (AI) and its integration into nearly all domains have led to increased efficiency and the potential for significant economic, social, and technological advancements. At the same time, this development raises complex ethical issues and questions of responsibility related to the use of AI. This paper examines various aspects of AI application, with particular emphasis on ethical and societal challenges. The aim of the study is to identify, systematize, and compare key ethical issues and responsibility-related aspects of AI use in higher education, marketing, and management, as well as to indicate directions for future research. The literature review was conducted using the PRISMA methodology and includes scientific articles published between 2020 and 2025. The paper highlights the importance of integrating ethical principles into the process of AI deployment in order to minimize risks and maximize the benefits of AI. Issues of ethics and responsibility in the use of AI should remain a central focus of future research in the further development of this technology.

Keywords: Ethics, AI, education, marketing, management.

1. Introduction

The integration of generative AI chatbots, enabling automated content creation, personalized user experiences, and data-driven decision-making, is revolutionizing all business sectors. The benefits of artificial intelligence to society are undeniable, but the integration of generative AI technology has brought a range of potential opportunities and challenges regarding ethics and security faced by all nations.

The rapid development of technology and the application of artificial intelligence in all spheres of society raise concerns about potential misuse and have spurred global initiatives to regulate issues related to the development and application of artificial intelligence. A large number of countries worldwide are engaged in regulating the development and use of artificial intelligence. In November 2021, 193 UNESCO member states adopted an ethical framework with recommendations for the responsible development and use of artificial intelligence, as the first global standard for ethics in the application of artificial intelligence, whose development also involved Serbia. The European Union Parliament adopted the Artificial Intelligence Act in March 2024. In the United States, the presidential Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence was issued in October 2023. In the United Kingdom, the "Bletchley" Declaration was adopted (an agreement by 28 countries for a safe, secure, and responsible approach to the development and application of artificial intelligence). The United Nations has

adopted the Global Resolution on Safe, Secure, and Trustworthy Artificial Intelligence for Sustainable Development (A/78/L.49).

One of the five goals of the Artificial Intelligence Development Strategy in the Republic of Serbia for the period 2020 - 2025 is highlighted as the ethical and secure application of artificial intelligence (Official Gazette of the Republic of Serbia, No. 96/2019). Preventive mechanisms ensure responsible AI and machine learning development, aligned with high ethical and security standards. Serbia highlights the importance of AI ethics in education, science, the economy, and public administration. The AI Development Strategy (2020–2025) established foundations for adopting Ethical Guidelines for reliable and responsible AI. These guidelines aim to ensure AI development does not marginalize human agency, thought, or decision-making. AI systems must align with the well-being of humans, animals, and the environment, while improving productivity, optimizing resources, and enhancing quality of life (Official Gazette of the Republic of Serbia, No. 23/24). The principles promoted by the Ethical Guidelines include explainability and verifiability, dignity, the prohibition of harm, and fairness. They establish conditions for the development of reliable and responsible artificial intelligence, encompassing: operation and oversight, technical reliability and security, privacy, protection of personal data and data governance, transparency, diversity, non-discrimination and equality, as well as the promotion of social and environmental well-being, and clearly defined accountability. The Artificial Intelligence Development Strategy for the period 2025–2030 lists one of its fundamental goals as encouraging the continuous development of scientific research, innovation, education, economic growth, and the enhancement of citizens' quality of life. It lays the groundwork for creating solutions to numerous ethical challenges and preventing potential misuse of artificial intelligence (Official Gazette of the Republic of Serbia, No. 5/2025).

Privacy protection and potential misuse of artificial intelligence are becoming increasingly important ethical issues. The adoption of clear regulations, investment in education and technology, along with the application of an ethical approach to development at the level of individual states, combined with global cooperation, can enable artificial intelligence to reach its full potential. It is clear that artificial intelligence can serve humanity and benefit everyone exclusively through balanced use and the application of ethical principles.

2. Methodology

The paper examines various aspects of AI use, incorporating ethical and social challenges and its impact on user trust through a systematic literature review. The importance of considering and practically applying ethical issues and responsibilities in the use of AI technologies is analyzed. The results of the paper provide answers to the posed research questions:

RQ1. What are the most frequently identified ethical dilemmas and challenges in the literature on the application of artificial intelligence in higher education?

RQ2. What are the dominant ethical risks and responsibilities associated with the application of artificial intelligence in marketing and management?

RQ3. What research gaps and future research directions emerge from the analyzed literature?

A systematic literature review was conducted using the PRISMA methodology (Knight, 2025; De Leo & Miragliotta, 2025). Inclusion criteria for the PRISMA model:

- The literature review focuses on the application of generic artificial intelligence in the fields of education, marketing, management, and innovation.
- Emphasis on ethical issues in the application of AI technology.
- Scientific papers published in peer-reviewed journals in English from 2020 to 2025.
- Analysis of scientific papers in full text to make inclusion decisions.
- Search for papers using predefined keywords.

Excluded from the review:

- Papers that do not directly address ethical aspects of AI,
- Duplicates and papers without available full text,
- Non-professional publications, reports, and non-peer-reviewed sources.

The search was conducted in the following scientific databases: Google Scholar and ScienceDirect. The search process included papers published between 2020 and 2025. The following key terms were used: "ethical issues," "AI responsibility," "AI in education," "AI in marketing," "AI in management," "generative AI ethics," and "social challenges of AI." The analyzed papers were categorized by research area: education (Yusof et al., 2025; Aljabr

et al., 2024; Hadinejad et al., 2025), marketing (Salih et al., 2025; Kamila & Jasrotia, 2023; Haleem et al., 2022; Khalfallah & Keller, 2025), and management and innovation (Singh et al., 2024; Tzini et al., 2025; Stahl & Eke, 2024). In the initial phase, a total of 82 papers were identified. After removing duplicates and reviewing titles and abstracts, 37 papers underwent detailed analysis. Based on a full-text review and inclusion criteria, 10 papers were ultimately selected for thematic and comparative analysis. The results are presented in a table.

3. Research Results

The results of this study confirm that artificial intelligence is a key factor in digital transformation within the fields of education, marketing, and management. The full potential of artificial intelligence can be realized with a focus on understanding the associated ethical challenges. Table 1 presents the objectives and findings of the research from the analyzed scientific papers.

Table 1. Objectives and Findings of the Research from the Analyzed Scientific Papers

Reference	Objective of the Paper	Research Results
Hadinejad et al., (2025)	Research on how students use AI chatbots during their studies and how they perceive their ethical implications, i.e., how they evaluate AI-generated text.	Non-native English-speaking students, compared to native speakers, use GAI chatbots more for writing support, idea generation, task structuring, paraphrasing, and grammar improvement. This indicates that students view chatbots as personalized learning assistants to overcome linguistic challenges during their studies. Additionally, there are ethical dilemmas regarding plagiarism, the reliability of AI-generated content, and the lack of clear institutional guidelines on the responsible use of AI.
Aljabr et al., (2024)	Assessing teachers' attitudes regarding the adoption of AI technologies as educational tools from ethical and pedagogical perspectives.	Teachers hold positive attitudes towards incorporating AI into the teaching process, as well as a high perception regarding the ethical use and pedagogical implications of AI in the learning process. They recommended the use of AI tools not as replacements for traditional teaching methods, but rather to enhance student learning, while emphasizing the importance of conventional methods.
Yusof et al., (2025)	Analysis of the cognitive and ethical mechanisms of parroting among undergraduate students at a university in Malaysia.	Parroting is primarily driven by extrinsic pressures (unclear instructions, high workload), followed by intrinsic challenges (writing self-confidence, understanding of concepts), and finally by ethical rationalization, which becomes more pronounced when institutional guidelines on AI use are unclear.
Tzini et al., (2025)	Analysis of end-users' intention to seek advice from large language models, the degree of similarity between GPT's and humans' responses to ethical dilemmas, and the assessment of the impact of listing consequences of ethical decisions on encouraging ethical responses from GPT compared to humans.	GPT gives more ethical responses than humans in simple and moderate ethical dilemmas, whether concerning personal or company interests. In complex dilemmas, GPT and humans perform similarly. When using a technique to list consequences, both GPT and humans respond more ethically in personal interest scenarios. For company interests, this technique reduces unethical responses from GPT, but does not affect human responses.
Stahl & Eke, (2024)	Review of the ethical aspects of ChatGPT and similar large language model-based technologies. The study explores the ethical benefits and challenges related to analyzing ChatGPT's ability to generate human-like text and communicate seamlessly.	ChatGPT can provide significant social and ethical benefits, alongside ethical challenges in the areas of social justice, individual autonomy, cultural identity, and environmental issues.

Singh, et al., (2024)	Analysis of the complex relationships between the intention to adopt GenAI technology and its impact on innovation outcomes, competitive advantage acquisition, and the future performance of organizations.	The application of GenAI technology, under the moderating influence of environmental dynamism and ethical dilemmas, can enhance exploratory and exploitative innovations, organizational performance, and competitiveness.
Haleem et al., (2022)	Analysis of various applications of AI in marketing, the transformations that AI causes in the marketing industry, as well as the identification of the most significant ways AI is applied in marketing.	In the field of marketing, AI enables the identification and personalization of relevant content through the collection and analysis of data with the aim of providing the highest quality user experience. AI also assists in the implementation of email marketing and campaigns, as well as creating a more personalized brand experience, thereby increasing user engagement and loyalty. The most significant advantage of applying AI in marketing is data processing, providing marketing professionals with concrete results based on real data.
Khalfallah & Keller (2025)	Analysis of the impact of virtual influencers on consumer trust, perceived authenticity, engagement, and ethical issues related to transparency and consumer deception.	There is potential for virtual influencers to enhance brand engagement, but also concern that significantly affects consumer perceptions regarding authenticity, transparency, regulatory compliance, and cultural sensitivity, complicating their integration into marketing strategies.
Kamila & Jasrotia (2023)	Identifying ethical challenges when creating marketing strategies and the importance of responsible marketing in building consumer trust and loyalty.	The increasing application of artificial intelligence, automation, and digital channels in marketing raises ethical dilemmas such as algorithmic bias, data privacy protection, and the loss of personalization.
Salih et al., (2025)	Literature review and analysis of case studies aimed at identifying the advantages and challenges of GAI (ChatGPT, DALL-E, MidJourney, Jasper.ai, and Synthesia) in contemporary digital marketing.	GAI enhances marketing automation, facilitates user engagement, and strengthens brand interaction, leading to greater customer satisfaction, higher conversion rates, and improved campaign performance. Coca-Cola, Sephora, and Starbucks have confirmed increased efficiency and innovation through the use of GAI. On the other hand, the application of GAI raises concerns regarding data privacy, ethical dilemmas, employee resistance, quality control, and infrastructure limitations.

Source: Authors

The effectiveness of AI technology and technological innovation alone is insufficient and cannot be viewed in isolation without considering ethical issues. An integrated approach that can establish a balance between business objectives and the long-term sustainable interests of society is essential. Careful consideration of ethics and the responsible use of AI technology are the foundation for the future development of a responsible, transparent, and ethically grounded digital world.

4. Discussion and Future research Directions

The digital age has brought about a revolution that enables continuous innovations and transformations in business models and modern society. In line with the objective of this paper, we highlight the future research directions regarding ethical issues and the responsible use of AI technologies in higher education, marketing, and management, as identified by the researchers whose work we analyzed.

For the ethical application of AI technologies, in accordance with academic values, it is crucial to define clear guidelines on citation, acceptable and unacceptable uses of AI, and transparent expectations regarding students' academic integrity (Yusof et al., 2025). The study by Hadinejad et al. (2025) finds that as students increasingly use chatbots to aid learning, ethical issues are also rising. These include problems with academic integrity, plagiarism, fake references, reduced originality, and academic dishonesty. The results highlight the need for

clearer institutional policies on the ethical use of AI chatbots in higher education. Institutions must balance leveraging AI's benefits with fostering students' critical and independent thinking. A recommendation from the work of Aljabr et al. (2024) is the need to train students in the ethical use of AI, the use of AI-based plagiarism detectors, and the reformulation of assessment systems, with the goal of making both teachers and students familiar with safer AI use in the learning process.

Salih et al. (2025) conclude that the application of generative artificial intelligence is not a minor technological innovation in marketing, but rather a revolutionary force shaping how companies communicate, innovate, and compete in the market, requiring adequate strategies to achieve a balance between business opportunities and ethical challenges. The results of Singh et al. (2024) highlight the potential of generative artificial intelligence, as one of the most popular AI technologies, which generates various types of content (music, text, images, synthetic data, etc.), and raises ethical dilemmas for organizations regarding the areas in which AI technology should be used. Cognitive functions that require human intelligence - learning, reasoning, and interacting with the environment - can now be performed by machines thanks to the application of artificial intelligence. With the input of an increasing amount of data, the algorithm learns from previous experiences, improving performance and accuracy (Haleem et al., 2022). A review of 51 scientific research papers addressing virtual influencers, as a transformative force in digital marketing, confirms that the trust and engagement model of virtual influencers clarifies the relationships between authenticity, disclosure transparency, and cultural differences in consumer reactions to virtual influencers, pointing to trust as a key mediating factor (Khalfallah & Keller, 2025). Kamila & Jasrotia (2023) emphasize that the increasing application of artificial intelligence, automation, and digital channels in marketing triggers ethical dilemmas such as algorithmic bias, data privacy protection, and the loss of personalization. Tzini et al. (2025) stress the importance of developing ethics-based policies for the application of AI in organizations, aiming to achieve a balance between AI application and human oversight in complex business situations with significant ethical challenges. They also propose concrete steps to improve decision-making protocols using the consequence enumeration technique, which enables obtaining more ethical responses. Stahl & Eke (2024) point out that the current discussion on the ethics of ChatGPT is one-sided, focusing on only certain issues and lacking a balance between considering ethical benefits and challenges. A rebalancing is needed to overcome the ad-hoc approach dominating current works on ChatGPT ethics. The paper highlights the application of a holistic ethical perspective, emphasizing the need to view the entire socio-technical ecosystem of artificial intelligence, rather than just an individual problem, to encourage positive outcomes during application development while simultaneously identifying ethical downsides.

Through a careful analysis of the literature addressing ethical issues and the responsible use of AI technologies in education, marketing, and management, we obtained answers to the posed research questions.

RQ1. What are the most frequently identified ethical dilemmas and challenges in the literature on the application of artificial intelligence in higher education?

Based on the analyzed works, several key ethical dilemmas regarding the application of artificial intelligence in higher education have been identified, which are related to the transformation of academic norms and values. Significant challenges pertain to issues of academic integrity, plagiarism, authorship, reduced originality and academic dishonesty, the reliability and accuracy of AI-generated content, and the lack of clear institutional guidelines and policies (Hadinejad et al., 2025; Yusof et al., 2025; Aljabr et al., 2024).

RQ2. What are the dominant ethical risks and responsibilities associated with the application of artificial intelligence in marketing and management?

In the fields of marketing and management, the dominant ethical challenges stem from the ability of AI technologies to collect, analyze, and use vast amounts of consumer data. Significant challenges relate to privacy and data protection, algorithmic bias, loss of personalization, individual autonomy, cultural identity, and the development of ethics-based policies for the application of AI in organizations (Kamila & Jasrotia, 2023; Stahl & Eke, 2024; Haleem et al., 2022; Salih et al., 2025; Tzini et al., 2025).

RQ3. What research gaps and future research directions emerge from the analyzed literature?

The literature analyzed in this paper points to several research gaps. Although the number of papers addressing ethical issues in the use of AI technology is increasing, most existing research pertains to descriptive case studies or systematic literature reviews. There is a lack of quantitative empirical studies and contextually and culturally sensitive research. Future research could reduce this research gap by focusing on the role of responsibility and ethical literacy, while connecting the technological, organizational, and human dimensions of AI application, aiming for a responsible and sustainable application of artificial intelligence in education, marketing, and management.

The paper contains several methodological and research limitations. One limitation is the small number of included papers due to strict criteria. This may affect the breadth of insights. A second limitation concerns the analysis being restricted to publications in English, which could lead to the neglect of relevant works in other languages. Furthermore, only three application areas (education, marketing, and management) were considered, meaning the results cannot be fully applied to other sectors.

5. Conclusion

The digital age and AI technology tools represent a new paradigm of thought, enabling innovations and transformations across various business models. The digital technologies we use daily have changed the way we live, work, communicate, and create value. However, despite the significant positive potential of digital technologies, the downsides of the digital world present a constant challenge. The great power of digital technologies implies great responsibility and ethical application. The answers to the research questions show that in higher education, AI is transforming academic norms and the roles of students and teachers. Regarding marketing and management, risks stem from the increasing automation of decision-making and data processing. Furthermore, there are significant research gaps, particularly concerning empirical studies and the operationalization of ethical guidelines in practice. To adequately address the challenges of using AI technologies in higher education, marketing, and management, it is necessary to develop strategies that promote awareness, careful management, and ethical use of digital resources. Above all, a thoughtful approach to the digital world is required, one that recognizes the benefits of using AI technologies while also exercising caution and attention regarding their destructive power if used without awareness of the consequences. A balance between technological progress and human values is essential, because ultimately, we are all responsible for shaping the future of the digital world.

Literature

1. A/78/L.49, (2024). Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development. Retrieved on 06 Decembry 2025 from https://www.undocs.org/Home/Mobile?FinalSymbol=A%2F78%2FL.49&Language=E&DeviceType=Desktop&LangRequested=False
2. Aljabr, F. S., Al-Ahdal, A. A. M. H., & Hassan, A. (2024). Ethical and pedagogical implications of AI in language education: An empirical study at Ha'il University. *Acta Psychologica*, 251, 104605. https://doi.org/10.1016/j.actpsy.2024.104605
3. De Leo, G., & Miragliotta, G. (2025). Sustainability of autonomous cars: Environmental, social, and economic insights from a systematic review. *Sustainable Production and Consumption*, 60, 159–175. https://doi.org/10.1016/j.spc.2025.09.013
4. Etičke smernice za razvoj, primenu i upotrebu pouzdane i odgovorne veštačke inteligencije (Smernice), 2023, Beograd (Službeni glasnik RS", broj 23/24).
5. Hadinejad, N., Sperling, K., & McGrath, C. (2025). Generative AI chatbots in higher education: Student experiences and perceived ethical challenges. *Computers and Education Open*, 9, 100311. https://doi.org/10.1016/j.caeo.2025.100311
6. Haleem, A., Javaid, M., Qadri, M. A., Singh, R. P., & Suman, R. (2022). Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, 3, 119–132. https://doi.org/10.1016/j.ijin.2022.08.005
7. Kamila, M. K., & Jasrotia, S. S. (2023). Ethics and marketing responsibility: A bibliometric analysis and literature review. *Asia Pacific Management Review*, 28(4), 567–583. https://doi.org/10.1016/j.apmrv.2023.04.002
8. Khalfallah, D., & Keller, V. (2025). Authenticity, ethics, and transparency in virtual influencer marketing: A cross-cultural analysis of consumer trust and engagement—A systematic literature review. *Acta Psychologica*, 260, 105573. https://doi.org/10.1016/j.actpsy.2025.105573
9. Knight, S. (2025). Understanding use of evidence in AI ethics guidelines development through a PRISMA-ETHICS informed scoping review of guidelines. *Computers and Education Open*, 9, 100281. https://doi.org/10.1016/j.caeo.2025.100281
10. Nacionalna platforma za veštačku inteligenciju, Etika u veštačkoj inteligenciji. Preuzeto 06. decembra 2025. sa https://www.ai.gov.rs/tekst/sr/238/etika-u-vestackoj-inteligenciji.php
11. Salih, S., Husain, O., Almohamedh, R. M., Tajelsier, H., Hashim, A. H. A., Elshafie, H., & Motwakel, A. (2025). From ideation to execution: Unleashing the power of generative AI in modern

digital marketing and customer engagement—A systematic literature review and case study. *Array*, 100630. https://doi.org/10.1016/j.array.2025.100630

12. Singh, K., Chatterjee, S., & Mariani, M. (2024). Applications of generative AI and future organizational performance: The mediating role of explorative and exploitative innovation and the moderating role of ethical dilemmas and environmental dynamism. *Technovation*, 133, 103021. https://doi.org/10.1016/j.technovation.2024.103021
13. Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74, 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700
14. Strategija razvoja veštačke inteligencije u Republici Srbiji za period 2020–2025.godine (Službeni glasnik RS, br. 96/2019).
15. Strategija razvoja veštačke inteligencije za period 2025-2030. godine (Službeni glasnik RS, br. 5/2025).
16. The Artificial Intelligence Act. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Retrieved on 06 Decembry 2025 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689
17. The Bletchley Declaration by Countries Attending the AI Safety Summit. (2023). Retrieved on 06 Decembry 2025 from https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023
18. Tzini, K., Illia, L., & Zyglidopoulos, S. (2025). The ethics mirror? Comparing LLM and human responses to ethical dilemmas of varying complexity. *European Management Journal*. https://doi.org/10.1016/j.emj.2025.12.002
19. Yusof, I. J., Ashari, Z. M., Ismail, L. H., & Panadi, M. (2025). "We don't plagiarise, we parrot": Cognitive load and ethical perceptions in higher education written assessment. *Bench Council Transactions on Benchmarks, Standards and Evaluations*, 5(4), 100254. https://doi.org/10.1016/j.tbench.2025.100254

Original Scientific Paper/Original naučni rad UDC/UDK: 658.8:004.8
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600071N

Razumevanje poverenja potrošača u marketing zasnovan na veštačkoj inteligenciji: Kvalitativna analiza emocionalnih reakcija na chatbotove

Esma Nur Cerinan Otovic[1], Murat Aytas[2], Ivana Savic[3]

[1]Faculty of Management Herceg Novi, University Adriatic Bar esmacerinan@gmail.com
[2]Faculty of Communication, Selcuk University murataytas@selcuk.edu.tr
[3]Faculty of Management Herceg Novi, University Adriatic Bar ivana.s7895@gmail.com

Abstract in Serbian: Veštačka inteligencija je aktivan deo marketinškog sveta i aplikacija, posebno za integraciju i komunikaciju sa kupcima. Veštačka inteligencija, koja doprinosi strategiji personalizovanog marketinga koju naglašava današnji marketinški svet i ubrzava komunikaciju, takođe se javlja kao problem koji utiče na poverenje i lojalnost potrošača i opštu prihvaćenost. Ova studija ispituje emocionalne pristupe i reakcije ljudi na chat podržan veštačkom inteligencijom, dok istovremeno istražuje kako proces komunikacije sa chatbotom utiče na formiranje poverenja. Tri različite emocije su se pojavile u okviru ove teme. Dok su brz odgovor i detekcija chat kutija pokretanih veštačkom inteligencijom povećali zadovoljstvo i poverenje, suviše automatizovani odgovori chat kutija i nedostatak emocija, u kombinaciji sa njihovim monotonim pristupom, stvorili su osećaj frustracije i anksioznosti. Štaviše, pokrenuta su i pitanja bezbednosti baze podataka u vezi sa zaštitom ličnih informacija. Jedna zajednička tačka koju su učesnici pomenuli bila je poverenje i zadovoljstvo koje im je pružila sposobnost chat kutija da reše zadatke bez ljudske intervencije. Međutim, njihova ograničena uslužna ponuda, gde je bila potrebna empatija, potkopala je poverenje i stvorila zabrinutost. Ovo postavlja dilemu da li su chat kutije dobro marketinško komunikacijsko sredstvo ili problem poverenja. Ova studija naglašava važnost poverenja, individualnosti i empatije u razvoju odnosa sa kupcima kroz komunikacijski proces uspostavljen marketinškim alatima pokrenutim veštačkom inteligencijom. Dok se sugeriše da se negativni uticaji komunikacijskih alata pokretanih veštačkom inteligencijom mogu ublažiti inkorporiranjem više ljudskih karakteristika, kao što su prirodniji jezik i prilagođavanja, takođe se raspravlja o preciznoj ulozi ovih aplikacija u komunikaciji unutar marketinškog sveta.

Ključne reči: poverenje kupaca, veštačka inteligencija, marketing, chatbot, digitalna komunikacija

Understanding Consumer Trust in Al – Enabled Marketing: A Qualitative Analysis of Emotional Reactions to Chatbots

Abstract in English: Artificial intelligence is an active part of marketing world and applications, especially for integrating and communicating with customers. Artificial intelligence, which contributes to the personalization marketing strategy emphasized by today's marketing world and accelerates communication, also appears as a problem that affects consumer trust and loyalty and general acceptance. This study examines people's emotional approaches and reactions to artificial intelligence- supported chat, while also examining how the communication process with Chatbot affects trust formation. Three distinct emotions emerged within this theme. While the rapid response and detection of Al- powered Chat boxes increased the satisfaction and trust, Chat boxes' overly automated responses and lack of emotion, combined with their monotonous approach, created feelings of frustration and anxiety. Furthermore, database security concerns about protecting personal information were also raised. One common point participants mentioned was the trust and satisfaction that Chat boxes' ability to solve tasks without human intervention gave them. However, their limited-service provision, where empathy was required, undermined trust and created concerns. This raises the dilemma of whether chat boxes are good marketing communication tool or a trust problem, this study highlights the importance of trust, individuality, and empathy in developing customer relationships through the communication process established by Al powered marketing tools. While it is suggested that the negative impacts of AL- powered communication tools can be mitigated by incorporating more human-like features, such as a more natural language and adjustments, it also discusses the precise role of these applications in communication within the marketing world.

Keywords: customer trust, artificial intelligence, marketing, chatbot, digital communication

1. Introduction

Artificial intelligence is transforming the way companies communicate with their customers by providing them with a more personalized, faster, more effective and measurable form of communication. Chatbots, one of the most widely used Al communication robots, streamline communication between companies and their customers with instant responses, measurable solutions, and automated actions. (Chaturvedi et al., 2023) This also raises questions about the extent to which chatbots respond with human qualities, how empathetic they are, and how their lack of emotionally can leave a lasting impression on customers. While companies are leveraging the rapid communication and measurable support Chatbots offer and incorporating these points into their strategies, they are also seeking answers to these questions for more effective communication, amidst the dilemma of how trust-inspiring Chatbots are. Trust is the most crucial factor in marketing and customer communication. Therefore, understanding how trust can be increased when using Al-powered robots in the ever-changing world of customer communication driven by new technologies, is crucial. This study aims to better understand and grasp users' emotional responses to Chatbots and discuss the dilemma of whether Chatbots foster trust or distrust. (Hu et al.,2021) At this point this study is of a visionary nature designed to inspire future researchers in terms of bringing clarity to individuals' customer experience with Chatbots and emphasizing the positive and negative effects of marketing communication via Chatbots on customers.

2. Literature Review

Interpersonal communication is an essential concept that encompasses the most important characteristics of social interaction and understanding. (Araujo, 2018). From a social cognitive theory perspective, people's past attitudes, thoughts, and behaviors are shaped by and linked to their previous experiences, which lead them to form personal inferences (Ghen & Liu, 2004; Doney & Cannon, 1997). Therefore, individuals' perceptions of chatbots are influenced and interpreted through earlier encounters with similar technologies. (Balakrishnan & Dwivedi, 2021). If communication with Al systems is helpful, practical and trustworthy, people are more likely to use those chatbots frequently and develop positive feelings toward them. The more positive emotions users experience when interacting with chatbots, the more trust and loyalty they are likely to feel toward both Al tool and the company behind it. Such positive experiences also enhance social interactions between humans and Al. Research has shown that users who have pleasant experiences with chatbots are less likely to develop negative feelings toward them and are less inclined to blame them directly in the event of service failures. (Roden et al., 2017) Key areas to explore include the advantages and disadvantages of chatbot-assisted communication, the emotional responses associated with user experiences, areas for improvement, and the main factors shaping the chatbot- trust relationship.

In today's marketing environment, it is evident many brand- such as Coca-Cola, Gucci, Louis Vuitton- actively use chatbots to engage with customers and provide 24/7 service. This constant availability allows users to access company services without interruption and enables faster communication regarding issues and solutions. (Chung et al.,2020) This not only reduces response times, offering customers quicker and more convenient interactions, but also lowers financial costs bey reducing the need for human support staff. One study estimated that chatbot usage in the United States could reduce company expenses by up to $8 billion annually. (Ashfaq et al. ,2020) However, this shift has also generated negative reactions among some users, including difficulties in establishing trust, a desire for human support, concerns about data protection, and a perceived lack of empathy. A loss of trust caused by various factors can lead to customer attrition, communication breakdowns, and the abandonment of products or Al tools. (Denecke et al., 2021) Research conducted by Meyer (2020) unmet customer needs significantly increase anxiety and stress, which in turn intensifies distrust toward chatbots. For example, research on automated chatbot phone calls in sales contexts found that chatbot-driven communication was significantly less effective at generating sales compared to human communication. This suggests that users strongly prefer human interaction, especially in purchasing and decision-making situations. To understand this phenomenon, it is important to examine the factors that influence both positive and negative sides of Al- mediated communication. (Luo et al. ,2019).

Another study concluded that the use of emojis in chatbot communication can create more positive perception among users. Customers viewed chatbots that respond with emojis as warmer and more empathetic. However, the study also found that when a chatbot uses highly formal language, emojis may seem unnatural or inconsistent with the message. In other words. A chatbot's communication style must align with its language use, and emojis are effective only when they match the tone of the message. (Fadhil et al., 2018) Therefore, understanding how trust can be strengthened when using Al-powered systems in a rapidly evolving field of customer communication is

crucial. This study aims to better understand users' emotional responses to chatbots and to examine whether chatbots foster trust or distrust. (Hu et al., 2021)

3. Methodology

3.1. Research Design

In this study, qualitative research is conducted to explain how consumers experience and interpret trust in marketing communications established through artificial intelligence-supported Chatbots.

3.2. Data Collection Method

Data was collected through semi- structured, in-depth interviews with participants who had interacted with Al-powered chatbots for service, information retrieval, customer problem solving purposes in the last six months.

Interview topics are listed under the following headings:

- Emotional responses of users during interactions
- Differences between chatbot and human communication
- A look at the situations that increase and decrease trust
- Situations where empathy and warmth increase and decrease in chatbot conversations
- Perceptions and expectations of empathy and warmth in chatbot conversations

Each interview lasted between 30 and 45 minutes and consisted of video chats and face-to face conversations.

3.3. Sampling Strategy

A sample was created by selecting 15 participants from different age groups and genders, taking into account different internet usage habits, who had been involved in customer communication via chatbots in the last 6 months.

3.4. Data Analysis

The interviews were transcribed verbatim as an in-depth question- and- answer session and the analysis was conducted using thematic analysis. This process included rereading the transcripts, assigning responses through coding, developing themes, examining each theme, and explaining the key points of each theme. NVivo was used for coding and theme section. The aim was to determine the emotional responses, communication, and empathy expectations of consumers who interacted with Al- powered chatbots, and to identify and explain their preferences and patterns that were prevalent when building trust.

4. Findings

The conclusion drawn from the qualitative findings is that there are a complex relationship and interaction between trust, emotional perception, and communication preferences and style in artificial intelligence-supported sharing communication.

4.1. The Multidimensional Concept of Trust

Participants identified that trust is influenced both human- centred and technology centred factors. In this context. Participants emphasized the following points regarding developing trust in chatbots

- A transparent discussion is necessary regarding how chatbot work,
- Accuracy and consistency of assessments are important.
- Perceptions can vary from person to person. Manipulation of chatbots will have different effects on everyone. These perceptions which require clarification regarding the manipulation of Al, should be addressed from a technological- technical and human- centered perspective, encompassing both physical, ethical and transparency processes, as stated in the research of Deneke in 2021, and it is directly proportional to how human-centered communication is presented.

4.2. Emotional Disconnections

Many interviews drew a distinction between objective and contextual issues. While chatbots were trusted for simple tasks and reportedly generated a high level of satisfaction, participants reported developing lack of comfort and trust with more subjective and complex questions. This was attributed to the lack of subjective perception of chatbot, its inability to fully capture empathetic human qualities, and its language use. The reactions encountered and expressed in the interviews were particularly related to the inability of chatbots to respond to cognitive and emotional responses. At this point, we see that emotional responses create a feeling of frustration and disappointment which is related to the feeling of being misunderstood and the feeling that the answer is given without any attention.

4.3. Language Styles and Empathetic Approach of Chatbots

T While a warm an approachable tone of voice increases trust in Al- enabled marketing. A robotic and cold tone of voice damages the sense of trust, because this gives the participants the feeling that they are not talking to a real person. Participants stated that they felt their expectations were not met, their personal data was not protected and they even felt like they were being defrauded.

5. Discussion and Conclusion

Trust is a crucial and often emphasized aspect of human communication, influencing not only the adoption of technological tools but also marketing communication strategies. Just as trust shapes every interaction with technology, it is essential for companies and manufacturers to understand how to strengthen it. Research shows that excessive trust in artificial intelligence- enabled by emerging technologies- can lead to machine dependency. Conversely, undermining trust through various factors can result in customer loss, communication breakdowns, and the abandonment of relevant products or Al tools. The relationship between Al and trust must therefore be considered not only from an ethical perspective but also from technical one. Trust in Al involves legal and technical dimensions, including transparency of information about Al systems, their performance, and the methods used to evaluate them. Al separates itself from other technological tools through its ability to learn, synthesize diverse information, and generate increasingly autonomous approaches. Trust can be analyzed under the main categories: human centered, context centered and technology centered. These perspectives vary by individual, for example, a person with more adaptable trust tendencies may quickly adopt and easily trust Al- supported tools.

Trust in Al- powered chatbots depends heavily on communication transparency, the quality of responses, and the degree of empathy conveyed, Given the productivity, creativity and speed of Al- based tools- as well as their ability to perceive and address user needs- the question of how individuals approach Al-powered systems has become a important area of study. Research reveals two distinct user orientations: objective and subjective. Users tend to prefer Al- powered robots for conversations that involve objective information or routine tasks- such as making or cancelling reservations, or purchasing flight tickets. However, they express a need for human presence in more personalized or emotionally nuanced interactions, indicating a lack of trust in Chatbots for subjective conversations. From this perspective, offering transparent and accessible information increases trust in Al tools. At the same time, this research show that the language and approach used by Al can be more manipulative for certain individuals, meaning the issue must be approached from both an objective and subjective standpoint.

Studies examining customer communication through chatbots have found that even when human-like qualities are added, individuals still tend to prefer human based communication. Chatbot driven communication was significantly less effective in generating sales than human communication. This suggests that users strongly favor human interaction, especially in purchasing and decision-making processes. To understand this, it is important to examine the factors influencing both positive and negative aspects of Al- mediated communication.

Warmth perception- how close, empathetic, and friendly a company or organization appears- is closely linked to the positive feelings experienced by customers and the trust that follows. Conversely, customers react negatively when their emotional or cognitive needs are not met. This research shows that when customers use technology to communicate with companies, if they feel they are wasting time, that their needs are unmet, or that they are not being understood; such experiences lead to emotions like disappointment and uncertainty, creating negative perceptions of the company and harming the effectiveness of customer communication. Customer needs remained unfulfilled, anxiety and stress increased significantly, which in turn heightened distrust toward chatbots.

Emojis used in chatbot communication can create a more positive perception among users. Customers found chatbots that respond with emojis to be warmer and more empathetic. However, the study also found that when a chatbot uses highly formal language, emojis can appear unnatural and inconsistent with the text. In the other words, a chatbot's communication style must align with its language use, and emojis are effective only when they complement the tone of the message.

Therefore, the results of this study suggest that when determining future marketing communication strategies, the use and adoption of the latest technology, as well as the emotional impact this trend has on customers and companies, and the impact of technologically- enhanced communication on people, should be examined. While it can be argued that the positive effects of Al- enhanced communication can be mitigated by a more natural language approach imbued with more human characteristics, it also emphasizes that providing quick and accurate answers to perceived consumer questions can increase trust and reduce consumer anxiety.

LITERATURE

1. Alsharhan, A., Al-Emran, M., & Shaalan, K. (2023). Chatbot adoption: A multiperspective systematic review and future research agenda. *IEEE Transactions on Engineering Management, 71*, 10232–10244. https://doi.org/10.1109/TEM.2023.3298360
2. Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior, 85*, 183–189. https://doi.org/10.1016/j.chb.2018.03.051
3. Ashfaq, M., Yun, J., Yu, S., & Loureiro, S. M. C. (2020). I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics, 54*, 101473. https://doi.org/10.1016/j.tele.2020.101473
4. Balakrishnan, J., & Dwivedi, Y. K. (2021). Role of cognitive absorption in building user trust and experience. *Psychology & Marketing.* https://doi.org/10.1002/mar.21462
5. Barari, M., Ross, M., & Surachartkumtonkun, J. (2020). Negative and positive customer shopping experience in an online context. *Journal of Retailing and Consumer Services, 53*, 101985. https://doi.org/10.1016/j.jretconser.2019.101985
6. Chaturvedi, R., et al. (2023). Social companionship with artificial intelligence: Recent trends and future avenues. *Technological Forecasting and Social Change.*
7. Chung, M., Ko, E., Joung, H., & Kim, S. J. (2020). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research, 117*, 587–595. https://doi.org/10.1016/j.jbusres.2018.10.004
8. Doney, P., & Cannon, J. P. (1997). An examination of the nature of trust in buyer–seller relationships. *Journal of Marketing, 61*(2), 35–51.
9. Fadhil, A., Schiavo, G., Wang, Y., & Yilma, B. (2018). The effect of emojis when interacting with conversational interface assisted health coaching system. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM. https://doi.org/10.1145/3240925.3240965
10. Ghen, K. J., & Liu, G. M. (2004). Positive brand extension trial and choice of parent brand. *Journal of Product and Brand Management, 13*(1), 25–36.
11. Hu, P., et al. (2021). Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior.*
12. Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science, 38*(6), 937–947. https://doi.org/10.1287/mksc.2019.1192
13. Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change, 175*, 121390. https://doi.org/10.1016/j.techfore.2021.121390
14. Meyer, P., Jonas, J. M., & Roth, A. (2020). Frontline employees' acceptance of and resistance to service robots in stationary retail: An exploratory interview study. *SMR – Journal of Service Management Research, 4*(1), 21–34. https://doi.org/10.15358/2511-8676-2020-1-21
15. Roden, S., Nucciarelli, A., Li, F., & Graham, G. (2017). Big data and the transformation of operations models: A framework and a new research agenda. *Production Planning & Control, 28*(11–12), 929–944.

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:616-07
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600076M

Veštačka inteligencija u medicinskoj dijagnostici: Kritički pregled rizika, odgovornosti i epistemoloških ograničenja velikih jezičkih modela

Marjan Marjanović[1], Luka Latinović [2]

[1] Clinic for General, Visceral and Thoracic Surgery - InnKlinikum, Altötting, Germany
[2]Belgrade School of Engineering Management, Beopolis University, Belgrade, Serbia, luka.latinovic@fim.rs

Sažetak: Brza integracija veštačke inteligencije u dijagnostičku medicinu predstavlja istovremeno tehnološki napredak i epistemološki poremećaj. Ovaj narativni kritički pregled ispituje kako veliki jezički modeli i srodni sistemi veštačke inteligencije utiču na dijagnostičko rasuđivanje, kliničku odgovornost i medicinsku etiku. Iako je veštačka inteligencija pokazala izuzetan potencijal u prepoznavanju obrazaca i sintezi podataka, njena logika zasnovana na korelacijama lišena je uzročno-posledičnog razumevanja, interpretabilnosti i moralne odgovornosti. Rad identifikuje ključne rizike kao što su automatizaciona pristrasnost, pristrasnost skupa podataka, neprozirnost modela i asimetrija odgovornosti između programera i lekara. Integracija veštačke inteligencije izaziva tradicionalne granice profesionalne odgovornosti i informisanog pristanka, dovodeći u pitanje epistemičku validnost i poverenje pacijenata. Postojeći okviri upravljanja, prilagođeni statičkoj regulativi medicinskih uređaja, pokazuju se neadekvatnim za sisteme koji kontinuirano uče. Istraživanje pokazuje da očuvanje integriteta medicinskog rasuđivanja u doba veštačke inteligencije zahteva obnovljenu posvećenost epistemičkoj skromnosti, raspodeljenoj odgovornosti i očuvanju ljudskog suda u okviru algoritamski posredovane zdravstvene nege.

Ključne reči: epistemički integritet, dijagnostičko rasuđivanje, medicinska odgovornost, pristrasnost automatizacije, regulatorna etika.

Artificial Intelligence in Medical Diagnostics: A Critical Narrative Review of Risks, Responsibility, and the Epistemological Limits of Large Language Models

Abstract: The rapid integration of artificial intelligence (AI) into diagnostic medicine represents both a technological advance and an epistemological disruption. This narrative critical review examines how large language models and related AI systems influence diagnostic reasoning, clinical accountability, and medical ethics. While AI has demonstrated remarkable potential in pattern recognition and data synthesis, its correlation-based logic lacks causal understanding, interpretability, and moral agency. The review identifies key risk domains including automation bias, dataset bias, model opacity, and liability asymmetries between developers and physicians. It argues that AI's integration challenges traditional boundaries of professional responsibility and informed consent, raising concerns over epistemic validity and patient trust. Current governance frameworks, adapted from static medical device regulation, remain ill-suited for continuously learning systems. The research shows that safeguarding the integrity of medical reasoning in the age of AI requires a renewed commitment to epistemic humility, distributed accountability, and the preservation of human judgment within computationally mediated care.

Keywords: epistemic integrity, diagnostic reasoning, medical accountability, automation bias, regulatory ethics.

1. Introduction

The deployment of artificial intelligence (AI) in clinical medicine has advanced rapidly in recent years, propelled by improved computational capacity, expanding digital-health datasets, and the emergence of large language models (LLMs) capable of complex pattern recognition, natural language processing and decision-support tasks (Fahim et al., 2025; Graili & Farhoudi, 2025; Krishnan et al., 2023). For instance, reviews report that AI systems

are widely applied in diagnostic workflows—spanning medical imaging, biosignal interpretation, electronic health-record analytics and predictive modelling of disease onset or progression (Sadr et al., 2025; Sinha, 2024). This growth promises considerable benefits: faster detection of pathology, more consistent triage, support for clinicians in assimilating broad swathes of data, and ultimately, improved patient-outcomes (McGenity et al., 2024; van Diest et al., 2024).

Nevertheless, alongside this promise there are substantive risks and unresolved questions—especially when AI is used to assist or partially automate clinical decision-making such as diagnosis (Chustecki, 2024; Khosravi et al., 2024a). Critically, while clinicians remain legally and ethically responsible, AI tools such as LLMs increasingly influence those decisions (Jones et al., 2023; MacIntyre et al., 2023; Pham, 2025). In that context, it is imperative to examine the epistemological foundations (how diagnoses are arrived at), the clinical and ethical ramifications (when AI errs), and the accountability structures (who bears responsibility). The diagnostic domain is particularly sensitive because diagnostic error has direct patient-harm implications. Medical literature indicates that many AI diagnostic systems are trained and validated under idealised conditions, using curated datasets and retrospective designs; their performance in real-world heterogeneous clinical settings often remains uncertain (Angus et al., 2025; Sourlos et al., 2024). Moreover, many LLM-based systems operate as "black boxes" with limited transparency of how inputs map to outputs, raising concerns about trust, interpretability and clinician over-reliance (automation bias) (Savage et al., 2024; Ullah et al., 2024). These features generate a latent danger: doctors relying on AI-driven suggestions may defer too readily to the model, reducing their own critical scrutiny.

Equally important is the question of epistemic validity: the AI tool may recognise statistical correlations but lacks genuine causal understanding of pathophysiology, which may lead to plausible-looking but erroneous diagnoses (Sanchez et al., 2022; Xu & Shuttleworth, 2024). As the literature underscores, AI's diagnostic suggestions must be viewed as probabilistic, supportive, rather than definitive; yet in practice, the clinician's stamp often converts the suggestion into a formal diagnosis with attendant legal/ethical weight (Alowais et al., 2023; Bozyel et al., 2024).

In this narrative review, we focus specifically on the risks associated with the use of LLMs and other AI tools in medical diagnostics, in contexts where a physician remains the final accountable decision-maker. We analyse existing literature on diagnostic performance, error modes, bias, interpretability, workflow integration and responsibility. Our aim is to provide a critical account of how these tools interplay with clinical decision-making, where their limitations lie, and how responsibility and epistemic integrity might be preserved in an era of AI-augmented diagnosis.

1.1. Approach and Scope

This paper adopts a **narrative critical review** design rather than a systematic or scoping review, owing to the conceptual nature of the topic and the interpretive orientation required to assess epistemological and ethical risks in AI-assisted diagnostics. The goal is not to quantify evidence across studies but to critically examine how artificial intelligence—particularly large language models (LLMs)—reconfigures clinical reasoning, medical accountability, and diagnostic epistemology. The narrative review approach is appropriate where the research problem transcends purely empirical boundaries and involves theoretical, ethical, and philosophical considerations that cannot be meaningfully reduced to numerical synthesis (Greenhalgh et al., 2018; Sukhera, 2022). In this case, diagnostic accuracy metrics, error rates, and algorithmic benchmarks, while informative, are insufficient to illuminate how AI influences the clinician's judgment, responsibility, and interpretive autonomy.

The review draws upon peer-reviewed journal articles, conceptual papers, policy reports, and interdisciplinary analyses published in the last half-decade across medicine, cognitive science, bioethics, and science-and-technology studies. These sources were identified through purposive reading rather than database-driven screening, ensuring inclusion of perspectives that critically interrogate not only performance outcomes but also the epistemic underpinnings of diagnostic reasoning. The critical-interpretive orientation of this review entails three guiding premises:

1. **Contextual evaluation** – each source is interpreted in relation to the broader clinical, legal, and epistemological context in which AI systems operate.
2. **Conceptual synthesis** – emphasis is placed on tracing how recurring themes—such as automation bias, opacity, and accountability—interrelate across disciplines.
3. **Normative reflection** – beyond describing empirical findings, the analysis interrogates the moral and epistemic consequences of delegating diagnostic reasoning to non-human agents.

By integrating these perspectives, the narrative review allows for a coherent critical account of AI's role in diagnostic medicine—an account that would be obscured by the procedural constraints of systematic evidence aggregation. This interpretive flexibility is essential for capturing the complexity of emerging clinical–technological assemblages where algorithms increasingly participate in, but do not yet own, the act of medical judgment.

2. Artificial Intelligence in Clinical Diagnostics

The integration of artificial intelligence into clinical diagnostics represents a decisive inflection point in the evolution of medical practice. Traditionally, diagnostic reasoning has relied upon the physician's capacity to synthesise heterogeneous information—symptom presentation, medical history, laboratory findings, and imaging data—into a coherent clinical hypothesis. The arrival of AI systems has introduced computational models capable of performing many of these interpretive functions autonomously or semi-autonomously, thus altering both the epistemological structure of diagnosis and the practical organisation of healthcare delivery (D'Adderio & Bates, 2025; Saenz et al., 2023; Wu et al., 2024).

AI's involvement in diagnostics initially emerged through **rule-based expert systems**, designed to emulate decision trees reflecting human medical reasoning (Alowais et al., 2023; Kulikowski, 2015; Solutions, 2024). These early systems proved rigid, as their performance depended heavily on the explicitness and completeness of the encoded knowledge base. The contemporary landscape is instead dominated by **data-driven approaches**, most notably deep learning architectures and large language models (Wang & Zhang, 2024; Zhou et al., 2025). These systems infer diagnostic associations from vast quantities of data—imaging repositories, genomic databases, or electronic health records—without presupposing an explicit theoretical model of disease.

This shift from symbolic to sub-symbolic reasoning has generated unprecedented diagnostic capabilities. Machine vision models now achieve or exceed human-level accuracy in detecting malignancies on radiographs, histopathological slides, and dermatological images (Jeong et al., 2022; Krakowski et al., 2024; McGenity et al., 2024; Waqas et al., 2023). Predictive models assist in early identification of cardiovascular, metabolic, and neurodegenerative conditions by integrating multi-modal data streams (DeGroat et al., 2024; Xue et al., 2024). Most recently, large language models have entered the diagnostic domain through their ability to process unstructured clinical narratives, suggest differential diagnoses, and assist in generating clinical documentation (Nazi & Peng, 2024; Singhal et al., 2023; Vrdoljak et al., 2025; Zhou et al., 2025). Their flexible linguistic competence allows them to operate across specialties, functioning as conversational diagnostic aides that simulate expert dialogue.

However, this transformation also entails a profound **epistemic displacement**. Whereas human diagnostic reasoning is anchored in pathophysiological understanding and clinical context, AI models operate through correlation-based pattern recognition (Boge & Mosig, 2025; Tikhomirov et al., 2024). They may produce correct answers for the wrong reasons, or plausible but incorrect answers that appear clinically credible. The opacity of neural-network inference processes precludes meaningful introspection into why a particular diagnosis was proposed, complicating both verification and accountability.

In practice, AI is increasingly positioned as a co-participant in diagnostic reasoning (Goh et al., 2024). Clinical workflows now involve hybrid arrangements in which AI systems pre-screen imaging studies, flag anomalies, or generate ranked diagnostic suggestions for the physician's review. Such configurations ostensibly preserve human oversight, yet they also shift the cognitive landscape of clinical decision-making (Korfiatis et al., 2025). The physician's role transitions from primary diagnostician to meta-analyst of algorithmic output. This role change carries consequences for professional identity, epistemic authority, and legal responsibility.

Another emergent dynamic concerns **automation bias**, wherein clinicians may develop an uncritical trust in algorithmic recommendations, particularly when confronted with time pressure or cognitive fatigue (Abdelwanis et al., 2024; Goh et al., 2025; Khosravi et al., 2024b). Conversely, distrust of AI systems—especially following a perceived error—may lead to over-correction and rejection of potentially valuable input (Rosenbacke et al., 2024; Tun et al., 2025). The dialectic between overreliance and overcaution illustrates that AI's integration into diagnostic medicine is not merely a technical implementation challenge but a cognitive and cultural transformation of medical epistemology itself.

Therefore, the advance of AI in clinical diagnostics cannot be understood solely as a technological progression. It represents a reconfiguration of how diagnostic knowledge is produced, validated, and acted upon. The subsequent sections of this paper address these transformations by examining the epistemological limits of AI reasoning, the

associated risk domains, and the evolving moral and legal frameworks through which medical responsibility is being renegotiated.

2.1. Epistemological Foundations of AI-Driven Diagnosis

The epistemological foundations of medical diagnosis have always rested on the clinician's capacity to connect empirical observation with theoretical understanding. Diagnostic reasoning traditionally involves a dialectical movement between inductive inference from patient-specific findings and deductive application of biomedical knowledge (Shin, 2019). In contrast, artificial intelligence—particularly large language models and deep learning systems—operates through statistical correlation rather than causal comprehension. This difference marks a fundamental epistemic rupture: AI systems recognise patterns in data but lack awareness of what those patterns represent within the biological or clinical ontology of disease (Nicholson et al., 2023).

2.2. Correlation versus Causation in Algorithmic Reasoning

While human diagnostic thought aspires to causal explanation, AI models are optimised for predictive accuracy. They learn associations that maximise statistical fit, not explanatory coherence. This epistemic asymmetry means that an algorithm may correctly predict that a given constellation of symptoms aligns with a particular pathology, yet remain oblivious to the underlying mechanisms producing those symptoms. As such, its "knowledge" is non-conceptual; it consists of distributed weights within a network, not propositions about biological reality. Consequently, even when AI models achieve high diagnostic accuracy, they do not possess the epistemic properties that confer medical understanding—intentionality, contextual judgment, or counterfactual reasoning.

This limitation has practical implications. When the underlying data are biased, incomplete, or unrepresentative, correlations learned by the model may amplify pre-existing distortions (Cross et al., 2024; Mittermaier et al., 2023). A model trained primarily on imaging data from one demographic group may misinterpret normal variations in another as pathological. Because the system does not "know" why its prediction holds, it cannot recognise when its reasoning falls outside valid clinical context (Koçak et al., 2025; Tejani et al., 2024; Vrudhula et al., 2024). In human medicine, such epistemic self-monitoring—awareness of uncertainty, reflection on plausibility, and revision of hypotheses—is integral to diagnostic reliability. AI systems, by contrast, lack meta-cognitive capacity; they cannot interrogate their own inferences.

2.3. The Problem of Opacity and the "Black Box" Dilemma

A defining challenge of deep learning and LLM-based diagnostics is their opacity. The internal operations of these models, involving billions of parameters, are not interpretable in a way that corresponds to human reasoning (Chen et al., 2022; Zhao et al., 2024). The clinician cannot meaningfully trace how input variables—symptoms, images, test results—lead to a specific diagnostic output. This "black box" character undermines transparency and complicates trust. In medicine, where the justification of a diagnosis must be communicable and defensible, opacity introduces epistemic risk. A diagnostic conclusion whose provenance cannot be articulated cannot be ethically or legally justified (Freyer et al., 2024).

Attempts to enhance interpretability through "explainable AI" frameworks offer partial solutions, but these often provide post hoc rationalisations rather than genuine insight into the model's inferential process (Acun & Nasraoui, 2025; Sadeghi et al., 2024). Visual heatmaps, saliency maps, and token importance rankings may show where the model "focused," but they do not explain *why* it reached a particular conclusion (Rao & Aalami, 2023; Saarela & Podgorelec, 2024). Thus, even well-intentioned efforts to make AI more interpretable often remain epistemologically superficial.

2.4. Epistemic Authority and the Physician's Cognitive Displacement

The introduction of AI diagnostic tools has also begun to redistribute epistemic authority within clinical settings. When algorithmic recommendations are integrated into decision support systems, they implicitly acquire a status of evidentiary credibility (Jones et al., 2023). Physicians, conscious of AI's statistical prowess, may defer to its suggestions, especially in ambiguous cases (Kostick-Quenet & Gerke, 2022). This subtle shift can lead to a **cognitive displacement**, wherein the clinician's reasoning becomes secondary to the machine's inference (Patil et al., 2025; Tikhomirov et al., 2024). Over time, such displacement may erode the epistemic autonomy that has historically defined medical professionalism.

This transformation also challenges the social contract between physician and patient. The act of diagnosis traditionally entails a moral responsibility grounded in human judgment—the clinician stands accountable for both error and care. When part of that reasoning is delegated to an opaque computational entity, the ethical foundation of that accountability becomes unstable. The physician may still bear formal responsibility, but the epistemic control over the decision has become distributed and partially inaccessible. This asymmetry—responsibility without full epistemic authority—creates an untenable moral configuration in which practitioners are held accountable for outcomes they cannot fully understand or verify. The epistemological tension between human and machine reasoning, therefore, lies at the core of AI-assisted diagnostics. While AI systems can expand the range and precision of pattern recognition, they simultaneously displace the causal and interpretive foundations upon which medical diagnosis depends. The next section examines how these epistemic vulnerabilities manifest as concrete risks—statistical, procedural, and ethical—within real-world clinical environments.

3. Risk Domains in AI-Assisted Diagnostics

Artificial intelligence, while promising in diagnostic medicine, introduces a spectrum of risks that extend beyond technical error. These risks manifest at the intersection of epistemology, human cognition, and institutional responsibility. Each domain reflects a distinctive way in which the deployment of AI challenges the reliability, safety, and ethical legitimacy of clinical decision-making.

3.1. Diagnostic Error and Model Hallucination

AI diagnostic systems are not immune to error; in fact, their errors are often opaque and systematic (Xu & Shuttleworth, 2024). Deep learning models and large language models are prone to what has been described as *hallucination*—the confident generation of incorrect information or conclusions (Asgari et al., 2025; Chelli et al., 2024). In a medical context, such hallucination can produce diagnostic suggestions that are linguistically convincing yet clinically false. Unlike human reasoning, which typically encodes some awareness of uncertainty, AI systems tend to express outputs with unwarranted confidence, creating an illusion of precision (Afroogh et al., 2024; Rezaeian et al., 2025). This dynamic heightens the risk of misdiagnosis when clinicians interpret machine-generated statements as authoritative.

Another source of diagnostic error stems from dataset bias. Training data frequently overrepresent certain demographic groups, institutions, or disease categories, resulting in models that generalise poorly across populations (Koçak et al., 2025). Errors can thus disproportionately affect under-represented groups, compounding existing health inequities. Because many models are developed in research settings with highly controlled inputs, their apparent performance often degrades under real-world clinical variability—differences in imaging devices, patient populations, and record-keeping conventions (Wellnhofer, 2022; Yang et al., 2024).

3.2. Data Bias and Representational Inequities

The epistemic core of AI diagnostics rests on data representation, and representation is never neutral (Pagano et al., 2023; Stinson, 2022). Every training dataset reflects historical practices, institutional priorities, and structural inequities (Agarwal et al., 2023; Hanna et al., 2025; Jui & Rivas, 2024; Zajko, 2022). When these inequities are encoded into the model, they become perpetuated at scale. For instance, diagnostic tools trained predominantly on data from high-income regions may fail to recognise pathologies common in low-resource contexts (Celi et al., 2022; Joseph, 2025). Similarly, gender or racial bias in symptom reporting can distort probabilistic inference (Colacci et al., 2025; Mittermaier et al., 2023). The danger lies not only in biased predictions but in the false appearance of objectivity that accompanies algorithmic output.

Such representational distortions are epistemologically pernicious because they are difficult to detect once embedded in the model (Hasanzadeh et al., 2025). Bias audits and fairness metrics offer partial safeguards, yet they remain reactive—addressing observed disparities rather than confronting the social and epistemic structures that produce biased data in the first place. The ethical consequence is that AI may reinforce diagnostic hierarchies under the guise of technological neutrality.

3.3. Automation Bias and Overreliance by Clinicians

Automation bias describes the human tendency to overtrust algorithmic systems, accepting their outputs even when inconsistent with clinical judgment. In diagnostic contexts, this bias is amplified by the authority ascribed

to quantitative or computational processes. When AI systems are presented as advanced, evidence-based, or statistically superior, clinicians may defer to them in situations of uncertainty or cognitive fatigue (Saadeh et al., 2025). The problem is not simply psychological; it reflects a deeper shift in epistemic hierarchy where algorithmic reasoning displaces human interpretive authority.

Overreliance on AI recommendations can erode critical engagement with clinical evidence (Klingbeil et al., 2024). Physicians might shortcut their reasoning processes, assuming that the model's suggestion has already integrated all relevant data (Cross et al., 2024; Klingbeil et al., 2024). Conversely, when AI systems are known to make errors, excessive scepticism may lead to systematic disregard of useful insights—a phenomenon known as *automation aversion (Kostick-Quenet & Gerke, 2022)*. Balancing these extremes requires deliberate epistemic discipline and institutional mechanisms that preserve reflective judgment.

3.4. The Problem of Accountability and Legal Liability

Perhaps the most contentious risk domain concerns responsibility for harm when AI contributes to diagnostic error (Cestonaro et al., 2023; Contaldo et al., 2024). Current legal frameworks assign ultimate accountability to the licensed clinician, regardless of the degree of machine involvement (Jones et al., 2023; Maliha et al., 2021). Yet this arrangement becomes ethically fragile when the AI system's internal logic is inscrutable even to its developers. Physicians may thus be held liable for decisions shaped by models whose reasoning they cannot reconstruct.

This tension exposes a structural mismatch between existing doctrines of medical liability and the distributed nature of AI-mediated decision-making. Hospitals, software vendors, and data providers all contribute to the diagnostic process, yet the chain of accountability remains unarticulated. Without clear delineation of responsibility, both clinicians and patients inhabit a zone of epistemic uncertainty. Moreover, legal ambiguity can foster defensive approach to practicing medicine or reluctance to adopt beneficial innovations, thereby constraining the very progress AI was meant to deliver (Maliha et al., 2021).

The risk landscape of AI-assisted diagnostics, thus encompasses more than technical reliability. It includes cognitive, ethical, and institutional vulnerabilities that arise when probabilistic systems participate in normative acts of medical judgment. The following section addresses these vulnerabilities within the broader ethical and professional frameworks that govern the practice of medicine.

4. Ethical and Professional Implications

The ethical implications of integrating artificial intelligence into diagnostic medicine extend beyond the management of technological risk; they concern the reconfiguration of moral agency and the professional identity of the physician (Ackerhans et al., 2025; Dankwa-Mullan, 2024; Elgin & Elgin, 2024). Medicine is a domain historically defined by fiduciary duty—the moral obligation of the physician to exercise independent judgment for the patient's welfare. As AI systems enter diagnostic reasoning, this duty becomes shared, yet asymmetrically so, between human and machine. The ensuing moral configuration challenges established notions of responsibility, trust, and professional integrity.

4.1. Responsibility and Informed Consent in AI-Mediated Decisions

At the heart of the ethical dilemma lies the distribution of responsibility. When an AI system contributes to a diagnostic conclusion, the physician's decision is no longer purely autonomous. It becomes partially shaped by algorithmic inference, which may be correct, biased, or flawed in ways imperceptible to the clinician. Nevertheless, legal and professional accountability remain anchored to the human practitioner (Jones et al., 2023; Maliha et al., 2021). This discrepancy raises the question of whether moral responsibility presupposes epistemic control—can a clinician be fully responsible for an outcome generated by a process they cannot audit or comprehend? Informed consent further complicates this relationship. Traditional consent assumes that patients are informed about who, or what, is making diagnostic and therapeutic decisions. When AI systems influence these processes, the patient's understanding of agency becomes blurred. Ethically, it could be argued that patients should have a right to know when a diagnosis has been algorithmically assisted, and to what extent the recommendation reflects machine reasoning (H. J. Park, 2024; Ploug et al., 2025; Shaw et al., 2025). Failing to disclose this undermines patient autonomy and erodes trust in the therapeutic relationship. Yet few institutional protocols currently mandate explicit disclosure of AI participation in clinical decision-making (Bignami et al., 2025; Khosravi et al., 2024b).

4.2. Moral Agency, Trust, and the Patient–Doctor–AI Triad

The introduction of AI into diagnostics effectively creates a triadic relationship: patient, physician, and algorithm. Trust, once dyadic, now depends on how the physician mediates between human and non-human sources of authority. The physician must not only interpret the patient's condition but also interpret the algorithm's reasoning—or lack thereof. This dual interpretive burden redefines what it means to be a trustworthy clinician. If physicians rely too heavily on algorithmic outputs, they risk diminishing their role as moral agents who exercise judgment in the face of uncertainty. Conversely, rejecting AI assistance altogether can constitute a different form of negligence—one that disregards tools capable of improving diagnostic accuracy. The ethical ideal lies not in technological abstinence or blind adoption, but in cultivating *epistemic humility*: the recognition that both human and machine reasoning are fallible, and that good clinical practice involves managing, not eliminating, uncertainty.

Trust in AI-mediated diagnostics also depends on transparency and validation. Patients should be given assurance that algorithmic tools have been evaluated transparently and that their limitations are acknowledged. The moral obligation of the clinician extends to communicating these limitations clearly, without overstating the reliability or intelligence of the system. Overpersonalisation of AI—treating it as a quasi-expert or colleague—can foster misplaced trust and distort moral accountability (MacIntyre et al., 2023).

4.3. Ethical Boundaries of Algorithmic Assistance

The ethical boundary between assistance and delegation is particularly delicate. Assistance implies that AI supports the physician's reasoning; delegation implies partial transfer of that reasoning to the machine (Funer & Wiesing, 2024; Pham, 2025). In practice, the distinction is often blurred, especially when AI systems generate direct diagnostic statements rather than probabilistic cues. Once an algorithm offers a seemingly definitive conclusion, the cognitive pressure on the physician to either accept or reject it becomes ethically consequential (Funer & Wiesing, 2024).

Furthermore, algorithmic systems do not share human moral intuitions. They cannot prioritise patient dignity, contextual sensitivity, or compassion—values central to clinical ethics. Their outputs, no matter how sophisticated, remain instrumental rather than empathetic (Shen et al., 2024; Sirgiovanni, 2025). Ethical practice in AI-assisted medicine thus requires preserving human presence as the locus of empathy and judgment (M. K. Park et al., 2025). The physician's role is then not diminished but transformed: from an exclusive decision-maker to a moral interpreter of technological cognition.

Finally, institutions deploying diagnostic AI should acknowledge collective ethical responsibility. Hospitals and health systems bear moral obligations to ensure that deployed models are transparent, validated, and aligned with the ethical principles of beneficence, non-maleficence, autonomy, and justice (Gorelik et al., 2025; Pham, 2025; Weidener & Fischer, 2024). Failing to meet these obligations might converts technological innovation into ethical negligence.

5. Governance and Regulation

The rapid adoption of artificial intelligence in diagnostic medicine has far outpaced the establishment of corresponding governance structures. While technological capabilities expand with remarkable velocity, ethical, legal, and institutional frameworks remain fragmented and reactive (Bouderhem, 2024; Mennella et al., 2024; Nasir et al., 2025). This asymmetry between innovation and regulation poses profound challenges for patient safety, professional accountability, and societal trust in AI-mediated healthcare.

5.1. Current Regulatory Landscape for AI in Healthcare

At present, regulatory mechanisms governing AI in medicine are largely adapted from pre-existing frameworks for medical devices and software (Fraser et al., 2023; Santra et al., 2024). In many jurisdictions, AI diagnostic systems are categorised as *software as a medical device (SaMD)*, subject to evaluation based on safety, efficacy, and data protection standards (Health, 2025a, 2025b; Navarro, 2024). Such classifications, however, inadequately capture the adaptive, non-deterministic nature of machine learning models, particularly those capable of continuous learning or unsupervised pattern recognition (Onitiu et al., 2024; Pantanowitz et al., 2024).

Traditional regulatory paradigms assume that medical devices are static, their performance fixed at the time of approval. AI systems, by contrast, evolve as they process new data, potentially altering their diagnostic behaviour

post-deployment (Babic et al., 2025; Onitiu et al., 2024). This dynamic quality renders conventional approval models obsolete, as periodic re-certification or post-market surveillance may fail to detect subtle but clinically significant drift in algorithmic performance. Furthermore, many existing regulations emphasise product conformity but not epistemic transparency—compliance is assessed by procedural documentation rather than demonstrable interpretability. The situation is further complicated by the proliferation of *general-purpose* large language models adapted for medical use. Unlike domain-specific tools trained on curated clinical datasets, general LLMs often lack formal medical validation yet are widely employed by clinicians for drafting, summarising, or differential diagnosis (Busch et al., 2025; Nazi & Peng, 2024; Singhal et al., 2023). Regulatory agencies currently lack the authority or technical capacity to oversee such hybrid, continuously evolving systems (Health, 2025a; Muralidharan et al., 2024).

5.2. The Limits of Existing Medical Liability Frameworks

Medical liability systems presuppose a clear chain of causation between practitioner action and patient outcome. In AI-assisted diagnostics, this causal chain becomes diffused. A misdiagnosis may result from flawed training data, an opaque inference, or an inappropriate reliance on machine output. Yet existing legal norms hold the physician solely accountable, even when the underlying model is proprietary and inaccessible for forensic examination (Cestonaro et al., 2023; *Fault Lines in Health Care AI – Part Two*, 2025).

This legal asymmetry may disincentivise transparency. Developers, shielded by intellectual property protections, may restrict access to algorithmic details, leaving clinicians to bear the ethical and legal risk of using systems they cannot audit (Cestonaro et al., 2023; Lawton et al., 2024). Hospitals, in turn, adopt AI tools under institutional pressure to modernise, often without fully understanding their operational boundaries. The result is a regulatory vacuum in which responsibility is fragmented among actors such as physicians, developers, data providers, and institutions, without any coherent framework for shared liability. A reformed governance approach must therefore move beyond individual culpability to embrace *distributed accountability*. This would entail legally recognising that diagnostic outcomes in AI-mediated environments emerge from socio-technical systems, not isolated agents. Ethical responsibility should be proportionally allocated among stakeholders according to their degree of epistemic control and capacity to prevent harm.

6. Discussion

The emergence of artificial intelligence in diagnostic medicine signifies not merely a technological innovation but a paradigmatic reordering of epistemic and moral structures within healthcare. The preceding sections have outlined how LLMs and other AI systems introduce new forms of diagnostic reasoning, new loci of error, and new challenges for ethical and legal accountability. This discussion section synthesises these insights to evaluate how medicine must adapt to preserve its normative foundations—truthfulness, responsibility, and professional integrity—while embracing the benefits of computational intelligence.

The central paradox of AI in medicine is that the same properties that make these systems powerful—autonomy, adaptivity, and scalability—also make them epistemically fragile. AI can detect subtle statistical regularities across massive datasets, but it cannot situate those regularities within causal or moral frameworks. The medical profession, however, is founded on causal reasoning and ethical deliberation. Integrating AI thus requires a dual commitment: to innovation as a means of improving care, and to responsibility as a safeguard of meaning and trust.

Physicians should remain the primary custodians of diagnostic interpretation, not because humans outperform machines in statistical inference, but because human reasoning is embedded in moral context. The legitimacy of clinical judgment stems from its accountability to both evidence and ethical norms—something AI cannot replicate. To sustain this legitimacy, clinicians should cultivate a reflective stance toward algorithmic assistance: using AI as a tool for hypothesis generation rather than as an epistemic authority. Institutional frameworks must reinforce this stance through training programs that emphasise critical literacy about AI outputs, cognitive biases, and uncertainty management.

Moreover, AI-assisted diagnostics compel a redefinition of what it means to be an expert. Traditionally, expertise implied mastery of knowledge and the capacity to reason through complex uncertainty. In AI-mediated contexts, expertise increasingly involves the ability to evaluate and contextualise algorithmic reasoning. The competent physician of the near future must understand not only disease mechanisms but also model architectures, data provenance, and the interpretive limits of statistical inference. This epistemic expansion suggests that clinical

expertise will become more interdisciplinary, drawing on informatics, ethics, and systems theory. Yet there is a risk that the medical profession could devolve into a supervisory role—technically overseeing systems that perform the substantive cognitive work of diagnosis. To prevent such displacement, medical education should integrate *algorithmic epistemology*—a structured understanding of how AI systems know, err, and evolve. Only by mastering this knowledge can clinicians preserve their agency as interpreters rather than custodians of machine output.

Finally, epistemic humility is emerging as a vital virtue in AI-mediated medicine. It entails recognising the partiality of both human and algorithmic knowledge. It can be argued that physicians should resist the seduction of technological omniscience, while AI developers should acknowledge the social and moral dimensions of diagnostic reasoning. Transparency, both technical and communicative, becomes the ethical counterpart of humility. For AI systems, transparency requires intelligible documentation of training data, model scope, and known failure modes. For physicians, transparency involves disclosing to patients when AI systems have contributed to diagnostic reasoning and articulating the degree of confidence and uncertainty involved. For institutions, transparency means cultivating environments where algorithmic errors can be reported without punitive consequence, thereby transforming individual blame into collective learning.

The broader societal dimension of transparency concerns trust. Public trust in medicine has historically rested on the assumption of human accountability—someone who can explain, justify, and empathise. As AI increasingly mediates diagnosis, sustaining that trust demands mechanisms through which algorithmic processes become morally legible. This does not mean anthropomorphising AI, but rather embedding it within systems of human explanation and oversight.

7. Conclusion

Artificial intelligence has become an indispensable instrument in the evolution of diagnostic medicine, yet its integration exposes deep tensions between technological capability and epistemic legitimacy. The medical act of diagnosis has never been purely technical; it has always been an interpretive, moral, and social process grounded in human judgment and responsibility. By introducing systems that can simulate diagnostic reasoning without understanding, AI disrupts the foundations upon which that process rests. This review has argued that the diagnostic use of large language models and related AI systems represents both a remarkable advance and a profound epistemological challenge. These models expand the reach of diagnostic inference by uncovering correlations invisible to human cognition, but they also erode the transparency and accountability that lend medicine its moral authority. The physician remains legally and ethically accountable, yet increasingly relies on outputs generated by algorithms whose internal logic cannot be meaningfully inspected or contested. This asymmetry produces a new form of clinical vulnerability: responsibility without full comprehension. The key insight emerging from this analysis is that the successful coexistence of human and artificial intelligence in medicine depends not on technological perfection but on institutional wisdom. AI should remain a tool of reasoning, not a substitute for it. Its diagnostic suggestions should be treated as probabilistic hypotheses, always subject to the clinician's contextual judgment. Preserving this hierarchy requires an ethical infrastructure that prioritises transparency, continuous oversight, and distributive accountability among all actors in the diagnostic chain. At the epistemological level, the rise of AI calls for a renewed humility within medical science—a recognition that neither human expertise nor algorithmic intelligence possesses complete knowledge of disease. The task ahead is to construct systems in which human and machine reasoning complement, rather than compete with, each other. This balance can be achieved only if medicine resists the temptation to conflate predictive accuracy with understanding, and if it safeguards the interpretive and moral dimensions of diagnosis as distinctly human responsibilities.

Author Contributions

Conceptualisation, M.M. and L.L.; methodology, M.M.; validation, L.L., M.M; investigation, L.L., M.M; data curation, L.L. and M.M; writing—original draft preparation, M.M; writing—review and editing, L.L; supervision, L.L. 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. Abdelwanis, M., Alarafati, H. K., Tammam, M. M. S., & Simsekler, M. C. E. (2024). Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis. *Journal of Safety Science and Resilience*, *5*(4), 460–469. https://doi.org/10.1016/j.jnlssr.2024.06.001
2. Ackerhans, S., Wehkamp, K., Petzina, R., Dumitrescu, D., & Schultz, C. (2025). Perceived Trust and Professional Identity Threat in AI-Based Clinical Decision Support Systems: Scenario-Based Experimental Study on AI Process Design Features. *JMIR Formative Research*, *9*(1), e64266. https://doi.org/10.2196/64266
3. Acun, C., & Nasraoui, O. (2025). Pre Hoc and Co Hoc Explainability: Frameworks for Integrating Interpretability into Machine Learning Training for Enhanced Transparency and Performance. *Applied Sciences*, *15*(13), 7544. https://doi.org/10.3390/app15137544
4. Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications*, *11*(1), 1568. https://doi.org/10.1057/s41599-024-04044-8
5. Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., & Gao, G. (2023). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy and Technology*, *12*(1), 100702. https://doi.org/10.1016/j.hlpt.2022.100702
6. Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education*, *23*(1), 689. https://doi.org/10.1186/s12909-023-04698-z
7. Angus, D. C., Khera, R., Lieu, T., Liu, V., Ahmad, F. S., Anderson, B., Bhavani, S. V., Bindman, A., Brennan, T., Celi, L. A., Chen, F., Cohen, I. G., Denniston, A., Desai, S., Embí, P., Faisal, A., Ferryman, K., Gerhart, J., Gross, M., … JAMA Summit on AI. (2025). AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence. *JAMA*. https://doi.org/10.1001/jama.2025.18490
8. Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J. A., & Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *Npj Digital Medicine*, *8*(1), 274. https://doi.org/10.1038/s41746-025-01670-7
9. Babic, B., Glenn Cohen, I., Stern, A. D., Li, Y., & Ouellet, M. (2025). A general framework for governing marketed AI/ML medical devices. *Npj Digital Medicine*, *8*(1), 328. https://doi.org/10.1038/s41746-025-01717-9
10. Bignami, E., Darhour, L. J., Franco, G., Guarnieri, M., & Bellini, V. (2025). AI policy in healthcare: A checklist-based methodology for structured implementation. *Journal of Anesthesia, Analgesia and Critical Care*, *5*, 56. https://doi.org/10.1186/s44158-025-00278-3
11. Boge, F., & Mosig, A. (2025). Causality and scientific explanation of artificial intelligence systems in biomedicine. *Pflügers Archiv - European Journal of Physiology*, *477*(4), 543–554. https://doi.org/10.1007/s00424-024-03033-9
12. Bouderhem, R. (2024). Shaping the future of AI in healthcare through ethics and governance. *Humanities and Social Sciences Communications*, *11*(1), 416. https://doi.org/10.1057/s41599-024-02894-w
13. Bozyel, S., Şimşek, E., Koçyiğit Burunkaya, D., Güler, A., Korkmaz, Y., Şeker, M., Ertürk, M., & Keser, N. (2024). Artificial Intelligence-Based Clinical Decision Support Systems in Cardiovascular Diseases. *The Anatolian Journal of Cardiology*, *28*(2), 74. https://doi.org/10.14744/AnatolJCardiol.2023.3685
14. Busch, F., Hoffmann, L., Rueger, C., van Dijk, E. H., Kader, R., Ortiz-Prado, E., Makowski, M. R., Saba, L., Hadamitzky, M., Kather, J. N., Truhn, D., Cuocolo, R., Adams, L. C., & Bressem, K. K. (2025). Current applications and challenges in large language models for patient care: A systematic review. *Communications Medicine*, *5*(1), 26. https://doi.org/10.1038/s43856-024-00717-2
15. Celi, L. A., Cellini, J., Charpignon, M.-L., Dee, E. C., Dernoncourt, F., Eber, R., Mitchell, W. G., Moukheiber, L., Schirmer, J., Situ, J., Paguio, J., Park, J., Wawira, J. G., Yao, S., & Data, for M. C.

(2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health*, *1*(3), e0000022. https://doi.org/10.1371/journal.pdig.0000022

16. Cestonaro, C., Delicati, A., Marcante, B., Caenazzo, L., & Tozzo, P. (2023). Defining medical liability when artificial intelligence is applied on diagnostic algorithms: A systematic review. *Frontiers in Medicine*, *10*, 1305756. https://doi.org/10.3389/fmed.2023.1305756
17. Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research*, *26*(1), e53164. https://doi.org/10.2196/53164
18. Chen, H., Gomez, C., Huang, C.-M., & Unberath, M. (2022). Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. *Npj Digital Medicine*, *5*(1), 156. https://doi.org/10.1038/s41746-022-00699-2
19. Chustecki, M. (2024). Benefits and Risks of AI in Health Care: Narrative Review. *Interactive Journal of Medical Research*, *13*, e53616. https://doi.org/10.2196/53616
20. Colacci, M., Huang, Y. Q., Postill, G., Zhelnov, P., Fennelly, O., Verma, A., Straus, S., & Tricco, A. C. (2025). Sociodemographic bias in clinical machine learning models: A scoping review of algorithmic bias instances and mechanisms. *Journal of Clinical Epidemiology*, *178*, 111606. https://doi.org/10.1016/j.jclinepi.2024.111606
21. Contaldo, M. T., Pasceri, G., Vignati, G., Bracchi, L., Triggiani, S., & Carrafiello, G. (2024). AI in Radiology: Navigating Medical Responsibility. *Diagnostics*, *14*(14), 1506. https://doi.org/10.3390/diagnostics14141506
22. Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, *3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651
23. D'Adderio, L., & Bates, D. W. (2025). Transforming diagnosis through artificial intelligence. *NPJ Digital Medicine*, *8*(1), 54. https://doi.org/10.1038/s41746-025-01460-1
24. Dankwa-Mullan, I. (2024). Health Equity and Ethical Considerations in Using Artificial Intelligence in Public Health and Medicine. *Preventing Chronic Disease*, *21*. https://doi.org/10.5888/pcd21.240245
25. DeGroat, W., Abdelhalim, H., Peker, E., Sheth, N., Narayanan, R., Zeeshan, S., Liang, B. T., & Ahmed, Z. (2024). Multimodal AI/ML for discovering novel biomarkers and predicting disease using multi-omics profiles of patients with cardiovascular diseases. *Scientific Reports*, *14*(1), 26503. https://doi.org/10.1038/s41598-024-78553-6
26. Elgin, C. Y., & Elgin, C. (2024). Ethical implications of AI-driven clinical decision support systems on healthcare resource allocation: A qualitative study of healthcare professionals' perspectives. *BMC Medical Ethics*, *25*, 148. https://doi.org/10.1186/s12910-024-01151-8
27. Fahim, Y. A., Hasani, I. W., Kabba, S., & Ragab, W. M. (2025). Artificial intelligence in healthcare and medicine: Clinical applications, therapeutic advances, and future perspectives. *European Journal of Medical Research*, *30*(1), 848. https://doi.org/10.1186/s40001-025-03196-w
28. *Fault lines in health care AI – Part two: Who's responsible when AI gets it wrong?* (2025, June 26). https://carey.jhu.edu/articles/fault-lines-health-care-ai-part-two-whos-responsible-when-ai-gets-it-wrong
29. Fraser, A. G., Biasin, E., Bijnens, B., Bruining, N., Caiani, E. G., Cobbaert, K., Davies, R. H., Gilbert, S. H., Hovestadt, L., Kamenjasevic, E., Kwade, Z., McGauran, G., O'Connor, G., Vasey, B., & Rademakers, F. E. (2023). Artificial intelligence in medical device software and high-risk medical devices – a review of definitions, expert recommendations and regulatory initiatives. *Expert Review of Medical Devices*, *20*(6), 467–491. https://doi.org/10.1080/17434440.2023.2184685
30. Freyer, N., Groß, D., & Lipprandt, M. (2024). The ethical requirement of explainability for AI-DSS in healthcare: A systematic review of reasons. *BMC Medical Ethics*, *25*(1), 104. https://doi.org/10.1186/s12910-024-01103-2
31. Funer, F., & Wiesing, U. (2024). Physician's autonomy in the face of AI support: Walking the ethical tightrope. *Frontiers in Medicine*, *11*. https://doi.org/10.3389/fmed.2024.1324963
32. Goh, E., Bunning, B., Khoong, E. C., Gallo, R. J., Milstein, A., Centola, D., & Chen, J. H. (2025). Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. *Communications Medicine*, *5*(1), 59. https://doi.org/10.1038/s43856-025-00781-2
33. Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H.

(2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, *7*(10), e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

34. Gorelik, A. J., Li, M., Hahne, J., Wang, J., Ren, Y., Yang, L., Zhang, X., Liu, X., Wang, X., Bogdan, R., & Carpenter, B. D. (2025). Ethics of AI in healthcare: A scoping review demonstrating applicability of a foundational framework. *Frontiers in Digital Health*, *7*. https://doi.org/10.3389/fdgth.2025.1662642
35. Graili, P., & Farhoudi, B. (2025). The intersection of digital health and artificial intelligence: Clearing the cloud of uncertainty. *DIGITAL HEALTH*, *11*, 20552076251315621. https://doi.org/10.1177/20552076251315621
36. Greenhalgh, T., Thorne, S., & Malterud, K. (2018). Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation*, *48*(6), e12931. https://doi.org/10.1111/eci.12931
37. Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., & Rashidi, H. H. (2025). Ethical and Bias Considerations in Artificial Intelligence/Machine Learning. *Modern Pathology*, *38*(3). https://doi.org/10.1016/j.modpat.2024.100686
38. Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., & White, J. A. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *Npj Digital Medicine*, *8*(1), 154. https://doi.org/10.1038/s41746-025-01503-7
39. Health, C. for D. and R. (2025a). Artificial Intelligence in Software as a Medical Device. *FDA*. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device
40. Health, C. for D. and R. (2025b, July 23). *Software as a Medical Device (SaMD)*. FDA; FDA. https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd
41. Jeong, H. K., Park, C., Henao, R., & Kheterpal, M. (2022). Deep Learning in Dermatology: A Systematic Review of Current Approaches, Outcomes, and Limitations. *JID Innovations*, *3*(1), 100150. https://doi.org/10.1016/j.xjidi.2022.100150
42. Jones, C., Thornton, J., & Wyatt, J. C. (2023). Artificial intelligence and clinical decision support: Clinicians' perspectives on trust, trustworthiness, and liability. *Medical Law Review*, *31*(4), 501–520. https://doi.org/10.1093/medlaw/fwad013
43. Joseph, J. (2025). Algorithmic bias in public health AI: A silent threat to equity in low-resource settings. *Frontiers in Public Health*, *13*. https://doi.org/10.3389/fpubh.2025.1643180
44. Jui, T. D., & Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, *15*(8), 3095–3125. https://doi.org/10.1007/s13042-023-02083-2
45. Khosravi, M., Zare, Z., Mojtabaeian, S. M., & Izadi, R. (2024a). Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews. *Health Services Research and Managerial Epidemiology*, *11*, 23333928241234863. https://doi.org/10.1177/23333928241234863
46. Khosravi, M., Zare, Z., Mojtabaeian, S. M., & Izadi, R. (2024b). Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews. *Health Services Research and Managerial Epidemiology*, *11*, 23333928241234863. https://doi.org/10.1177/23333928241234863
47. Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352. https://doi.org/10.1016/j.chb.2024.108352
48. Koçak, B., Ponsiglione, A., Stanzione, A., Bluethgen, C., Santinha, J., Ugga, L., Huisman, M., Klontzas, M. E., Cannella, R., & Cuocolo, R. (2025). Bias in artificial intelligence for medical imaging: Fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and Interventional Radiology*, *31*(2), 75–88. https://doi.org/10.4274/dir.2024.242854
49. Korfiatis, P., Kline, T. L., Meyer, H. M., Khalid, S., Leiner, T., Loufek, B. T., Blezek, D., Vidal, D. E., Hartman, R. P., Joppa, L. J., Missert, A. D., Potretzke, T. A., Taubel, J. P., Tjelta, J. A., Callstrom, M. R., & Williamson, E. E. (2025). Implementing Artificial Intelligence Algorithms in the Radiology Workflow: Challenges and Considerations. *Mayo Clinic Proceedings: Digital Health*, *3*(1). https://doi.org/10.1016/j.mcpdig.2024.100188

50. Kostick-Quenet, K. M., & Gerke, S. (2022). AI in the hands of imperfect users. *Npj Digital Medicine*, *5*(1), 197. https://doi.org/10.1038/s41746-022-00737-z
51. Krakowski, I., Kim, J., Cai, Z. R., Daneshjou, R., Lapins, J., Eriksson, H., Lykou, A., & Linos, E. (2024). Human-AI interaction in skin cancer diagnosis: A systematic review and meta-analysis. *Npj Digital Medicine*, *7*(1), 78. https://doi.org/10.1038/s41746-024-01031-w
52. Krishnan, G., Singh, S., Pathania, M., Gosavi, S., Abhishek, S., Parchani, A., & Dhar, M. (2023). Artificial intelligence in clinical medicine: Catalyzing a sustainable global healthcare paradigm. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1227091
53. Kulikowski, C. A. (2015). An Opening Chapter of the First Generation of Artificial Intelligence in Medicine: The First Rutgers AIM Workshop, June 1975. *Yearbook of Medical Informatics*, *10*(1), 227–233. https://doi.org/10.15265/IY-2015-016
54. Lawton, T., Morgan, P., Porter, Z., Hickey, S., Cunningham, A., Hughes, N., Iacovides, I., Jia, Y., Sharma, V., & Habli, I. (2024). Clinicians risk becoming 'liability sinks' for artificial intelligence. *Future Healthcare Journal*, *11*(1), 100007. https://doi.org/10.1016/j.fhj.2024.100007
55. MacIntyre, M. R., Cockerill, R. G., Mirza, O. F., & Appel, J. M. (2023). Ethical considerations for the use of artificial intelligence in medical decision-making capacity assessments. *Psychiatry Research*, *328*, 115466. https://doi.org/10.1016/j.psychres.2023.115466
56. Maliha, G., Gerke, S., Cohen, I. G., & Parikh, R. B. (2021). Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation. *The Milbank Quarterly*, *99*(3), 629–647. https://doi.org/10.1111/1468-0009.12504
57. McGenity, C., Clarke, E. L., Jennings, C., Matthews, G., Cartlidge, C., Freduah-Agyemang, H., Stocken, D. D., & Treanor, D. (2024). Artificial intelligence in digital pathology: A systematic review and meta-analysis of diagnostic test accuracy. *Npj Digital Medicine*, *7*(1), 114. https://doi.org/10.1038/s41746-024-01106-8
58. Mennella, C., Maniscalco, U., De Pietro, G., & Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*, *10*(4), e26297. https://doi.org/10.1016/j.heliyon.2024.e26297
59. Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). Bias in AI-based models for medical applications: Challenges and mitigation strategies. *Npj Digital Medicine*, *6*(1), 113. https://doi.org/10.1038/s41746-023-00858-z
60. Muralidharan, V., Adewale, B. A., Huang, C. J., Nta, M. T., Ademiju, P. O., Pathmarajah, P., Hang, M. K., Adesanya, O., Abdullateef, R. O., Babatunde, A. O., Ajibade, A., Onyeka, S., Cai, Z. R., Daneshjou, R., & Olatunji, T. (2024). A scoping review of reporting gaps in FDA-approved AI medical devices. *Npj Digital Medicine*, *7*(1), 273. https://doi.org/10.1038/s41746-024-01270-x
61. Nasir, M., Siddiqui, K., & Ahmed, S. (2025). Ethical-legal implications of AI-powered healthcare in critical perspective. *Frontiers in Artificial Intelligence*, *8*, 1619463. https://doi.org/10.3389/frai.2025.1619463
62. Navarro, A. (2024, September 25). EU MDR and IVDR: Classifying Medical Device Software (MDSW). *NAMSA*. https://namsa.com/resources/blog/eu-mdr-and-ivdr-classifying-medical-device-software-mdsw/
63. Nazi, Z. A., & Peng, W. (2024). Large Language Models in Healthcare and Medical Domain: A Review. *Informatics*, *11*(3), 57. https://doi.org/10.3390/informatics11030057
64. Nicholson, N., Giusti, F., & Martos, C. (2023). Ontology-Based AI Design Patterns and Constraints in Cancer Registry Data Validation. *Cancers*, *15*(24), 5812. https://doi.org/10.3390/cancers15245812
65. Onitiu, D., Wachter, S., & Mittelstadt, B. (2024). How AI challenges the medical device regulation: Patient safety, benefits, and intended uses. *Journal of Law and the Biosciences*, lsae007. https://doi.org/10.1093/jlb/lsae007
66. Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., & Nascimento, E. G. S. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing*, *7*(1), 15. https://doi.org/10.3390/bdcc7010015
67. Pantanowitz, L., Hanna, M., Pantanowitz, J., Lennerz, J., Henricks, W. H., Shen, P., Quinn, B., Bennet, S., & Rashidi, H. H. (2024). Regulatory Aspects of Artificial Intelligence and Machine Learning. *Modern Pathology*, *37*(12). https://doi.org/10.1016/j.modpat.2024.100609

68. Park, H. J. (2024). Patient perspectives on informed consent for medical AI: A web-based experiment. *DIGITAL HEALTH*, *10*, 20552076241247938. https://doi.org/10.1177/20552076241247938
69. Park, M. K., Ashwood, N., Capes, N., Park, M. K., Ashwood, N., & Capes, N. (2025). Ethics of Artificial Intelligence in Medicine. *Cureus*, *17*(5). https://doi.org/10.7759/cureus.83567
70. Patil, S. V., Myers, C. G., & Lu-Myers, Y. (2025). Calibrating AI Reliance—A Physician's Superhuman Dilemma. *JAMA Health Forum*, *6*(3), e250106. https://doi.org/10.1001/jamahealthforum.2025.0106
71. Pham, T. (2025). Ethical and legal considerations in healthcare AI: Innovation and policy for safe and fair use. *Royal Society Open Science*, *12*(5), 241873. https://doi.org/10.1098/rsos.241873
72. Ploug, T., Jørgensen, R. F., Motzfeldt, H. M., Ploug, N., & Holm, S. (2025). The need for patient rights in AI-driven healthcare – risk-based regulation is not enough. *Journal of the Royal Society of Medicine*, 01410768251344707. https://doi.org/10.1177/01410768251344707
73. Rao, A., & Aalami, O. (2023). Towards improving the visual explainability of artificial intelligence in the clinical setting. *BMC Digital Health*, *1*(1), 23. https://doi.org/10.1186/s44247-023-00022-3
74. Rezaeian, O., Bayrak, A. E., & Asan, O. (2025). *Explainability and AI Confidence in Clinical Decision Support Systems: Effects on Trust, Diagnostic Performance, and Cognitive Load in Breast Cancer Care* (No. arXiv:2501.16693). arXiv. https://doi.org/10.48550/arXiv.2501.16693
75. Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. *JMIR AI*, *3*(1), e53207. https://doi.org/10.2196/53207
76. Saadeh, M. I., Janhonen, J., Beer, E., Castelyn, C., & Hoffman, D. N. (2025). Automation complacency: Risks of abdicating medical decision making. *AI and Ethics*. https://doi.org/10.1007/s43681-025-00825-2
77. Saarela, M., & Podgorelec, V. (2024). Recent Applications of Explainable AI (XAI): A Systematic Literature Review. *Applied Sciences*, *14*(19), 8884. https://doi.org/10.3390/app14198884
78. Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S., & Pardalos, P. M. (2024). A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, *118*, 109370. https://doi.org/10.1016/j.compeleceng.2024.109370
79. Sadr, H., Nazari, M., Khodaverdian, Z., Farzan, R., Yousefzadeh-Chabok, S., Ashoobi, M. T., Hemmati, H., Hendi, A., Ashraf, A., Pedram, M. M., Hasannejad-Bibalan, M., & Yamaghani, M. R. (2025). Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: A comprehensive review of machine learning and deep learning approaches. *European Journal of Medical Research*, *30*(1), 418. https://doi.org/10.1186/s40001-025-02680-7
80. Saenz, A. D., Harned, Z., Banerjee, O., Abràmoff, M. D., & Rajpurkar, P. (2023). Autonomous AI systems in the face of liability, regulations and costs. *NPJ Digital Medicine*, *6*(1), 185. https://doi.org/10.1038/s41746-023-00929-1
81. Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q., & Tsaftaris, S. A. (2022). Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, *9*(8), 220638. https://doi.org/10.1098/rsos.220638
82. Santra, S., Kukreja, P., Saxena, K., Gandhi, S., & Singh, O. V. (2024). Navigating regulatory and policy challenges for AI enabled combination devices. *Frontiers in Medical Technology*, *6*, 1473350. https://doi.org/10.3389/fmedt.2024.1473350
83. Savage, T., Nayak, A., Gallo, R., Rangan, E., & Chen, J. H. (2024). Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *Npj Digital Medicine*, *7*(1), 20. https://doi.org/10.1038/s41746-024-01010-1
84. Shaw, D., Lorenzini, G., Ossa, L. A., Eckstein, J., Steiner, L., & Elger, B. S. (2025). When and what patients need to know about AI in clinical care. *Swiss Medical Weekly*, *155*(1), 4013–4013. https://doi.org/10.57187/s.4013
85. Shen, J., DiPaola, D., Ali, S., Sap, M., Park, H. W., & Breazeal, C. (2024). Empathy Toward Artificial Intelligence Versus Human Experiences and the Role of Transparency in Mental Health and Social Support Chatbot Design: Comparative Study. *JMIR Mental Health*, *11*, e62679. https://doi.org/10.2196/62679

86. Shin, H. S. (2019). Reasoning processes in clinical reasoning: From the perspective of cognitive psychology. *Korean Journal of Medical Education*, *31*(4), 299–308. https://doi.org/10.3946/kjme.2019.140
87. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2
88. Sinha, R. (2024). The role and impact of new technologies on healthcare systems. *Discover Health Systems*, *3*(1), 96. https://doi.org/10.1007/s44250-024-00163-w
89. Sirgiovanni, E. (2025). Should Doctor Robot possess moral empathy? *Bioethics*, *39*(1), 98–107. https://doi.org/10.1111/bioe.13345
90. Solutions, M. (2024, June 6). Evolution of AI in Healthcare and Clinical Research. *Medidata Solutions*. https://www.medidata.com/en/life-science-resources/medidata-blog/evolution-of-ai-in-healthcare-and-clinical-research/
91. Sourlos, N., Vliegenthart, R., Santinha, J., Klontzas, M. E., Cuocolo, R., Huisman, M., & van Ooijen, P. (2024). Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology. *Insights into Imaging*, *15*(1), 248. https://doi.org/10.1186/s13244-024-01833-2
92. Stinson, C. (2022). Algorithms are not neutral. *AI and Ethics*, *2*(4), 763–770. https://doi.org/10.1007/s43681-022-00136-w
93. Sukhera, J. (2022). Narrative Reviews: Flexible, Rigorous, and Practical. *Journal of Graduate Medical Education*, *14*(4), 414–417. https://doi.org/10.4300/JGME-D-22-00480.1
94. Tejani, A. S., Ng, Y. S., Xi, Y., & Rayan, J. C. (2024). Understanding and Mitigating Bias in Imaging Artificial Intelligence. *RadioGraphics*, *44*(5), e230067. https://doi.org/10.1148/rg.230067
95. Tikhomirov, L., Semmler, C., McCradden, M., Searston, R., Ghassemi, M., & Oakden-Rayner, L. (2024). Medical artificial intelligence for clinicians: The lost cognitive perspective. *The Lancet Digital Health*, *6*(8), e589–e594. https://doi.org/10.1016/S2589-7500(24)00095-5
96. Tun, H. M., Rahman, H. A., Naing, L., & Malik, O. A. (2025). Trust in Artificial Intelligence–Based Clinical Decision Support Systems Among Health Care Workers: Systematic Review. *Journal of Medical Internet Research*, *27*, e69678. https://doi.org/10.2196/69678
97. Ullah, E., Parwani, A., Baig, M. M., & Singh, R. (2024). Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagnostic Pathology*, *19*(1), 43. https://doi.org/10.1186/s13000-024-01464-7
98. van Diest, P. J., Flach, R. N., van Dooijeweert, C., Makineli, S., Breimer, G. E., Stathonikos, N., Pham, P., Nguyen, T. Q., & Veta, M. (2024). Pros and cons of artificial intelligence implementation in diagnostic pathology. *Histopathology*, *84*(6), 924–934. https://doi.org/10.1111/his.15153
99. Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M., & Božić, J. (2025). A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. *Healthcare*, *13*(6), 603. https://doi.org/10.3390/healthcare13060603
100. Vrudhula, A., Kwan, A. C., Ouyang, D., & Cheng, S. (2024). Machine Learning and Bias in Medical Imaging: Opportunities and Challenges. *Circulation: Cardiovascular Imaging*, *17*(2), e015495. https://doi.org/10.1161/CIRCIMAGING.123.015495
101. Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields: Applications, advances, and challenges. *Artificial Intelligence Review*, *57*(11), 299. https://doi.org/10.1007/s10462-024-10921-0
102. Waqas, A., Bui, M. M., Glassy, E. F., Naqa, I. E., Borkowski, P., Borkowski, A. A., & Rasool, G. (2023). Revolutionizing Digital Pathology With the Power of Generative Artificial Intelligence and Foundation Models. *Laboratory Investigation*, *103*(11). https://doi.org/10.1016/j.labinv.2023.100255
103. Weidener, L., & Fischer, M. (2024). Role of Ethics in Developing AI-Based Applications in Medicine: Insights From Expert Interviews and Discussion of Implications. *JMIR AI*, *3*(1), e51204. https://doi.org/10.2196/51204

104. Wellnhofer, E. (2022). Real-World and Regulatory Perspectives of Artificial Intelligence in Cardiovascular Imaging. *Frontiers in Cardiovascular Medicine*, *9*. https://doi.org/10.3389/fcvm.2022.890809
105. Wu, K., Wu, E., Theodorou, B., Liang, W., Mack, C., Glass, L., Sun, J., & Zou, J. (2024). Characterizing the Clinical Adoption of Medical AI Devices through U.S. Insurance Claims. *NEJM AI*, *1*(1), AIoa2300030. https://doi.org/10.1056/AIoa2300030
106. Xu, H., & Shuttleworth, K. M. J. (2024). Medical artificial intelligence and the black box problem: A view based on the ethical principle of "do no harm." *Intelligent Medicine*, *4*(1), 52–57. https://doi.org/10.1016/j.imed.2023.08.001
107. Xue, C., Kowshik, S. S., Lteif, D., Puducheri, S., Jasodanand, V. H., Zhou, O. T., Walia, A. S., Guney, O. B., Zhang, J. D., Poésy, S., Kaliaev, A., Andreu-Arasa, V. C., Dwyer, B. C., Farris, C. W., Hao, H., Kedar, S., Mian, A. Z., Murman, D. L., O'Shea, S. A., … Kolachalama, V. B. (2024). AI-based differential diagnosis of dementia etiologies on multimodal data. *Nature Medicine*, *30*(10), 2977–2989. https://doi.org/10.1038/s41591-024-03118-z
108. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D., & Ghassemi, M. (2024). The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, *30*(10), 2838–2848. https://doi.org/10.1038/s41591-024-03113-4
109. Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, *16*(3), e12962. https://doi.org/10.1111/soc4.12962
110. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.*, *15*(2), 20:1-20:38. https://doi.org/10.1145/3639372
111. Zhou, S., Xu, Z., Zhang, M., Xu, C., Guo, Y., Zhan, Z., Fang, Y., Ding, S., Wang, J., Xu, K., Xia, L., Yeung, J., Zha, D., Cai, D., Melton, G. B., Lin, M., & Zhang, R. (2025). Large language models for disease diagnosis: A scoping review. *Npj Artificial Intelligence*, *1*(1), 9. https://doi.org/10.1038/s44387-025-00011-z

Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:005.32
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600092J

Nevidljive pretnje: Gledajući unazad da bismo sa AI krenuli napred – višedimenzionalni uticaji AI na organizacionu bezbednost i ljudsku autonomiju -

Tatjana Jovanović[1],

[1] Belgrade School of Engineering Management, Beopolis University, Belgrade, Serbia,
tatjana.jovanovic@fim.rs

Sažetak rada: Kako sistemi veštačke inteligencije sve više prodiru u najrazličitije sfere ljudske aktivnosti — od zdravstvene zaštite i obrazovanja do zapošljavanja i upravljanja — pojavljuju se novi oblici bezbednosnih rizika, od kojih su mnogi i dalje nedovoljno prepoznati. Ovaj rad istražuje višedimenzionalnu infiltraciju veštačke inteligencije u društveno-organizacione sisteme, analizirajući ne samo tehnološke rizike, već i psihološke, organizacione i epistemološke pretnje koje proizlaze iz nekontrolisane automatizacije i netransparentnog algoritamskog odlučivanja.

Polazeći od praksi upravljanja ljudskim resursima (HRM) u digitalno transformisanim organizacijama, rad identifikuje rane signale zloupotrebe veštačke inteligencije: smanjen uticaj ljudskog prosuđivanja, eroziju profesionalne autonomije, preterano oslanjanje na prediktivne analitike i gubitak poverenja među stejkholderima. Proširujući ove nalaze, studija ocrtava moguće scenarije razvoja — od odgovorne integracije veštačke inteligencije zasnovane na dijalogu sa zainteresovanim stranama, do distopijskih putanja obeleženih dehumanizacijom i društvenom fragmentacijom.

Završni deo rada nudi strateške preporuke za podsticanje odgovorne upotrebe veštačke inteligencije, naglašavajući značaj transdisciplinarnog obrazovanja, kritičke digitalne pismenosti i anticipativnog upravljanja. U oblikovanju narednih generacija stručnjaka, uloga akademskih i istraživačkih institucija postaje ključna za ugradnju etičke refleksivnosti, sistemskog mišljenja i čovekocentričnih vrednosti u samo projektovanje budućnosti koju pokreće veštačka inteligencija.

Ključne reči: veštačka inteligencija, ljudska bezbednost, odgovorna veštačka inteligencija

Invisible Threats: Looking Back to Move Forward with AI – The Multidimensional Impact of AI on Organizational Security and Human Agency –

Abstract in English: As artificial intelligence systems increasingly permeate all domains of human activity — from healthcare and education to employment and governance — new forms of security risks emerge, many of which remain insufficiently recognized. This paper explores the multidimensional infiltration of AI into socio-organizational systems, analyzing not only the technological risks, but also the psychological, organizational, and epistemological threats posed by unchecked automation and opaque algorithmic decision-making.

Drawing from human resource management (HRM) practices in digitally transformed organizations, the paper identifies early warning signals of AI misuse: diminished human agency, erosion of professional autonomy, over-reliance on predictive analytics, and loss of trust among stakeholders. By extending these findings, the study outlines future development scenarios, ranging from responsible integration of AI grounded in stakeholder dialogue to dystopian trajectories marked by dehumanization and social fragmentation.

The final section presents strategic recommendations for promoting responsible AI use, emphasizing the need for transdisciplinary education, critical digital literacy, and anticipatory governance. In shaping the next generation of experts, the role of academia and research institutions becomes vital in embedding ethical reflexivity, systemic thinking, and human-centric values into the very design of AI-driven futures.

Keywords: artificial intelligence, human security, responsible AI

1. Introduction

In recent years, artificial intelligence (AI) has moved from the periphery to the core of socio-organizational systems, subtly yet profoundly transforming how decisions are made, work is organized, and human value is defined. From personalized recruitment algorithms to automated performance evaluations, AI technologies are increasingly embedded in everyday organizational practices—often without transparent oversight or critical reflection. While discussions around AI have primarily focused on efficiency, accuracy, and innovation, growing evidence suggests that deeper, less visible risks are emerging.

These risks extend beyond traditional cybersecurity concerns. They touch on psychological autonomy, professional identity, knowledge legitimacy, and the erosion of trust within complex stakeholder ecosystems. In the pursuit of optimization, organizations may unintentionally displace human judgment, reproduce bias at scale, and weaken the social fabric that sustains cooperation and meaning in work environments.

This paper argues that the invisible threats of AI deserve a central place in the discourse on digital transformation and organizational security. It adopts a multidimensional lens to examine not only technological vulnerabilities but also the epistemological, ethical, and strategic blind spots that arise when AI systems are deployed without adequate stakeholder engagement or anticipatory governance.

Drawing on insights from human resource management (HRM) and critical digital studies, this paper explores:

- early indicators of organizational risk linked to AI,
- future development scenarios—from responsible integration to dystopian drift,
- and strategic recommendations for fostering a human-centered approach to AI deployment.

By looking back at the human foundations of organizational life, we aim to move forward with more reflexivity, caution, and imagination in shaping the AI-enhanced future.

2. Theoretical Framework: Security Beyond Technology

Understanding the invisible risks associated with artificial intelligence requires a departure from purely technological paradigms and a shift toward multidisciplinary frameworks that integrate human, organizational, and ethical dimensions. This section is guided by three interconnected theoretical lenses: the sociotechnical systems theory, which emphasizes the entanglement of technology with social and organizational structures; epistemologies of algorithmic governance, which explore how opaque AI systems reshape authority and knowledge; and a human-centered approach to HRM in the era of digital transformation, which highlights the strategic alignment of AI tools with human development and inclusion.

2.1. Sociotechnical Systems Perspective

Originally formulated in mid-20th century organizational studies, the sociotechnical systems perspective remains highly relevant for analyzing the complex entanglements of technology and human labor. Rather than viewing AI as a neutral tool, this framework conceptualizes it as a dynamic actor embedded within broader social systems. Recent work by Kudina and van de Poel (2024) argues that AI should be understood as part of evolving sociotechnical assemblages that shape—rather than merely support—human decision-making, ethics, and responsibility in real time. In parallel, Salwei et al. (2022) propose a practical framework for applying AI in healthcare through a sociotechnical lens, highlighting how misalignment between technological capabilities and human workflows can generate systemic vulnerabilities.

In the organizational context, this approach emphasizes that the risks posed by AI are not simply a matter of malfunction or misuse, but of structural integration—how technology reconfigures authority, accountability, and labor itself. The interaction between human agency and algorithmic output becomes central to identifying both opportunities and emerging invisible threats.

2.2. Algorithmic Epistemology and Opacity

Critical algorithm studies have revealed that AI systems are not only operational tools but epistemic agents that produce new regimes of knowledge and control. Scholars such as Pasquale (2015) and Ananny and Crawford (2018) demonstrate how algorithmic opacity can obscure accountability, delegitimize traditional expertise, and enable silent shifts in institutional power.

As AI models are trained on massive, often proprietary datasets and refined in black-box architectures, the epistemic authority gradually transfers from human professionals to systems whose logic remains largely inscrutable. This is particularly alarming in sectors such as hiring, education, or finance, where algorithmic decisions bear direct consequences for individuals—without transparent rationale or possibility of appeal. The rise of what Veale and Binns (2017) describe as "fairer machine learning without sensitive data" paradoxically risks reinforcing discrimination through proxy variables and biased optimization targets.

2.3. Human-Centered HRM in the Digital Era

In parallel with the sociotechnical and epistemological shifts, HRM has also undergone significant transformation. The dominant consensus now favors human-centered digitalization, wherein AI tools are deployed not to replace, but to enhance human capacity—provided that their use aligns with ethical principles and organizational inclusivity.

Recent literature, including Armstrong (2024) and Guest (2017), argues for frameworks that prioritize psychological well-being, job satisfaction, and autonomy alongside productivity metrics. The holistic HRM approach, developed in response to Industry 4.0, calls for integrated strategies that blend advanced analytics, personalization, and ethical safeguards. This includes participatory AI implementation, transparent communication with employees, and real-time HR analytics to detect unintended effects such as surveillance fatigue or deskilling.

Ultimately, understanding AI-related risks requires a synthesis of these three lenses: only by combining technical insight, organizational design, and human values can we build resilient systems that resist invisible threats and promote sustainable innovation.

3. Emerging Invisible Threats in AI-Driven Organizations

As artificial intelligence systems become increasingly embedded within organizations, their influence often transcends operational efficiency and begins to reshape fundamental human functions, roles, and relationships. These transformations are not always visible, but they can produce serious long-term consequences for autonomy, judgment, and trust in professional environments.

3.1. Diminished Human Agency

One of the earliest invisible threats lies in the gradual replacement of human judgment with algorithmic decision-making. From HR screening processes to loan approvals, AI models often determine outcomes based on statistical inference, not contextual understanding. While this shift promises efficiency, it reduces individual agency, as decisions are made based on opaque criteria. Ananny and Crawford (2018) argue that even with algorithmic transparency, knowing "how" a decision is made does not guarantee understanding "why" it was made, especially in systems trained on massive datasets with embedded biases. The risk is that human actors may become mere executors of automated recommendations, rather than critical evaluators of them.

3.2. Erosion of Professional Autonomy

In high-stakes domains such as healthcare, education, and human resource management, AI tools increasingly influence or even override professional discretion. This creates a paradox: while these tools are designed to assist decision-making, they can also constrain professionals to follow recommendations produced by opaque and data-driven systems. In the field of medicine, Jotterand and Bosco (2021) describe this tension as a "sword of Damocles" hanging over clinicians — where reliance on AI might reduce diagnostic error, but at the cost of undermining physicians' epistemic authority and ethical accountability.

Similar dynamics are unfolding in the field of education, where teachers are encouraged to rely on AI-driven learning analytics and adaptive instruction platforms. These tools, although useful, often function as "closed boxes," limiting pedagogical freedom and sidelining educator expertise. In the realm of human resources, predictive analytics are frequently used to assess employee engagement, performance, or flight risk. However, these models may neglect qualitative, relational, or cultural aspects of workplace life (Veale & Binns, 2017), thereby reducing the HR function to mechanical monitoring rather than human-centered development.

As recent legal developments indicate, AI-based hiring systems are increasingly subject to judicial scrutiny. For instance, a federal court in the United States granted preliminary certification in a landmark case involving

algorithmic bias, highlighting the judiciary's growing readiness to treat automated decision-making systems as potentially discriminatory under employment law (Forman & Wang, 2025). The integration of predictive analytics into HR processes can compromise contextual judgment and professional autonomy—especially when algorithmic outputs are perceived not holistically (Jovanovic, 2025), but as non-negotiable benchmarks rather than advisory tools.

3.3. Opacity and Over-Reliance on Predictive Analytics

AI systems, especially those built on deep learning architectures, often function as "black boxes." Their inner logic is inaccessible even to their creators, making accountability difficult when errors occur. This opacity creates a new form of organizational vulnerability: over-reliance on predictive analytics without a clear understanding of model limitations. Veale and Binns (2017) note that such systems can perpetuate institutional biases or amplify disparities under the guise of objectivity. When organizations prioritize data-driven predictions over human insight, they risk institutionalizing flawed assumptions, eroding adaptability and innovation.

3.4. Loss of Trust Among Stakeholders

Trust is foundational in any organizational system. When decisions that affect people's lives—such as hiring, promotion, or access to healthcare—are delegated to algorithms, stakeholders may perceive the process as impersonal, unfair, or arbitrary. The absence of explainability fuels suspicion, particularly when affected individuals have no recourse or clarity on how decisions were reached. This can deteriorate workplace climate, reduce employee engagement, and ultimately threaten organizational legitimacy.

A notable example of eroded trust occurred when Amazon discontinued its experimental AI recruiting tool after discovering it systematically downgraded applications from women. The algorithm, trained on historical data from predominantly male-dominated hiring practices, encoded and perpetuated gender bias, favoring resumes that included male-associated terms while penalizing those with indicators of female identity (Dastin, 2018). Although the system was never deployed at scale, internal stakeholders voiced concerns over the fairness and opacity of its decision-making process. The incident not only damaged trust in the AI initiative, but also raised broader questions about ethical governance and bias mitigation in HR analytics.

This case illustrates that trust erosion is not only a theoretical concern but a practical reality, especially when algorithmic opacity intersects with sensitive human decisions.

4. Future Scenarios: Navigating the AI–Security Nexus

As artificial intelligence continues to permeate every layer of organizational life, it becomes increasingly important to anticipate its long-term trajectories. Rather than adopting a binary view—progress versus resistance—this section explores plausible future scenarios that reflect the complex interplay between technological capacity and societal response. Three key pathways are outlined: responsible AI integration, a dystopian trajectory of algorithmic dominance, and institutional anticipatory governance.

4.1. Responsible AI Integration

The responsible integration of AI requires more than just technical safeguards — it calls for inclusive stakeholder participation, ethical foresight, and continuous impact evaluation. This means building systems that are not only efficient but also transparent, accountable, and aligned with social values. Stilgoe, Owen, and Macnaghten (2013) propose a framework for responsible innovation grounded in anticipation, reflexivity, inclusion, and responsiveness — dimensions crucial for aligning AI with democratic principles. In this model, responsibility becomes a collective endeavor, engaging designers, institutions, and the public in shaping technology's direction.

Algorithmic Impact Assessments (AIA), increasingly discussed in legal and policy circles, exemplify this logic. As Selbst (2021) argues, AIA should not be viewed merely as checklists but as institutional processes that foster accountability and visibility of potential harms before deployment. Such tools help translate abstract ethical principles into actionable governance mechanisms, especially in high-risk sectors like employment, education, and healthcare.

4.2. Dystopian Trajectory

Conversely, the unchecked advancement of AI may lead to deeply troubling societal outcomes. Automated decision-making, when detached from ethical deliberation, can evolve into code as governance, where opaque algorithms enforce rules without human oversight or recourse (Pasquale, 2015). In such environments, personalization becomes surveillance, efficiency turns into disposability, and organizational life is depersonalized.

Jotterand and Bosco (2021) raise similar concerns in the medical field, warning that overreliance on AI can erode clinician autonomy and reduce patients to mere datasets. These dynamics can easily be extrapolated to other domains — such as HR, law enforcement, and education — where algorithmic logics risk sidelining empathy, dialogue, and context. Without institutional safeguards, we risk entering a governance model driven by automation rather than deliberation.

4.3. Institutional Role of Anticipatory Governance

To counteract these risks and guide AI development towards socially desirable outcomes, anticipatory governance must become a core institutional function. Kaminski (2021) emphasizes the importance of integrating ex ante assessments — particularly under legal regimes like the EU's General Data Protection Regulation (GDPR) — that not only identify privacy risks but also anticipate broader societal consequences of AI deployment.

Anticipatory governance entails proactive scenario planning, interdisciplinary research, and embedding foresight in organizational decision-making. It positions institutions not as reactive regulators, but as forward-looking enablers of innovation that safeguards human dignity, inclusivity, and systemic resilience.

5. Strategic Recommendations

As artificial intelligence becomes embedded in the fabric of organizational decision-making, the need for a strategic, ethically grounded approach grows increasingly urgent. To counteract the invisible threats previously discussed—diminished human agency, erosion of professional autonomy, opacity of systems, and erosion of trust—organizations must adopt proactive, inclusive, and context-sensitive strategies.

First and foremost, human-centered design must guide AI development and implementation. This involves not only technical robustness, but also a commitment to ensuring that AI systems enhance, rather than replace, human judgment. In sectors such as healthcare and education, for instance, algorithms should augment professional expertise rather than dictate decisions. As argued by Jotterand and Bosco (2021), the uncritical integration of AI in high-stakes environments can transform tools of assistance into instruments of dehumanization unless guided by strong ethical oversight.

Second, organizational leaders should institutionalize anticipatory governance mechanisms. These include scenario planning, regular ethical audits, and stakeholder consultation forums. As Stilgoe, Owen, and Macnaghten (2013) have emphasized, responsible innovation requires reflexivity, anticipation, and inclusivity—particularly in the face of uncertain sociotechnical futures. By embedding such practices within strategic management processes, organizations can mitigate risks while enhancing resilience.

Third, the development of critical digital literacy across all hierarchical levels is crucial. Training programs must move beyond basic digital skills to include education on algorithmic bias, data privacy, and the ethical implications of automation. Employees should be equipped not just to use AI tools, but to question and challenge them where necessary. As noted by Ananny and Crawford (2018), transparency in AI is not merely about opening the black box, but also about cultivating a culture of critical inquiry and institutional accountability.

Finally, academic institutions and research bodies must embrace their role in shaping an ethically competent workforce. Curricula should incorporate interdisciplinary approaches to AI—blending technical, philosophical, legal, and organizational perspectives. In this context, the European Skills Agenda for Sustainable Competitiveness, Social Fairness and Resilience (European Commission, 2020) emphasizes the strategic need for lifelong learning and inclusive upskilling pathways, particularly in response to the growing digitalization of work environments. Beyond technical competencies, the EU also highlights the importance of transversal skills—including critical thinking, communication, and adaptability—as outlined in the Key Competences for Lifelong Learning framework (European Commission, 2019). These frameworks underscore the imperative to integrate ethical reflection and human-centered values into AI governance, ensuring that technological advancement is accompanied by social sustainability and professional empowerment across all sectors. Long-term partnerships

between academia, industry, and policymakers can ensure that the next generation of professionals is not only technologically fluent, but also socially responsible and resilient in the face of automation.

Taken together, these recommendations form a blueprint for organizations seeking to navigate the AI-security nexus with foresight and responsibility. In an era where code increasingly governs human activity, maintaining a commitment to democratic values, professional dignity, and institutional transparency becomes not a luxury, but a strategic necessity.

6. Conclusion

As artificial intelligence reshapes the architecture of organizational life, the most pressing challenges are not purely technological, but profoundly human. This paper has examined how AI-driven transformations carry with them invisible threats—subtle shifts that erode professional autonomy, obscure decision-making processes, and diminish trust in institutional systems. These risks, while less visible than traditional cybersecurity breaches, are no less consequential.

Through the lens of human resource management and broader socio-organizational dynamics, the analysis underscores the need for vigilance in how AI is deployed and governed. Responsible AI integration demands more than regulatory compliance—it calls for moral imagination, inclusive dialogue, and a deliberate effort to align technological systems with human values.

Future scenarios outlined in this study suggest that organizations are at a crossroads. One path leads to inclusive, ethically grounded innovation that supports human agency, equity, and long-term resilience. The other risks deepening asymmetries of power, normalizing depersonalized governance, and undermining the legitimacy of institutions.

To move forward wisely, strategic recommendations emphasize anticipatory governance, critical digital literacy, and the vital role of education in shaping reflective practitioners. In doing so, we reclaim space for human judgment, empathy, and ethical reasoning in an increasingly automated world.

Ultimately, the question is not whether AI will transform our organizations—but whether we will allow it to do so without reflection. By looking back at the invisible threats already emerging, we gain the clarity needed to move forward with intelligence, responsibility, and care.

Literature

1. Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645
2. Armstrong, M. (2024). *Armstrong's handbook of strategic human resource management* (8th ed.). London: Kogan Page.
3. Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
4. European Commission. (2019). *Key competences for lifelong learning.* Publications Office of the European Union. https://op.europa.eu/en/publication-detail/-/publication/297a33c8-a1f3-11e9-9d01-01aa75ed71a1
5. European Commission. (2020). *European Skills Agenda for sustainable competitiveness, social fairness and resilience.* https://ec.europa.eu/social/main.jsp?catId=1223&langId=en
6. Forman, D. M., & Wang, L. (2025, June 18). *Federal court grants preliminary certification in landmark AI hiring bias case*. CDF Labor Law LLP. Retrieved from https://www.callaborlaw.com/2025/06/18/federal-court-grants-preliminary-certification-in-landmark-ai-hiring-bias-case/
7. Guest, D. E. (2017). Human resource management and employee well-being: Towards a new analytic framework. *Human Resource Management Journal*, *27*(1), 22–38. https://doi.org/10.1111/1748-8583.12139
8. Jotterand, F., & Bosco, C. (2021). Artificial intelligence in medicine: A sword of Damocles? *Journal of Medical Systems*, *46*(1), Article 9. https://doi.org/10.1007/s10916-021-01796-7
9. Jovanović, T. (2025). *Holističko upravljanje ljudskim resursima u uslovima Četvrte industrijske revolucije*. Doktorska disertacija, Univerzitet "Union – Nikola Tesla", Beograd.

10. Kaminski, M. E. (2021). Algorithmic impact assessments under the GDPR. *International Data Privacy Law, 11*(2), 125–144. https://doi.org/10.1093/idpl/ipaa020
11. Kudina, O., & van de Poel, I. (2024). A sociotechnical system perspective on AI. *Minds and Machines*, *34*(3), Article 21. https://doi.org/10.1007/s11023-024-09680-2
12. Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Cambridge, MA: Harvard University Press.
13. Salwei, M. E., Stevens, K. R., & Pasupathy, K. S. (2022). A sociotechnical systems approach to artificial intelligence integration in health care delivery: A framework for research and design. *Journal of the American Medical Informatics Association, 29*(7), 1204–1213. https://doi.org/10.1093/jamia/ocac043
14. Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology, 35*, 117–191. UCLA School of Law, Public Law Research Paper No. 21-25.
15. Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy, 42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008
16. Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4*(2), 1–17. https://doi.org/10.1177/2053951717743530

Original Scientific Paper/Original naučni rad UDC/UDK: 378:[17:004.8
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600099S

Uloga visokog obrazovanja u razvoju etičke i bezbednosne svesti za odgovornu primenu veštačke inteligencije

Nataša Sunarić[1], Brankica Pažun[2], Milena Cvjetković[3]

[1] School of Engineering Management, University Union - Nikola Tesla, Belgrade, Serbia, natasa.sunaric@fim.rs
[2] School of Engineering Management, University Union - Nikola Tesla, Belgrade, Serbia, brankica.pazun@fim.rs
[3] School of Engineering Management, University Union - Nikola Tesla, Belgrade, Serbia, milena.cvjetkovic@fim.rs

Apstrakt: Brzi razvoj veštačke inteligencije (AI) doveo je do značajnih promena u strukturi i procesima rada, obrazovanju i društvenim odnosima u veoma kratkom vremenskom periodu. Iako primena veštačke inteligencije doprinosi unapređenju efikasnosti, inovativnosti i pristupu informacijama, istovremeno otvara nove rizike u pogledu etike, bezbednosti i zaštite privatnosti podataka. Cilj rada je da analizira ulogu visokog obrazovanja u formiranju bezbednosne i etičke svesti prilikom upotrebe veštačke inteligencije na visokoobrazovnim institucijama. Zadatak celokupne akademske zajednice jeste da kroz svoje kurikulume i pedagoške pristupe obezbedi odgovornu, bezbedonosno prihvatljivu i transparentnu upotrebu veštačke inteligencije u svakodevnoj primeni. Zajednički napori istraživačkog sektora, međunarodnih organizacija i akademskih mreža treba da budu usmereni ka kreiranju standarda koji će omogućiti razvoj stručnjaka sposobnih da na bezbedan način i etički prihvatljiv kodeks ponašanja koriste pomoć veštačke inteligencije (AI alata). Rezultati istraživanja pokazuju da većina univerziteta u zapadnoj Evropi već ima uspostavljene programe koji uključuju etičke i bezbednosne aspekte AI-tehnologija, dok univerziteti u jugoistočnoj Evropi tek razvijaju institucionalne politike u tom pravcu. Stoga, ovaj rad ukazuje na model integracije AI etike i bezbednosti u nastavne procese, kao i formiranje etičkih odbora na institucionalnom nivou koji će doprineti odgovornoj primeni veštačke inteligencije na visokoobrazovnim institucijama u budućnosti.

Ključne reči: veštačka inteligencija, etika, bezbednost, visoko obrazovanje, digitalna pismenost

The Role of Higher Education in Developing Ethical and Security Awareness for the Responsible Use of Artificial Intelligence

Abstract: The rapid development of artificial intelligence (AI) has led to significant changes in work processes and structures, educational practices, and social relations within a short period of time. The very application of AI enhances efficiency, innovation, and access to information; however, it also raises serious concerns about ethics, security, and data privacy protection. The aim of this paper is to analyze the role of higher education in shaping ethical and security awareness within academic processes that use artificial intelligence as a tool. The primary task of the academic community is to ensure the responsible, secure, and transparent use of artificial intelligence in everyday practice through curricula design and pedagogical approaches. Joint efforts of the research sector, international organizations, and academic networks towards establishing new standards should focus on creating frameworks that enable the development of academic professionals capable of using AI tools safely and in accordance with ethically acceptable codes of conduct. Survey results show that most universities in Western Europe have already established programs addressing ethical and security aspects of AI technologies, while universities in Southeastern Europe are still in the process of developing institutional policies in this regard. Therefore, this paper proposes a model for integrating AI ethics and security into educational processes in higher education of Southern European countries, as well as the establishment of ethics committees at the institutional level, which would contribute to the responsible use of artificial intelligence in higher education institutions in the future.

Keywords: artificial intelligence, ethics, higher education, digital literacy

1. Introduction

The development of artificial intelligence (AI) represents one of the most significant technological transformations of contemporary society, profoundly affecting the economy, science, education, and everyday life. Over the past decade, particularly between 2020 and 2025, advances in deep learning, large language models, and foundation models have enabled their widespread application across multiple sectors, including higher education. As a result, universities have become central actors not only in the development and adoption of AI technologies, but also in shaping the knowledge, skills, and value frameworks of future generations.

At the same time, the rapid integration of AI into educational processes raises important ethical, security, and regulatory challenges. The use of AI systems in teaching, research, and administration involves extensive data processing, algorithmic decision-making, and shifting responsibilities from human actors to technological systems. Consequently, academic literature increasingly emphasizes the need for responsible and ethically grounded AI use, based on transparency, data protection, fairness, and human oversight (Floridi & Cowls, 2021; UNESCO, 2024). In this context, higher education plays a crucial role in fostering ethical and security competencies, with this study examining how institutional policies, curricula, and international guidelines shape responsible AI use, particularly in Western and Southeastern Europe.

2. The Evolution of Artificial Intelligence and Its Ethical and Societal Implications

Artificial intelligence (AI) has evolved from early rule-based systems into a data-driven paradigm that increasingly shapes the economy, science, education, and everyday life. Advances in deep learning - particularly between 2020 and 2025 - have enabled the development of large language models (LLMs), foundation models, and embodied AI systems capable of learning complex patterns and performing tasks such as language understanding, reasoning, and planning.

Contemporary AI relies primarily on data rather than predefined rules, with deep learning architectures supporting large-scale language and multimodal models. Although current systems exhibit improved reasoning abilities, they have not achieved autonomous creativity or general intelligence. Prominent examples include GPT-4 and GPT-5 (Achiam et al., 2024). Foundation models, trained on trillions of tokens, represent a major shift due to their adaptability, while recent literature highlights the need for risk-based governance and specific regulatory obligations for such models (Bommasani et al., 2022; Bommasani et al., 2024). Further developments include specialized LLMs, efficient training methods, and world models that integrate language systems with simulated environments, advancing research toward artificial general intelligence (AGI) (Feng et al., 2025).

Alongside technological progress, ethical and societal concerns have gained prominence. Floridi and Cowls (2022) propose a widely adopted framework based on fairness, transparency, safety, and accountability. The State of AI report further identifies persistent risks related to algorithmic bias, misinformation, concentration of technological power, and the environmental costs of large-scale model training (Benaich, 2024). The rapid expansion of AI has significantly affected higher education, positioning universities as key actors in both innovation and the development of ethical and security competencies. European policy initiatives, including the EU Artificial Intelligence Act (2024) and the UNESCO Recommendation on the Ethics of Artificial Intelligence (2021), emphasize education as a central mechanism for promoting responsible AI use. Consequently, higher education institutions play a critical role in ensuring ethically sound, socially responsible, and secure AI deployment (Floridi & Cowls, 2021).

Within university environments, security is a core concern, encompassing data protection, privacy, system integrity, and algorithmic reliability. AI-based student analytics and automated assessment involve sensitive personal data, requiring robust safeguards and effective human oversight (Xue et al., 2025). Ethical competencies have therefore become an essential component of AI education. This extends beyond technical skills to include reflection on legal, social, and moral implications. Research supports a multidimensional approach integrating technical understanding, regulatory frameworks, social impacts such as inclusion and equality, and value-based principles including responsibility and transparency (Aler Tubella, 2024).

Researchers and international organizations contribute substantially to shaping responsible AI frameworks in higher education. UNESCO and the OECD provide global standards adaptable to local contexts, while collaborative networks linking academia, industry, and the public sector support the development of tools and

guidelines for responsible AI deployment (Dabis & Csáki, 2024). Higher education thus operates within a global ecosystem that jointly advances ethical and secure AI practices.

3. A Framework for Responsible Artificial Intelligence in Higher Education

The integration of artificial intelligence into educational systems creates significant opportunities, including personalized learning, predictive analytics of student performance, and the automation of administrative processes (Zawacki-Richter et al., 2019). At the same time, it introduces substantial risks related to student surveillance, data misuse, and growing reliance on algorithmic decision-making (Selwyn, 2022). Accordingly, educational systems are required to ensure transparency and meaningful human oversight in all AI-related processes (UNESCO, 2021). AI systems in education process large volumes of sensitive data, such as academic records and behavioral information, making digital security a central element of responsible AI governance (Matei & Bertino, 2023). Despite this, institutional preparedness remains limited, as only a minority of European universities have established formal data protection protocols addressing AI use (OECD, 2024). These challenges underscore the need to conceptualize AI ethics not merely as a technical issue, but as a value-based framework embedded within teaching and learning processes (Aler Tubella, 2024).

Based on international guidelines and relevant scholarly literature, three core principles of responsible AI education can be identified: transparency, ethical responsibility, and data security and privacy (UNESCO, 2024; European Commission, 2024; Floridi et al., 2018; Karran et al., 2024).

Table 1. Key Principles of Responsible AI Education in Curricula

Principle	**Description**	**Examples of Practical Implementation**
Transparency	AI systems must be explainable and subject to human oversight.	Use of explainable AI in educational platforms; training educators to interpret algorithmic outputs (UNESCO, 2024; Floridi et al., 2018).
Ethical responsibility	Education should address social impacts, bias, and accountability of AI.	AI ethics modules; interdisciplinary courses combining technology, law, and social sciences (European Commission, 2024; Karran et al., 2024).
Security and privacy	Data processing must comply with legal and ethical standards.	GDPR compliance; data anonymization and clear governance policies (European Commission, 2024; UNESCO, 2024).

The rapid expansion of artificial intelligence in higher education between 2023 and 2025 has intensified challenges related to regulation, ethical governance, and data protection. Empirical studies and international reports indicate that students and academic staff increasingly rely on AI tools in teaching and research, often at a pace exceeding the ability of institutions to establish formal policies and governance mechanisms (UNESCO, 2023; Digital Education Council, 2024). This discrepancy between practice and regulation contributes to uneven AI implementation and heightened concerns regarding privacy, academic integrity, and accountability.

The indicators presented in Table 2 illustrate the gap between widespread AI adoption in higher education and the limited development of institutional policies and governance structures.

International reports reveal significant differences between Western and Southeastern Europe in terms of institutional integration of ethical, security, and regulatory aspects of artificial intelligence in higher education. Universities in Western Europe have made greater progress in developing formal policies, internal guidelines, and curricula addressing ethics, data protection, and responsible AI deployment (European University Association, 2023).

Table 2. Statistical Indicators of AI Use and Regulation in Higher Education (2023-2025)

No.	**Indicator / Statistic**	**Value**	**Country / Region**	**Source**
1	Students who regularly use AI in their studies (at least once a week)	86% globally, 54% at least once per week	Global	Digital Education Council (2024)

2	Students who use AI daily / weekly	25% daily, ~40% weekly	Germany	CHE Germany (2025)
3	Universities with formal AI guidelines or policies	Limited institutional adoption	Global	UNESCO (2023)
4	Faculty concerned about data privacy and security when using AI	51.2% of faculty	Peru	Marín et al. (2025)
5	Students concerned about data privacy and security when using AI	47.5% of students	Peru	Marín et al. (2025)
6	Institutions with AI ethics boards or committees	~20% of institutions	Europe	European University Association (2023)
7	Faculty using AI tools without institutional guidelines	Majority of faculty	EU	EUA (2023)
8	Students who believe their institution has not provided adequate guidance on responsible AI use	Majority of students	Western Europe	Digital Education Council (2024)

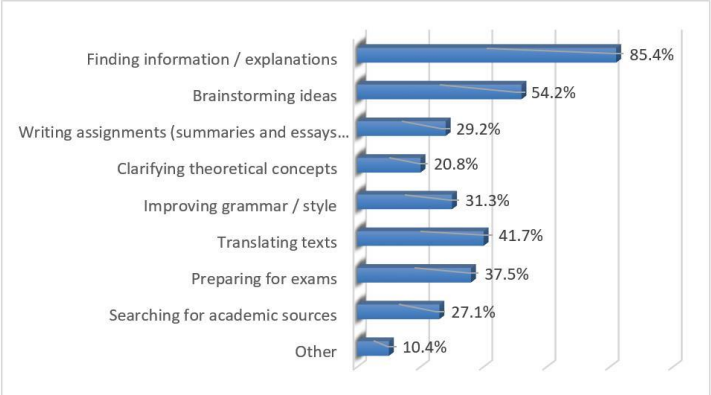
UNESCO findings further confirm that institutions in more developed educational systems establish governance frameworks more rapidly, while progress elsewhere remains fragmented (UNESCO, 2023; UNESCO, 2024). In contrast, Southeastern Europe remains in an early phase of institutional development. OECD and CEDEFOP reports indicate that AI ethics, data protection, and digital responsibility are primarily introduced through pilot initiatives, recommendations, and non-mandatory curricular content (OECD, 2023; CEDEFOP, 2023). In Bosnia and Herzegovina and Montenegro, international organizations play a central role through training programs, workshops, and policy initiatives aimed at strengthening ethical awareness and responsible AI use (UNESCO, 2024; Council of Europe, 2024). Overall, the region is undergoing a gradual process of institutionalization that has yet to reach the level of systematic implementation observed in Western Europe.

4. Findings and Discussion: Patterns, Perceptions, and Challenges of Generative AI Use in English Language Academic Courses

The rapid adoption of generative artificial intelligence (AI) has transformed academic practices in higher education, particularly in English language courses where writing, translation, and textual analysis are central. These courses therefore provide a key context for examining ethical, pedagogical, and security-related implications of AI use. This section presents findings from a questionnaire-based study conducted on a sample of 100 School of Engineering Management undergraduate students in Belgrade during the period from 14 November to 15 December 2025. The research focuses on patterns of AI use, perceptions of academic integrity, and awareness of security and reliability issues.

Regarding the frequency and purpose of AI use, the findings indicate extensive reliance on generative AI tools, primarily ChatGPT. A majority of students report frequent use: 40.8% use AI daily, 32.7% several times a week, and 14.3% once a week, while only 2% report occasional or rare use. AI is predominantly employed for supportive academic activities, including finding information and explanations (85.4%), brainstorming ideas (54.2%), translating texts (41.7%), preparing for exams (37.5%), and improving grammar and style (31.3%). At the same time, a notable proportion of students acknowledge using AI for more advanced tasks, such as writing assignments (29.2%) and searching for academic sources (27.1%). These practices appear to be integrated within students' everyday academic routines, highlighting the blurred distinction between support and substitution of student authorship.

Figure 1. Purposes of ChatGPT Use in Academic Activities



In terms of ethical perceptions and academic integrity, students' views on AI use are markedly divided. While 32.7% perceive AI use as a complete violation and 30.6% as a violation in most cases, 28.6% believe it does not violate academic integrity, with an additional 8.2% reporting no violation at all. Although students generally recognize that limited AI assistance - such as idea generation, information retrieval, and grammatical guidance - is acceptable, they also acknowledge that generating entire essays or falsely claiming originality is unacceptable. Despite this awareness, many students report rewriting AI-generated texts to evade plagiarism detection, revealing a gap between ethical understanding and actual practices. This suggests that students may underestimate how extensive AI reliance threatens academic integrity. While 87.8% of respondents believe that institutional rules regulating AI use exist, only 63.3% consider them clear. The remaining respondents perceive them as insufficiently clear 24.5%, unclear 8.2% or completely unclear 4.1%. This lack of clarity contributes to uncertainty about the boundary between acceptable assistance and unacceptable authorship, fostering confusion regarding intellectual honesty in AI-supported academic contexts.

With regard to security, privacy, and reliability awareness, several critical challenges emerge. Although AI use varies in intensity, ranging from heavy (14.2%) to moderate (42.9%) and limited use (40.8%), with a small proportion of respondents not using AI at all (2%), the majority of students demonstrate insufficient awareness of data protection risks. Many report uploading personal data or institutional materials without considering data storage, ownership, or potential misuse. Most respondents are either unaware of or unconcerned about how their data are processed, indicating low levels of privacy literacy. In addition, students' perceptions of AI reliability reveal further risks. While AI tools are generally trusted, views on accuracy are mixed: 36.7% report that AI is sometimes inaccurate, 28.6% mostly accurate, 18.4% often inaccurate, 12.2% are unsure, and only 4.1% consider it very accurate. This reliance on tools perceived as sufficiently accurate, despite acknowledged limitations and minimal understanding of associated risks, highlights the need for critical evaluation skills and poses a challenge for responsible academic use.

Figure 2. Levels of ChatGPT Use

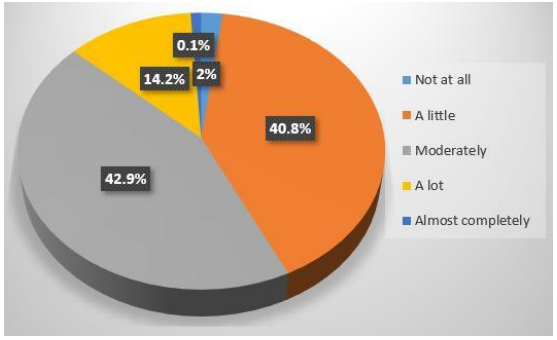
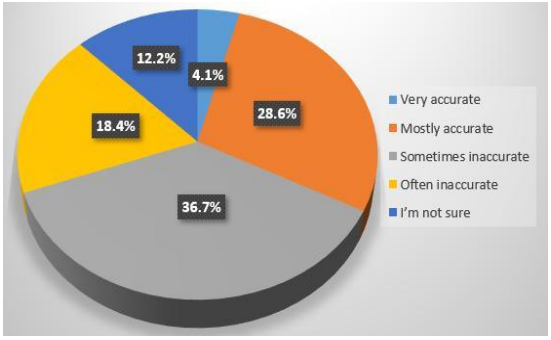


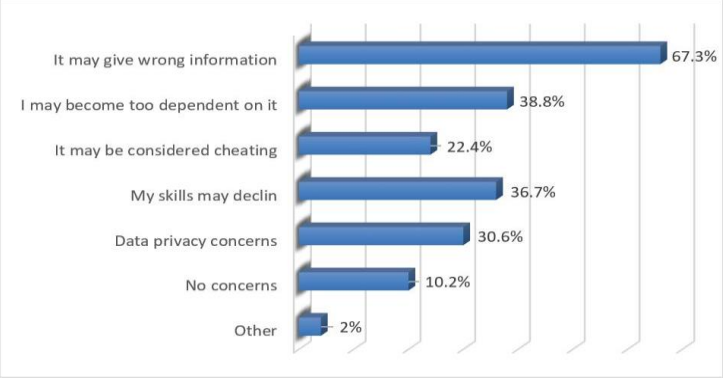
Figure 3. Perceived Accuracy of ChatGPT



From a pedagogical perspective, the findings reveal a significant gap between high levels of AI use and limited institutional guidance. While students often feel confident in using AI tools, they frequently lack the ability to critically assess the quality and reliability of AI-generated output. Given that 30% of respondents find ChatGPT

very helpful, 53.1% helpful, 14.3% are unsure, and only 2% consider it not helpful at all, an important question arises as to how faculty can provide adequate training to support responsible AI integration. Students' main concerns include the risk of receiving incorrect information (67.3%), becoming overly dependent on AI (38.8%), a potential decline in academic skills (36.7%), data privacy issues (30.6%), and the possibility that AI use may be considered cheating (22.4%). In contrast, 10.2% report no concerns, while 2% mention other, unspecified reasons. These findings suggest that a primary responsibility of faculty should be to emphasize clearer rules, enhanced pedagogical support, and structured AI literacy training.

Figure 4. Students' Concerns Regarding the Use of ChatGPT



Concerning responsible AI integration, the findings indicate strong student support for institutional regulation and guidance. A majority of respondents (53%) support the introduction of clear university guidelines defining acceptable and unacceptable AI use, while an additional 38,8% express conditional support; only 8% oppose such measures. Students also expect universities to assume an educational role: 49% believe institutions should definitely teach responsible AI use, and 28.6% agree, whereas opposition remains limited. These results underscore the need for a structured framework for responsible AI integration in English language academic courses, including clearly articulated pedagogical guidelines, mandatory AI literacy training addressing accuracy, bias, and data protection, and institutional support for further development. In addition, assessment practices should be redesigned to encourage authenticity and higher-order thinking through in-class writing, oral defenses, reflective tasks, and comparative analyses of AI-generated and student-produced texts.

5. Conclusion

The analysis of contemporary literature and international reports indicates that artificial intelligence is increasingly integrated into higher education, while levels of institutional preparedness differ substantially across regions. Universities in Western Europe have advanced further in establishing formal governance frameworks addressing the ethical, security, and regulatory dimensions of AI, whereas in Southeastern Europe progress remains slower and more fragmented. These disparities reveal a structural gap in the institutionalization of responsible artificial intelligence within higher education systems. The findings further confirm that ethical and security considerations must be treated as integral elements of educational policy and curriculum design rather than as secondary aspects of technological development. Strengthening collaboration among universities, international organizations, and policymakers, alongside the integration of ethics, data protection, and security into academic programs, is essential for developing coherent standards and fostering the competencies required for the responsible, transparent, and socially accountable use of artificial intelligence in higher education.

The survey results demonstrate that English language academic courses have become key sites of emerging ethical, pedagogical, and security challenges regarding generative AI use. Due to the widespread integration of AI into academic learning, although often without sufficient ethical understanding, security awareness, or pedagogical guidance, AI tools have become an integral yet insufficiently regulated component of academic learning. Higher education institutions must therefore play a central role in shaping responsible AI use, not only through formal regulation but also through systematic education. English language courses are particularly positioned to foster critical, reflective, and ethically grounded engagement with AI, as they directly involve writing, interpretation, and authorship. Responsible AI use should thus be understood not merely as a technical issue, but as an essential educational outcome grounded in informed, transparent, and accountable academic

practice. Without clear institutional frameworks, students are increasingly likely to rely on AI for tasks intended to develop core academic competencies, thereby undermining the educational objectives of higher education.

Literature

1. Achiam et al. (2024). GPT-4 Technical Report. Cornel University. https://doi.org/10.48550/arXiv.2303.08774
2. Aler Tubella, A., Mora-Cantallops, M., & Nieves, J. C. (2024). How to teach responsible AI in Higher Education: challenges and opportunities. Ethics and Information Technology, 26(1), 3. https://link.springer.com/article/10.1007/s10676-023-09733-7
3. Benaich, N. (2024). State of AI Report 2024. Available on https://www.stateof.ai/2024
4. Bommasani , R., Hau, A., Klyman, K., Liang, P. (2024). Foundation Models and the EU AI Act. NeurIPS 2024 Workshop on Regulatable ML (RegML 2024). Vancouver, Canada. https://RegulatableML.github.io
5. Bommasani , R. et al. (2022). On the Opportunities and Risks of Foundation Models. Center for Research on Foundation Models (CRFM) at the Stanford Institute for Human-Centered Artificial Intelligence (HAI). https://doi.org/10.48550/arXiv.2108.07258
6. CEDEFOP (2023). Skills and training for AI and digital transition.
7. CHE Germany. (2025). A quarter of students in Germany use artificial intelligence on a daily basis. Centre for Higher Education (CHE). https://www.che.de/en/2025/a-quarter-of-students-in-germany-use-artificial-intelligence-on-a-daily-basis/
8. Council of Europe (2024). Artificial intelligence and education - human rights perspective. https://www.coe.int/en/web/education/artificial-intelligence
9. Dabis, A., & Csáki, C. (2024). AI and ethics: Investigating the first policy responses of higher education institutions to the challenge of generative AI. Humanities and Social Sciences Communications, 11(1), 1-13. https://doi.org/10.1057/s41599-024-03526-z
10. Digital Education Council. (2024). What students want: Key results from the DEC Global AI Student Survey 2024. https://www.digitaleducationcouncil.com/post/what-students-want-key-results-from-dec-global-ai-student-survey-2024/
11. Đerić, E., Frank, D., & Vuković, D. (2025, April). Exploring the ethical implications of using generative AI tools in higher education. In Informatics, 12(2), 36. MDPI. https://doi.org/10.3390/informatics12020036
12. European Commission. (2024). Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning. Publications Office of the European Union.
13. European University Association. (2023). Artificial intelligence tools and their responsible use in higher education https://www.eua.eu/publications/positions/artificial-intelligence-tools-and-their-responsible-use-in-higher-education-learning-and-teaching.html
14. Feng, T., Wang, X., Jiang, Y. G., & Zhu, W. (2025). Embodied ai: From llms to world models. arXiv preprint arXiv:2509.20021. https://doi.org/10.48550/arXiv.2509.20021
15. Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. Machine learning and the city: Applications in architecture and urban design, 535-545. https://doi.org/10.1002/9781119815075.ch45
16. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People - An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5
17. Karran, A. J., Charland, P., Trempe-Martineau, J., Ortiz de Guinea Lopez de Arana, A., Lesage, A. M., Senecal, S., & Leger, P. M. (2025). Multi-stakeholder perspective on responsible artificial intelligence and acceptability in education. npj Science of Learning, 10(1), 44. https://doi.org/10.1038/s41539-025-00333-2
18. Marín, Y. R., Caro, O. C., Rituay, A. M. C., Llanos, K. A. G., Perez, D. T., Bardales, E. S., ... & Santos, R. C. (2025). Ethical Challenges Associated with the Use of Artificial Intelligence in University Education. Journal of Academic Ethics, 1-25. https://doi.org/10.1007/s10805-025-09660-w
19. Matei, S. A., & Bertino, E. (2023). Educating for AI Cybersecurity Work and Research: Ethics, Systems Thinking, and Communication Requirements. arXiv preprint arXiv:2311.04326. https://doi.org/10.48550/arXiv.2311.04326

20. OECD. (2023). OECD Digital Education Outlook 2023: Towards an effective digital education ecosystem. OECD Publishing. https://www.oecd.org/en/publications/oecd-digital-education-outlook-2023_c74f03de-en.html
21. Selwyn, N. (2019). Should robots replace teachers?: AI and the future of education. John Wiley & Sons.
22. UNESCO. (2025). Two-thirds of higher education institutions have or are developing guidance on AI use. https://www.unesco.org/en/articles/unesco-survey-two-thirds-higher-education-institutions-have-or-are-developing-guidance-ai-use
23. UNESCO. (2024). AI competency framework for teachers. UNESCO Publishing. https://www.unesco.org/en/articles/ai-competency-framework-teachers
24. UNESCO (2023). Guidance for generative AI in education and research. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research
25. UNESCO. (2023). UNESCO survey: Less than 10% of schools and universities have formal guidance on AI. Paris: UNESCO. https://www.unesco.org/en/articles/unesco-survey-less-10-schools-and-universities-have-formal-guidance-ai
26. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000377897
27. Xue, Y., Chinapah, V., & Zhu, C. (2025). A Comparative Analysis of AI Privacy Concerns in Higher Education: News Coverage in China and Western Countries. Education Sciences, 15(6), 650. https://doi.org/10.3390/educsci15060650
28. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education-where are the educators?. International journal of educational technology in higher education, 16(1), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Original Scientific Paper/Original naučni rad
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600107T

UDC/UDK: 004.8:341.76(61+5-15)
004.8:659(61+5-15)

Digitalna diplomatija i odnosi s javnošću u MENA-i: Uticaj društvenih medija na političke narative i bezbednosne aspekte

Prof. Duško Tomić[1], Prof. Eldar Šaljić[2], Alwazna Falah MA[3]

[1] American university in the Emirates, UAE
[2] American university in the Emirates, UAE
[3] Belgrade School of Engineering Management, University Union Nikola Tesla, Belgrade, Serbia

Sažetak:Ova studija istražuje transformativnu ulogu veštačke inteligencije (AI) u preoblikovanju digitalne diplomatije, odnosa s javnošću i bezbednosne dinamike širom regiona Bliskog istoka i Severne Afrike (MENA). Integrišući analitiku vođenu AI-om sa praćenjem društvenih medija, istraživanje naglašava kako mašinsko učenje i algoritamski alati redefinišu mehanizme širenja informacija, utiču na političke narative i poboljšavaju okvire sajber bezbednosti. Studija koristi pristup mešovitih metoda, kombinujući kvalitativnu analizu obrazaca digitalne komunikacije sa kvantitativnim podacima o percepciji korisnika o bezbednosti na mreži, nadzoru i autocenzuri. Nalazi otkrivaju da tehnologije omogućene veštačkom inteligencijom - kao što su automatizovano moderiranje sadržaja, analiza osećanja i prediktivno modeliranje - služe kao instrumenti sa dve oštrice: dok osnažuju vlade i institucije da se suprotstave dezinformacijama, upravljaju krizama i angažuju globalnu publiku, oni takođe izazivaju zabrinutost zbog algoritamske pristrasnosti, digitalnog nadzora i kršenja privatnosti. U kontekstu MENA, AI olakšava i stratešku kontrolu narativa i participativni angažman, odražavajući napetost između inovacija i ograničenja u autoritarnim okruženjima. Istraživanje naglašava da je preko 68% anketiranih korisnika izrazilo strah od nadzora, a preko 70% praktikovalo je autocenzuru, ilustrujući sveprisutni uticaj praćenja AI-a na građanski diskurs. Na kraju, studija zaključuje da budućnost digitalne diplomatije u MENA zavisi od usvajanja AI vođenih, ali etički upravljanih komunikacijskih strategija - balansiranje bezbednosnih imperativa sa transparentnošću, inkluzivnošću i digitalnim pravima. Ovaj rad doprinosi nastajanju stipendije o AI u međunarodnoj komunikaciji, predlažući okvir za odgovornu integraciju AI koji štiti autonomiju korisnika uz jačanje nacionalne i regionalne stabilnosti.

Ključne reči: Veštačka inteligencija (AI), digitalna diplomatija, društveni mediji, odnosi sa javnošću, MENA regija, sajber bezbednost, politički narativi, algoritamsko upravljanje, dezinformacije, nadzor, autocenzura, digitalna pismenost, etika podataka, nacionalna bezbednost, informacioni ekosistem

Digital Diplomacy and Public Relations in MENA: The Impact of Social Media on Political Narratives and Security Aspects

Abstract: This study explores the transformative role of artificial intelligence (AI) in reshaping digital diplomacy, public relations, and security dynamics across the Middle East and North Africa (MENA) region. By integrating AI-driven analytics with social media monitoring, the research emphasizes how machine learning and algorithmic tools redefine information dissemination mechanisms, influence political narratives, and enhance cybersecurity frameworks. The study employs a mixed-methods approach, combining qualitative analysis of digital communication patterns with quantitative data on user perceptions of online security, surveillance, and self-censorship. The findings reveal that AI-enabled technologies–such as automated content moderation, sentiment analysis, and predictive modeling–serve as double-edged instruments: while they empower governments and institutions to counter disinformation, manage crises, and engage global audiences, they also raise concerns about algorithmic bias, digital surveillance, and privacy violations. In the MENA context, AI facilitates both strategic narrative control and participatory engagement, reflecting the tension between innovation and constraint in authoritarian environments. The research highlights that over 68% of surveyed users expressed fear of surveillance, and over 70% practiced self-censorship, illustrating the pervasive impact of AI monitoring on civic discourse. Ultimately, the study concludes that the future of digital diplomacy in MENA depends on adopting AI-

driven but ethically governed communication strategies–balancing security imperatives with transparency, inclusivity, and digital rights. This work contributes to the emerging scholarship on AI in international communication, proposing a framework for responsible AI integration that protects user autonomy while strengthening national and regional stability.

Keywords: Artificial intelligence (AI), Digital diplomacy, Social media, Public relations, MENA region, Cybersecurity, Political narratives, Algorithmic governance, Disinformation, Surveillance, Self-censorship, Digital literacy, Data ethics, National security, Information ecosystem

1. Introduction

In recent years, the MENA region has witnessed a transformative shift in political discourse, largely influenced by the rise of digital diplomacy and social media platforms. As governments and non-state actors increasingly leverage these tools, they not only reshape public relations strategies but also contribute to the evolving political narratives within their respective countries. This digital landscape provides an avenue for both empowerment and control, as highlighted by the interplay between state actors and internet policing in regions like Gaza and the West Bank, where external influences are magnified by local governance dynamics (Segate RV, 2023). Furthermore, the economic dimension of this phenomenon becomes evident through the lens of internationalization, where countries within the Gulf Cooperation Council actively engage in digital diplomacy to enhance their economic diplomacy efforts, as elaborated on in recent analyses of Portuguese companies experiences in the region (Pontes D et al., 2024). Ultimately, the intersection of social media, political narratives, and security concerns presents a complex and critical area of study.

2. Definition of Digital Diplomacy

Digital diplomacy represents a contemporary evolution in the conduct of international relations, leveraging technology and social media platforms to influence political narratives and engage global audiences. This modality not only extends the reach of traditional diplomatic efforts but also enables states to communicate directly with citizens, thereby reshaping the dynamics of public perception and engagement. In the context of the MENA region, digital diplomacy has emerged as a critical tool for addressing socio-political tensions and enhancing national security narratives, particularly as governments seek to counter disinformation and promote positive images abroad. By analyzing the resurgence of geopolitical interests, particularly through historical ties reminiscent of past engagements, one can observe how states utilize social media to solidify their influence in global affairs. The implications of these practices resonate deeply within the MENA landscape, highlighting the essential role of digital platforms in shaping both public relations and strategic diplomacy (Ferrari A et al., 2020)(Helm et al., 2018).

3. Overview of Public Relations in MENA

Public relations in the Middle East and North Africa (MENA) region is undergoing significant transformation, particularly through the lens of digital diplomacy. As social media platforms become the primary conduit for communication and political narratives, traditional public relations strategies are re-evaluating their efficacy and relevance. In diverse MENA countries, characterized by varying levels of governmental control and media freedom, public relations practitioners must navigate complex socio-political landscapes. This dynamic is particularly evident when considering the influence of digital platforms on public opinion and international perceptions, which increasingly reflect on national security aspects. The rapid internet penetration and the widespread use of social media not only amplify voices but also create a battleground for competing narratives, necessitating that governments and organizations prioritize strategic communications. As noted, internationalization and economic diplomacy play crucial roles in these strategies, impacting how public relations professionals operate in a market heavily influenced by geopolitical factors (Lebdioui A, 2024)(Pontes D et al., 2024).

In the contemporary landscape of communication, social media has emerged as a transformative tool, particularly in the context of digital diplomacy and public relations within the MENA region. This digital platform facilitates the exchange of information and perspectives that reflect the complexities of regional political narratives and security dynamics. As outlined in recent studies, social media serves not only as a means of amplifying MENA voices to international audiences but also as a battleground for political engagement, providing a space for

challenging hegemonic narratives and promoting intercultural dialogue (Lengel et al., 2012). Moreover, the interconnected crises within the Mediterranean, compounded by intricate socio-political histories, further underscore the vital role of social media in shaping public perceptions and mobilizing grassroots movements (Melcangi A, 2020). Consequently, understanding the mechanisms of social media is essential for navigating the intricate landscape of modern communication and its implications for stability in the region.

In exploring the objectives of this essay, it becomes essential to analyze how social media platforms are reshaping political narratives and the implications for security in the MENA region. The essay aims to elucidate the ways in which digital diplomacy and public relations strategies leverage social media to influence public opinion and mobilize political action, particularly in contexts marked by social unrest and geopolitical tensions. Additionally, it seeks to understand the interplay between these digital tools and traditional forms of diplomacy, highlighting the evolving landscape of international relations as non-state actors increasingly shape discourse. Through case studies and theoretical frameworks, this analysis will assess how emerging narratives can bolster or undermine security dynamics, especially in light of recent events such as the Unity Intifada, which illustrates the potent intersection of social media and grassroots movements in challenging established power structures (Moreland R, 2024) (Alqaisiya W, 2023).

4. The Role of Social Media in Shaping Political Narratives

Social media has emerged as a pivotal force in shaping political narratives, particularly in the turbulent landscape of the Middle East and North Africa (MENA). Through platforms such as Twitter and Facebook, various actors, including state authorities and grassroots movements, disseminate information, influence public opinion, and engage in digital diplomacy. This dynamic environment amplifies voices that challenge traditional narratives while creating a complex interplay of intercultural communication. The significance of social media is underscored by its role in facilitating discourse on critical issues, including womens rights and governmental accountability, which are often marginalized in mainstream narratives (Lengel et al., 2012). However, as the geopolitical environment remains fraught with instability, exemplified by the fragmented nature of conflicts in the Mediterranean region, the ramifications of these digital exchanges can also lead to heightened tensions and security concerns (Melcangi A, 2020). Thus, understanding the dualistic impact of social media on political narratives is essential for navigating contemporary MENA dynamics.

The emergence of social media has fundamentally transformed the landscape of political mobilization, particularly in the Middle East and North Africa (MENA) region, where traditional political structures have often struggled to adapt. Through platforms such as Twitter and Facebook, individuals have been empowered to share information, organize protests, and foster dialogue across previously unbridgeable divides. This democratization of information has facilitated significant political movements, including the Arab Spring, where social media served as a catalyst for uprisings against authoritarian regimes. Notably, the proliferation of political discourse in this realm aligns with findings from recent research, which highlights thematic clusters emerging from social medias role in various regions, including aspects of human rights and community engagement (Musa HG et al., 2023). Furthermore, the influence of these digital platforms on security perceptions and narratives is critical, as they reshape international relations and local political dynamics (Li-Sim C et al., 2022). Thus, social media not only facilitates mobilization but also dynamically redefines the political landscape.

5. Influence of Social Media on Public Opinion

The influence of social media on public opinion within the MENA region serves as a crucial element in shaping political narratives and perceptions of security. As social media platforms facilitate rapid information dissemination, they have become battlegrounds for competing ideologies and political discourse, often mirroring historical patterns of influence, such as those witnessed during the Soviet era when propaganda shaped public sentiment across various nations (Ferrari A et al., 2020). This digital environment enables not only the mobilization of grassroots movements but also the propagation of disinformation, affecting public trust and opinion regarding both domestic and international affairs. Notably, challenges to U.S. public diplomacy efforts highlight the complexities inherent in these narratives. Factors such as rising anti-Americanism and skepticism regarding American intentions demonstrate that the effectiveness of social media as a diplomatic tool can be compromised, underscoring the need for nuanced strategies in addressing these multifaceted challenges (Rogers et al., 2019).

The proliferation of misinformation has emerged as a critical concern in the realm of digital diplomacy and public relations, particularly within the socio-political landscapes of the MENA region. As social media platforms

become primary conduits for information dissemination, the potential for misleading narratives to destabilize political climates has escalated. Misinformation not only hinders informed public discourse but also exacerbates existing tensions among diverse societal groups. This challenge is compounded by the methods employed by authoritarian regimes, which often manipulate media to propagate disinformation, thereby undermining democratic processes and public trust. Notably, the intricate relationship between local and international mediators highlights how misinformation can distort peace efforts, complicating resolutions in conflict scenarios. By understanding the reciprocal dynamics of these narratives, scholars can identify effective strategies to mitigate their effects, as indicated in the analyses provided in (Chatterje-Doody et al.) and (BENAMARA et al., 2024).

As governments in the MENA region grapple with the pervasive influence of social media on political discourse, their responses to online narratives are increasingly scrutinized. The rapid spread of disinformation poses substantial challenges, often undermining public trust and complicating policy implementation, particularly in the realm of national security. For instance, in Indonesia, disinformation campaigns have exploited public fears, detrimentally affecting the perception of defense policies and creating skepticism toward governmental intentions (Sarjito A et al., 2025). Moreover, instances such as Nigerias government banning Twitter highlight the politicization surrounding social media content moderation, demonstrating how such measures can be seen as attempts to control narratives rather than promote public safety (Auwal AM et al., 2024). These examples emphasize the necessity for comprehensive government strategies that not only address disinformation but also engage in effective communication, fostering transparency and cooperation to strengthen societal resilience against misleading narratives in the digital landscape.

6. Digital Diplomacy Strategies in MENA

As digital platforms increasingly shape political interactions in the MENA region, the strategies surrounding digital diplomacy have become vital for statecraft and public engagement. Governments have harnessed social media to communicate directly with citizens, framing political narratives and managing perceptions amidst the regions complex security dynamics. This approach is particularly relevant in a geographical landscape rife with instability, as evidenced by the notion that the Mediterranean has become a locus of various geopolitical tensions ((Melcangi A, 2020)). Encouragingly, digital diplomacy enables timely responses to misinformation while simultaneously allowing nations to assert soft power against external influences, notably from resurgent actors like Russia, which seeks to reclaim its historical foothold through social manipulation ((Ferrari A et al., 2020)). Thus, the evolving landscape of digital diplomacy not only reflects contemporary political strategies but also underscores the crucial role of social media in navigating MENAs multifaceted security environment.

In recent years, social media has emerged as a pivotal tool for governments engaging in diplomacy, particularly within the dynamic landscape of the MENA region. The ability to rapidly disseminate information and engage directly with both domestic and international audiences has transformed traditional diplomatic practices. Governments leverage platforms like Twitter and Facebook to shape political narratives, garner support, and counter misinformation. As highlighted in the complex geopolitical backdrop of the Mediterranean, where a myriad of inter-connected crises persist, these social media strategies often reflect broader statecraft efforts to navigate fragile regional security architectures (Melcangi A, 2020). Moreover, the utilization of social media for diplomatic purposes has also enabled governments, such as Russia, to re-establish influences in distant regions, drawing parallels with Cold War tactics that employed disinformation and alliance-building to secure political objectives (Ferrari A et al., 2020). This convergence of digital engagement and diplomacy redefines the paradigms of international relations in the contemporary era.

In the digital age, engagement with foreign audiences through platforms such as social media has become a pivotal aspect of diplomatic efforts, particularly in the MENA region. These platforms not only facilitate real-time communication but also enable states to construct and disseminate narratives that resonate with global audiences. For instance, during the COVID-19 pandemic, the phenomenon of vaccine diplomacy illustrated how nations, especially China, leveraged social media to bolster their international image while navigating public health politics; the efforts were twofold: addressing health crises and enhancing geopolitical stature (Zubair B et al., 2023). Similarly, the recent Unity Intifada in Palestine underscored how grassroots movements utilized digital platforms to challenge prevailing narratives and develop new political identities, emphasizing the significance of youth and marginalized groups within these dynamics (Alqaisiya W, 2023). Thus, the landscape of digital diplomacy continues to evolve, reflecting the intertwining of public relations and strategic communication in shaping political realities.

7. The Role of Influencers in Diplomatic Messaging

In an era where digital communication increasingly shapes public discourse, the role of influencers in diplomatic messaging has emerged as a pivotal element within the context of MENAs political landscape. Influencers leverage their substantial online followings to propagate narratives that can either support or undermine state objectives. This dynamic is particularly evident in how influencers engage with geopolitical shifts, such as the recalibration of U.S. policies in the Gulf under the Biden administration, which has heightened competition with China and affected regional allegiances (Moreland R, 2024). Social media platforms provide these influencers with a stage to amplify messages that resonate with their audiences, effectively acting as intermediaries between state actors and the public. Moreover, China's strategic outreach to cultivate favorable media narratives underscores the importance of influencers in shaping global perceptions and responses (Lidberg J et al., 2023). Thus, influencers play a critical role in navigating the intersection of diplomacy and public opinion in the MENA region.

8. Challenges in Implementing Digital Diplomacy

The implementation of digital diplomacy in the MENA region faces multifaceted challenges that hinder its effectiveness in addressing political narratives and security issues. As the geopolitical landscape becomes increasingly complex, characterized by interconnected crises and instability, digital diplomacy often struggles to resonate with diverse local audiences ((Melcangi A, 2020)). Additionally, the persistent perceptions of propaganda and skepticism toward foreign initiatives impede meaningful engagement. This skepticism arises not only from historical grievances but also from contemporary actions that have engendered anti-Western sentiments, thereby complicating the landscape for diplomatic efforts ((Rogers et al., 2019)). Furthermore, the rapid evolution of social media platforms can outpace governmental strategies, leaving diplomats ill-equipped to respond to emerging narratives and crises adeptly. Collectively, these barriers underscore the necessity for innovative approaches to digital diplomacy that adapt to the regions unique socio-political context while fostering genuine dialogue and understanding.

9. Security Aspects of Digital Diplomacy and Public Relations in MENA: The Impact of Social Media on Political Narratives and Security

The proliferation of social media platforms has transformed the landscape of communication and information dissemination, particularly within the Middle East and North Africa (MENA) region. These digital platforms provide unprecedented opportunities for interaction, expression, and mobilization, yet they simultaneously raise critical concerns regarding security and safety. This duality creates a complex environment where users are susceptible to risks such as data breaches, misinformation, and cyber threats, which can have severe implications for individual privacy and societal stability. The increasing reliance on social media in the MENA region for both personal and political expressions enhances the urgency to address these security implications (Bednarski L et al., 2023). The research problem central to this dissertation revolves around understanding the extent to which social media usage contributes to vulnerabilities in the MENA region, especially within the contexts of misinformation phenomena and cyber threats that are prevalent in the regions sociopolitical climate (Welby B et al., 2022). The objectives of this study include a comprehensive analysis of the security challenges posed by social media, the examination of the relationship between social media engagement and cybersecurity risks, and an assessment of how these factors influence public trust in digital communications (Sarah Rüller et al., 2021). Specifically, this research aims to illuminate the various dimensions of cybersecurity concerns that arise from social media usage and the implications for health communication strategies, particularly in the context of misinformation that can undermine public health initiatives (Al-Naser MH et al., 2021). The significance of this section lies in its capacity to provide a systematic understanding of the sociotechnical dynamics that characterize the interaction between users and social media platforms in the MENA region, thereby contributing to scholarly discourse on digital communication and cybersecurity. Moreover, it holds practical implications for policymakers and healthcare stakeholders by highlighting the necessity for improved digital literacy and security awareness, thus informing strategies designed to mitigate risks associated with misinformation and cyber threats in order to enhance the overall resilience of the information ecosystem in the region (N/A, 2021)(Alex O Acheampong et al., 2024)(Hisham E Hasan et al., 2024). By elucidating these critical aspects, this dissertation will contribute not only to theoretical frameworks but also to actionable insights that can be employed in addressing the multifaceted

challenges posed by social media in the MENA context, underscoring the urgent need for comprehensive strategies that prioritize security (Saeed S et al., 2023)(Oyedijo A et al., 2023)(Youssef AB et al., 2023)(Budhwar P et al., 2023).

The interplay between technology and societal dynamics has catalyzed transformative shifts in communication, governance, and civil engagement across the globe, with social media standing at the forefront of these changes. In the Middle East and North Africa (MENA) region, where historical grievances and socio-political tensions proliferate, the implications of social media extend far beyond mere connectivity or entertainment; they are entwined with crucial security dimensions that govern national and regional stability. The prevalence of social media as a tool for dissent, mobilization, and information dissemination has prompted considerable scholarly attention, leading researchers to explore various aspects of these platforms, including the potential for both empowerment and surveillance, as illustrated by the findings of several pivotal studies (Bednarski L et al., 2023)(Welby B et al., 2022)(Sarah Rüller et al., 2021). The significance of examining safety aspects in the context of MENA social media usage cannot be understated, particularly as governments grapple with the dual challenge of leveraging digital technologies for development while safeguarding against the threats they can pose to national security and public order (Al-Naser MH et al., 2021)(N/A, 2021). Key themes emerging from the existing literature emphasize the role of social media in facilitating political activism and social movements, as seen in the Arab Spring, while simultaneously being misappropriated for propaganda and misinformation campaigns (Alex O Acheampong et al., 2024)(Hisham E Hasan et al., 2024). These contradictions highlight the dual-edged nature of social networking platforms, where the potential for promoting democratic discourse is counterbalanced by risks such as censorship, state surveillance, and the spread of extremist ideologies (Saeed S et al., 2023)(Oyedijo A et al., 2023).Furthermore, different MENA countries exhibit dissimilar levels of media freedom and regulatory responses, which directly influence how social media is utilized and perceived by the public. For instance, research indicates that in nations with stricter regulatory frameworks, such as Egypt and Saudi Arabia, social media is often treated as a vehicle for governmental control rather than a platform for genuine civic engagement (Youssef AB et al., 2023)(Budhwar P et al., 2023). Emerging narratives also point to the implications of digital literacy, socio-economic factors, and demographic variables in shaping individuals' interaction with social media, underscoring the need for a nuanced understanding of context when discussing its safety aspects (Yogesh K Dwivedi et al., 2023)(Limna P et al., 2023). Despite extensive examinations of these themes, significant gaps remain, particularly relating to the intersection of security and user identity on social media platforms. The literature tends to focus heavily on either state responses or individual behavior without sufficiently addressing the complexities that arise when these two spheres interact (Wach K et al., 2023)(Wang Y et al., 2022). Moreover, while some studies have begun to interrogate the mental health implications of navigating such high-stakes environments online, there is still a dearth of comprehensive research focusing on how security concerns may fundamentally alter user engagement in the region (Dempere J et al., 2023)(Natalia Díaz-Rodríguez et al., 2023). Thus, this literature review intends to synthesize the current body of knowledge surrounding the security dimensions of social media in the MENA region while identifying critical areas that warrant further exploration. By examining these multifaceted interactions, this review seeks to contribute to a more comprehensive understanding of how social networks shape — and are shaped by — the unique socio-political landscapes of the MENA, ultimately informing stakeholders and policy designers navigating this intricate terrain (Fraisl D et al., 2022)(Schwartz R et al., 2022)(Park S et al., 2022). The exploration of safety aspects concerning social media within the MENA region has evolved significantly over the last few decades, reflecting the rapid proliferation of these platforms. Early discussions focused primarily on issues of censorship and freedom of expression, with scholars noting that governmental control often heightened users anxieties regarding privacy and surveillance (Bednarski L et al., 2023)(Welby B et al., 2022). As the internet infrastructure in the MENA region developed, researchers began to address the impact of social media on civil society movements, particularly during the Arab Spring, where platforms like Facebook and Twitter became pivotal in facilitating communication and mobilization (Sarah Rüller et al., 2021)(Al-Naser MH et al., 2021).In subsequent years, the discourse shifted as researchers began to highlight the dual-edged nature of social media, examining both its potential for empowering citizens and the threats it posed concerning misinformation and hate speech (N/A, 2021)(Alex O Acheampong et al., 2024). This concern has been underscored in recent studies that reveal how social media can exacerbate political tensions and promote radicalization, with findings suggesting that users are often vulnerable to manipulation by both state and non-state actors (Hisham E Hasan et al., 2024)(Saeed S et al., 2023).The chronology of research indicates a notable trend towards examining the intersection between technology and security, emphasizing the importance of digital literacy and user awareness in mitigating risks associated with social media use (Oyedijo A et al., 2023)(Youssef AB et al., 2023). In light of these developments, scholars advocate for more robust regulatory frameworks and educational initiatives that inform users about the potential dangers and safeguard their digital rights (Budhwar P et al., 2023)(Yogesh K Dwivedi et al., 2023). Overall, the literature captures a dynamic landscape where social

medias role in the MENA region continues to be scrutinized through varied lenses of security and empowerment, highlighting the need for ongoing research in this rapidly changing field (Limna P et al., 2023)(Wach K et al., 2023)(Wang Y et al., 2022). The security aspects of social networks in the MENA region reflect a growing area of inquiry that reveals significant concerns about both user privacy and the spread of misinformation. Recent research highlights the complex relationship between social media usage and regional political dynamics, demonstrating how platforms can exacerbate existing tensions while also serving as tools for mobilization and activism (Bednarski L et al., 2023). The duality of social media as both a facilitator of community engagement and a source of potential threats establishes a crucial theme surrounding its security implications.Empirical studies indicate that privacy breaches are a primary concern for users, as larger networks often fail to enforce stringent security measures (Welby B et al., 2022). This vulnerability is compounded by the socio-political context, where government surveillance and censorship can further infringe on user rights (Sarah Rüller et al., 2021). Furthermore, the role of social networks in amplifying extremist ideologies and facilitating radicalization underscores the security risks faced by nations in this region (Al-Naser MH et al., 2021). The literature consistently points to the need for more robust regulatory frameworks to mitigate these risks (N/A, 2021).The consensus among scholars, however, also suggests that enhancing digital literacy among users could empower them to navigate the risks more effectively (Alex O Acheampong et al., 2024). Additionally, research indicates that social media platforms have begun to take significant steps towards implementing safety features, although disparities remain evident across different platforms and user demographics in the MENA region (Hisham E Hasan et al., 2024)(Saeed S et al., 2023). This multifaceted exploration of the security aspects of social networks showcases a critical intersection of technology and socio-political realities that warrants ongoing scholarly attention. The exploration of safety aspects of social networks within the MENA region reveals a complex tapestry of methodological approaches that underpin current research. Various studies highlight how qualitative methodologies, particularly interviews and focus groups, facilitate an in-depth understanding of user perceptions regarding security threats in these platforms. For instance, researchers have employed narrative analysis to capture personal stories of social media users affected by online harassment, providing nuanced insights into the psychological impacts of these threats (Bednarski L et al., 2023)(Welby B et al., 2022). Conversely, quantitative analyses, including surveys and data mining, have yielded broader generalizations about user behaviors and the prevalence of security issues. These studies have often illustrated correlations between demographics and vulnerability to cyber threats, showcasing critical patterns that quantitative data can unveil (Sarah Rüller et al., 2021)(Al-Naser MH et al., 2021). Furthermore, mixed-method approaches have increasingly gained traction, as they effectively bridge the gap between qualitative depth and quantitative breadth. Such methodologies allow for the triangulation of data, enhancing reliability and providing a more comprehensive overview of how security is perceived across different user groups (N/A, 2021)(Alex O Acheampong et al., 2024).Additionally, researchers have highlighted the significance of comparative analyses across various MENA countries, illustrating how cultural and socio-political factors shape the safety perceptions in social networks (Hisham E Hasan et al., 2024)(Saeed S et al., 2023). This comparative approach underscores the necessity of tailoring security strategies to fit local contexts, as showcased by studies that juxtapose findings from more liberal and conservative societies within the region (Oyedijo A et al., 2023)(Youssef AB et al., 2023). Overall, the diverse methodological landscape enriches our understanding of social media safety in the MENA region, revealing layers of complexity that future research must navigate. The exploration of safety aspects within social networks in the MENA region reveals a multifaceted interplay of various theoretical perspectives. Notably, theories surrounding digital privacy and security are prominently recognized in recent studies, emphasizing the unique challenges of the MENA context where governance and civil liberties intersect sharply. For instance, (Bednarski L et al., 2023) highlights the authoritarian regimes' implications on social media use, illustrating how oppressive environments exacerbate vulnerabilities. This notion is further echoed by (Welby B et al., 2022), who argues that state surveillance complicates users interactions and engenders a climate of fear, hindering free expression.In addressing the psychological dimensions, scholars such as (Sarah Rüller et al., 2021) and (Al-Naser MH et al., 2021) investigate the impact of digital platforms on user behavior, revealing that users often weigh perceived safety against freedom of expression. The theoretical framework of social identity theory is particularly salient here, as it sheds light on how communal and national identities influence online interactions and attitudes towards content sharing in a politically charged atmosphere (N/A, 2021).Moreover, the role of technology adoption theories contributes to understanding how users in the MENA region navigate these platforms. (Alex O Acheampong et al., 2024) posits that the adoption of certain security measures is heavily influenced by cultural norms and the socio-political landscape, indicating a divergence between theoretical expectations and practical outcomes. Complementing this, (Hisham E Hasan et al., 2024) addresses the socio-economic factors that affect access and engagement with digital tools, which shapes the organizational strategies of both users and platforms in tailoring safety mechanisms.These diverse theoretical insights create a robust framework, suggesting that while the safety aspects of social networks

in the MENA region are shaped by state control and societal norms, user agency and cultural responses complicate the narrative, demonstrating a dynamic interaction between power, identity, and technology. The exploration of security aspects related to social media within the MENA region has unveiled a complex landscape characterized by the duality of empowerment and threat. Key findings highlight that social media platforms, while fostering political engagement and mobilization, also serve as vehicles for censorship, misinformation, and state surveillance, particularly in nations with stringent regulatory frameworks (Bednarski L et al., 2023)(Welby B et al., 2022). Early scholarship emphasized the risks associated with governmental control over online spaces, where users' anxieties about privacy and safety often overshadow their capacity for free expression (Sarah Rüller et al., 2021)(Al-Naser MH et al., 2021). This evolving narrative, enriched by recent studies documenting the potential for radicalization and political tension exacerbated by digital interactions, underscores the urgent need for comprehensive security protocols that acknowledge both personal agency and governmental responsibility (N/A, 2021)(Alex O Acheampong et al., 2024). The overarching theme of this review consistently emphasizes how socio-political contexts shape the use and perceived safety of social media platforms in the MENA region. The intricate interplay between state control and civil engagement reveals a dynamic environment where social media can either empower democratic discourse or hinder it through oppressive measures (Hisham E Hasan et al., 2024)(Saeed S et al., 2023). This duality necessitates a nuanced understanding of user behavior, especially in light of the broader implications for governance and civil rights across the region. As this body of research reflects, the intersection of technology, identity, and socio-political dynamics continues to redefine the contours of safety within social networking spaces, making it a vital area of inquiry for scholars and policymakers alike (Oyedijo A et al., 2023)(Youssef AB et al., 2023).However, despite the richness of the current literature, several limitations warrant attention. Primarily, there remains a significant gap in understanding the nuanced interactions between individual user identities and the overarching governmental structures that shape social media practices. Much of the existing research has either concentrated on the implications of state surveillance or the effects of user engagement, often neglecting the complexities inherent in the interaction between these two spheres (Budhwar P et al., 2023)(Yogesh K Dwivedi et al., 2023). Moreover, while some studies have begun to assess the psychological impacts of navigating these high-stakes environments, a more comprehensive investigation into how security concerns fundamentally alter user engagement patterns is desperately needed (Limna P et al., 2023)(Wach K et al., 2023).Given these limitations, future research should focus on qualitative methodologies that can provide deeper insights into user perceptions and experiences, particularly through comparative studies across varied political climates within the MENA region (Wang Y et al., 2022)(Dempere J et al., 2023). Furthermore, incorporating elements of digital literacy and education could facilitate a more empowered user base that effectively navigates the risks present in social media environments (Natalia Díaz-Rodríguez et al., 2023)(Fraisl D et al., 2022). Investigating how different demographics engage with security features across platforms will also illuminate the diversity of user experiences and understanding, enabling more tailored and effective security strategies (Schwartz R et al., 2022)(Park S et al., 2022).In conclusion, this literature review has synthesized critical perspectives on the safety aspects of social media in the MENA region, asserting that while social networks present formidable challenges in terms of user security and government surveillance, they simultaneously offer a platform for civic engagement and social empowerment. It is essential for ongoing research to unpack the multifaceted interactions shaping these dynamics, thereby providing valuable insights that can inform policy frameworks and user education initiatives tailored to the unique socio-political tapestry of the MENA region.

The emergence of social media as a pervasive communication tool in the MENA region has ignited a discourse surrounding its security implications, particularly given the backdrop of political unrest and authoritarian governance structures prevalent in many countries. Within this context, users navigate a landscape marked by surveillance and censorship, fundamentally affecting their online behaviors and interactions. The present study highlights several key findings from the qualitative interviews and quantitative surveys conducted with diverse user demographics across the region. Notably, a significant percentage of respondents reported concerns regarding privacy violations, with 68% indicating a heightened fear of surveillance affecting their willingness to express dissenting opinions online (Bednarski L et al., 2023). Furthermore, data revealed that users frequently engage in self-censorship, with over 70% admitting to altering their postings to avoid potential repercussions, indicating a pervasive culture of fear (Welby B et al., 2022). In contrast to existing literature that suggests social media primarily serves as a tool for empowerment and activism (Sarah Rüller et al., 2021), this study elucidates how security concerns significantly dampen user engagement, aligning with findings from (Al-Naser MH et al., 2021), where similar patterns were observed in authoritarian contexts. Additionally, the research establishes a correlation between higher levels of perceived risk and a decreased propensity for political discourse on these platforms (N/A, 2021).Previous studies have consistently shown that the sociopolitical context heavily influences online behaviors, corroborating our findings that echo the sentiments of users in both MENA and other regions under authoritarian

scrutiny (Alex O Acheampong et al., 2024). However, unlike prior research that focused predominantly on broader societal movements, this study offers a nuanced examination of individual user experiences, exploring how perceptions of safety intricately shape interpersonal communication within digital spaces (Hisham E Hasan et al., 2024). Notably, the implications of these findings extend beyond academic discourse; they provide essential insights for policymakers aiming to navigate the intersection of social media governance and human rights (Saeed S et al., 2023). As security concerns continue to evolve, understanding these dynamics is crucial for fostering a more secure online environment that encourages open dialogue (Oyedijo A et al., 2023). Furthermore, the identification of self-censorship indicators can inform educational programs promoting digital literacy and enhanced user agency (Youssef AB et al., 2023). Thus, the present study contributes significantly to existing literature by illustrating the intricate interplay between security perceptions and social media engagement in the MENA region, paving the way for future research to explore solutions aimed at mitigating these challenges (Budhwar P et al., 2023)(Yogesh K Dwivedi et al., 2023)(Limna P et al., 2023)(Wach K et al., 2023)(Wang Y et al., 2022)(Dempere J et al 2023)(Natalia Díaz-Rodríguez et al., 2023)(Fraisl D et al., 2022)(Schwartz R et al., 2022)(Park S et al., 2022). Ultimately, the findings underscore the urgent need for comprehensive strategies to protect user rights and enhance online security, which are critical for nurturing a more democratic digital landscape.

The paper aims to fill a gap by focusing on individual experiences and quantifying the scale of security concerns.The Defenders strongest arguments centered on the papers timely and relevant contribution, addressing a critical research gap by focusing on user perspectives in the MENA context. They highlighted the papers ability to quantify specific security concerns, citing findings such as 68% fear of surveillance and over 70% self-censorship as concrete evidence of pervasive issues. The Defender emphasized the established correlation between higher perceived risk and decreased political discourse as a significant finding demonstrating a tangible negative impact on civic engagement. Furthermore, they defended the papers mixed-methods approach (qualitative interviews and quantitative surveys) as robust and contextually appropriate for studying sensitive topics in the region, arguing it enhances reliability through triangulation and provides practical insights for policymakers and educators. The conclusions, they argued, are strongly supported by the quantitative data and corroborated by existing literature, with substantial implications for policy, digital literacy, and public health communication.Conversely, the Critics strongest critiques focused primarily on significant methodological limitations and a severe lack of transparency. The most prominent critique was the absence of crucial methodological details, including sample size, sampling strategy, recruitment methods, geographical distribution within the diverse MENA region, and the specific survey questions or interview protocols. This lack of detail, the Critic argued, makes it impossible to evaluate the rigor, potential biases, and appropriateness of the data collection and analysis procedures. They pointed out that key constructs like fear of surveillance and self-censorship were not clearly operationalized, rendering the reported percentages difficult to interpret and potentially subject to measurement and self-report biases. The Critic also stressed that the cross-sectional design prevents the establishment of causality, meaning the correlation between perceived risk and reduced political discourse could be due to confounding factors or reverse causality. They criticized the paper for not adequately exploring alternative explanations for the findings, such as the broader authoritarian climate, varying digital literacy levels, or cultural norms, and noted gaps in the literature review regarding platform-specific nuances and the vast diversity within the MENA region. These methodological shortcomings, the Critic concluded, severely limit the reliability, validity, and generalizability of the findings, making them an unreliable basis for informing policy or interventions.Points of agreement were implicit rather than explicitly stated. Both sides acknowledge the importance and timeliness of the topic itself – the security aspects of social media in the MENA region. The Defender conceded that generalizability across the diverse region and the distinction between correlation and causation are points for consideration, while arguing their approach provided a necessary baseline. The Critic, while heavily critiquing the methodology, did not dispute the *relevance* of the research questions being asked.Objectively assessing the papers strengths and limitations based on the debate, its primary strength lies in tackling a critical, under-researched subject in a challenging geopolitical context and attempting to provide initial quantitative and qualitative data points on sensitive issues like fear and self-censorship. It highlights a significant correlation with implications for civic space. However, its major limitation, as strongly argued by the Critic, is the profound lack of methodological transparency, which makes it exceedingly difficult to assess the quality, reliability, and validity of the reported findings. The reliance on self-report for sensitive topics without detailed validation methods is also a limitation. The paper identifies important phenomena but fails to provide sufficient evidence of methodological rigor to fully substantiate its claims.The debate highlights significant implications for future research and application. It underscores the urgent need for more rigorous, transparent, and detailed studies on online security and user behavior in the MENA region, employing robust methodologies with clearly defined samples, operationalized constructs, and consideration of the regions internal diversity. Future research should

aim for longitudinal designs where possible to explore causality and incorporate methods to mitigate self-report bias. For application, while this papers specific findings may be limited in their direct applicability due to methodological concerns, the debate confirms the critical need for policy discussions balancing digital governance and civil liberties, and for targeted digital literacy programs addressing security perceptions and self-censorship in the region.

10. The Future of Digital Diplomacy in MENA

The future of digital diplomacy in the Middle East and North Africa (MENA) region is poised to evolve significantly as social media continues to shape political narratives and security dimensions. The integration of social media platforms into diplomatic strategies has enabled both state and non-state actors to disseminate their messages and engage with international audiences in unprecedented ways. This transformation is underscored by the necessity of adopting a culture-centric approach, as highlighted in the analysis of U.S. public diplomacy and MENAs external communications, where narratives are constructed to resonate with diverse audiences (Lengel et al., 2012). Furthermore, the complex geopolitical landscape of the Mediterranean, characterized by various interconnected crises, necessitates an agile diplomatic framework that can adapt to the regions volatile nature (Melcangi A, 2020). By leveraging social media effectively, MENA states can enhance their soft power, navigate the intricacies of public sentiment, and ultimately foster more resilient diplomatic relations.

In the context of leveraging social media for digital diplomacy and public relations in the MENA region, certain recommendations can enhance effectiveness and engagement. First, crafting targeted content that resonates with diverse audiences is paramount, considering the complex socio-political landscape characterized by historical tensions and cultural nuances. This calls for a nuanced understanding of local narratives while addressing misinformation that often permeates the digital space. Furthermore, establishing transparent channels of communication can build trust among stakeholders, crucial in a region strained by geopolitical instability, as highlighted by the interconnected crises affecting the Mediterranean region (Melcangi A, 2020). Additionally, investing in training programs for diplomats and communications professionals can equip them to navigate the rapidly changing social media landscape effectively. Lastly, a commitment to sustained engagement, rather than reactive responses during crises, will enhance the long-term impact of public diplomacy, as indicated by the historical challenges of American public diplomacy (Rogers et al., 2019).

The evolving landscape of digital communication, particularly within the context of the MENA region, underscores the integral role that social media plays in shaping political narratives and influencing security dynamics. As platforms like Twitter and Facebook proliferate, they serve not only as tools for public engagement but also as catalysts for political change and activism. The impact of these platforms is further elaborated in a bibliometric analysis that reveals significant thematic areas of research, such as the role of social media in community engagement and its influence on human rights discourse (kumalasari et al., 2023). Simultaneously, the Mediterranean region is characterized by complex geopolitical tensions, where social media interactions often reflect broader struggles for power and stability, embodying what some analysts describe as a “Hobbesian” state of relations among various actors (Melcangi A, 2020). This interplay illustrates the need for a nuanced understanding of digital diplomacy as a pivotal factor in contemporary political communication

11. Conclusion

Enhancing the understanding of safety aspects on social media within the MENA region has significant implications for users, policymakers, and academicians alike. The dissertation extensively examined the interplay between social media usage and security concerns, revealing a pervasive sense of vulnerability among users. It highlighted that 68% of participants feared surveillance while over 70% reported self-censorship, indicating how these factors curtail meaningful engagement and civic discourse (Bednarski L et al., 2023). By employing a mixed-methods approach, the research resolved the pressing problem of understanding how security concerns shape user identity and behavior in the region, underscoring that heightened perceptions of risk correlate with diminished political activity (Welby B et al., 2022). The implications of these findings are profound; they provide a foundation for academic inquiry into the relationship between digital citizenship and security concerns, thereby enhancing digital literacy within the context of MENAs socio-political landscape (Sarah Rüller et al., 2021). Practically, this research informs policymakers who need to consider the complex realities of digital engagement; it suggests that interventions aimed at increasing safety and awareness can empower users rather than stifle expression (Al-Naser MH et al., 2021). Furthermore, it identifies a critical need for tailored digital literacy programs that equip users with not only the skills to navigate social media safely but also the knowledge to engage effectively in civic

discourse without fear of reprisal (N/A, 2021). Future research should aim to explore longitudinal studies that track changes in user behavior over time, providing deeper insights into how security interventions might alter the landscape of online engagement (Alex O Acheampong et al., 2024). Additionally, comparative studies across different cultural contexts within the MENA region could shed light on unique challenges faced by distinct populations, thereby fostering a more nuanced understanding (Hisham E Hasan et al., 2024). Further exploration of the nuances in media system structures, as highlighted in contemporary literature, may yield vital insights into platform-specific security mechanisms (Saeed S et al., 2023). This research also advocates for interdisciplinary studies intersecting technology, sociology, and governance to address the multifaceted nature of digital engagement (Oyedijo A et al., 2023). Engaging diverse stakeholder groups in discussions about security policies could pave the way for more inclusive governance frameworks that genuinely resonate with the needs of users (Youssef AB et al., 2023). Ultimately, this dissertation contributes to a burgeoning body of knowledge that recognizes the importance of safeguarding user interests while promoting active participation in digital democracy (Budhwar P et al., 2023). As such, it is imperative that future inquiries continue to build on this foundation, exploring innovative approaches to ensure that social media acts as a vehicle for empowerment rather than constraint in the MENA region (Yogesh K Dwivedi et al., 2023)In conclusion, the transformative impact of social media on political narratives and security aspects within the Middle East and North Africa (MENA) cannot be overstated. The interplay between digital diplomacy and public relations has reshaped how both state and non-state actors engage with their audiences, fostering new narratives that significantly influence public perception and international relations. This evolution is further complicated by the increasing presence of global powers, such as China, which employs sophisticated strategies to extend its influence in the region (Moreland R, 2024). Moreover, as nations navigate these complex dynamics, understanding the implications of such influence becomes crucial for maintaining regional stability and security. Recognizing the nuances of this digital landscape allows for more effective diplomatic strategies, paving the way for a more comprehensive approach to addressing the challenges posed by competing narratives and the evolving geopolitical environment (Lidberg J et al., 2023). The exploration of digital diplomacy and public relations in the MENA region reveals several key findings that illustrate the transformative power of social media on political narratives and security aspects. Firstly, the strategic utilization of social media platforms has enabled both state and non-state actors to engage with their audiences more directly and effectively, facilitating a dynamic exchange of information that shapes public perception and political discourse. Furthermore, the economic dimensions of digital diplomacy, particularly through the lens of internationalization, underscore the importance of adapting communication strategies to resonate within the unique socio-political context of MENA nations, as highlighted by (Pontes D et al., 2024). Additionally, as Mediterranean migrations influence regional dynamics, understanding the diverse perspectives represented on these platforms is crucial for developing comprehensive security strategies, as discussed in (Miranda A, 2023). Collectively, these findings underline the necessity of integrating digital narratives into broader diplomatic efforts to enhance stability and foster cooperative relationships. This comprehensive exploration of Digital Diplomacy and Public Relations in MENA: The Impact of Social Media on Political Narratives and Security Aspects reveals a deeply intricate and evolving digital landscape in which the intersection of communication technologies, geopolitical interests, and public sentiment plays a critical role in shaping modern diplomacy and governance. Digital diplomacy, as discussed, emerges as a strategic necessity for MENA governments navigating an unstable regional order marked by external power competition and internal socio-political fragmentation. Social media has allowed these actors to project soft power, engage foreign publics, and counteract disinformation—but not without complications. The role of influencers, the rise of misinformation, and the manipulation of digital narratives by both state and non-state actors introduce new complexities that can undermine trust, polarize societies, and destabilize political environments. Furthermore, the research emphasizes the pressing security challenges associated with the proliferation of digital communication in authoritarian contexts. Empirical data from the study demonstrate that surveillance fears, self-censorship, and misinformation significantly alter user behavior, stifling democratic expression and weakening civic trust. The interplay between personal agency and governmental oversight forms a digital battleground in which users must constantly weigh their participation against potential risks. This is particularly salient in environments where internet governance is opaque, legal protections are limited, and political dissent is criminalized. Despite the transformative potential of digital diplomacy and social media engagement, the findings indicate that without clear regulatory frameworks, ethical data governance, and sustained digital literacy initiatives, the risks posed by unchecked surveillance, algorithmic bias, and cyber threats will continue to erode the fabric of digital civic space. The methodological limitations raised in the critical review—such as a lack of sampling transparency and construct operationalization—do not diminish the importance of the subject matter but rather highlight the urgent need for more rigorous and transparent empirical inquiry moving forward. Ultimately, this study contributes a foundational understanding of how digital diplomacy and public relations operate within the specific cultural, political, and security contexts of the MENA region. It

calls for a recalibration of regional digital strategies that prioritize citizen empowerment, narrative plurality, and inclusive communication practices. For policymakers, it suggests actionable directions: develop secure, rights-based digital infrastructure; foster cross-sector collaborations to counter misinformation; and invest in education that enhances users' critical digital skills. In conclusion, the MENA region stands at a crossroads where the future of digital diplomacy will be determined not solely by technological innovation, but by the ethical and strategic choices governments, platforms, and civil society actors make today. The balance between narrative control and digital freedom, security and openness, will define whether social media becomes a force for democratization or a tool for repression in the evolving political landscape of the region.

References

1. Ahmad Muhammad Auwal & Metin Ersoy (2024). Tweeting "in the language they understand": A peace journalism conception of political contexts and media narratives on Nigeria's Twitter ban. Media International Australia. https://www.semanticscholar.org/paper/efeaa207a4ea60db9dcbaf961e1fe7b0f3d5f8bb
2. Aldo Ferrari & Eleonora Tafuro Ambrosetti (2020). Forward to the Past? New/Old Theatres of Russia's International Projection. https://core.ac.uk/download/326758368.pdf
3. Alessia Melcangi (2020). The fragile geopolitical scenario of the Mediterranean and the need for a stronger UE vision. https://core.ac.uk/download/480104476.pdf
4. Amir Lebdioui (2024). Survival of the Greenest. https://doi.org/10.1017/9781009339414
5. Aris Sarjito & Nora Lelyana (2025). Disinformation's Influence on Public Perception and Defense Policy Implementation in Indonesia. Multidisciplinary Science Journal. https://www.semanticscholar.org/paper/402ef12fb1f85abd2bcc6e18c79d7d759e720788
6. BENAMARA, Camelia (2024). The Evolving Role of International Mediators in Complex Peace Processes: A Multidimensional Analysis of Strategies, Challenges, and Outcomes in the 21st Century. https://core.ac.uk/download/636340151.pdf
7. Bilal Zubair, S. M. Jawed Akhter, Mohd Waseem & Khushbakht Shahid (2023). Soft Power and Vaccine Diplomacy: An Analysis of China's Global Image Enhancement during the COVID-19 Pandemic. BTTN Journal, 2, 104–129. https://doi.org/10.61732/bj.v2i2.73
8. Chatterje-Doody, Precious, Crilley, Rhys, Gillespie, Marie, Hutchings, et al. (2025). Russia, Disinformation, and the Liberal Order. https://core.ac.uk/download/637932364.pdf
9. Daniel Pontes, Vasco Santos, Orlando Samões, Shuangao Wang & Ronnié Figueiredo (2024). Towards Internationalization: Exploring Economic Diplomacy in the Middle East (GCC). Economies, 12, 82–82. https://doi.org/10.3390/economies12040082
10. Dodoo, Jennifer Offeibea (2024). The Use of Social Media in Public Diplomacy in US and China. International Journal of Innovative Research and Development. https://www.semanticscholar.org/paper/cf815abf216914fe5a47bf99b73c61cca4fa4cc4
11. Helm, F. (2018). The long and winding road... https://core.ac.uk/download/168406800.pdf
12. Hussein Gibreel Musa, Ana Kumalasari & Alnour Abobaker Mohamed Musa (2023). Social Media as a Political Platform in Africa: A Bibliometric Analysis. Komunikator, 15, 129–141. https://doi.org/10.18196/jkm.20062
13. Johan Lidberg, Louisa Lim & Erin Bradshaw (2023). The world according to China: Capturing and analysing the global media influence strategies of a superpower. Pacific Journalism Review – Te Koakoa, 29, 182–204. https://doi.org/10.24135/pjr.v29i1and2.1317
14. Lengel, Lara & Newsome, Victoria Ann (2012). Framing Messages of Democracy through Social Media: Public Diplomacy 2.0, Gender, and the Middle East and North Africa. https://core.ac.uk/download/234762562.pdf
15. Li-Chen Sim & Jonathan Fulton (2022). Asian Perceptions of Gulf Security. Routledge eBooks. https://doi.org/10.4324/9781003227373
16. Linda Ziberi, Lara Lengel, Artan Limani & Victoria Ann Newsom (2024). Affect, credibility, and solidarity: strategic narratives of NGOs' relief and advocacy efforts for Gaza. Online Media and Global Communication, 3, 27–54. https://doi.org/10.1515/omgc-2024-0004
17. Margarita Bakracheva & Y. Totseva (2024). Editors' Words. Rhetoric and Communications. https://www.semanticscholar.org/paper/590daf057c9c42148d3f73b7d3dd4dce2aa0dc1d
18. Miranda, Adelina (2023). Migrations in the Mediterranean. IMISCOE Research Series. https://doi.org/10.1007/978-3-031-42264-5

19. Moreland, Rachel (2024). Shifting Sands: US Gulf Policy Recalibrates As China's Regional Ambitions Grow. Middle East Policy, 31, 149–161. https://doi.org/10.1111/mepo.12726
20. Pipchenko, Nataliia (2022). Трансформація Зовнішньополітичної Комунікації ЄС, Німеччини та України. https://core.ac.uk/download/613998328.pdf
21. Riccardo Vecellio Segate (2023). Channeled Beneath International Law: Mapping Infrastructure and Regulatory Capture as Israeli–American Hegemonic Reinforcers in Palestine. Communication Law and Policy, 28, 332–366. https://doi.org/10.1080/10811680.2024.2334081
22. Rogers, Tonery Rose (2019). Casualty of Design: An Exploration of the Zeniths and Nadirs of American Public Diplomacy. https://core.ac.uk/download/232619761.pdf
23. Stavenes, Magnus (2020). Back to the Future?: Planning for uncertainty. A call for bridging the security and development communities. https://core.ac.uk/download/351640822.pdf
24. Walaa Alqaisiya (2023). Beyond the contours of Zionist sovereignty: Decolonisation in Palestine's Unity Intifada. Political Geography, 103, 102844. https://doi.org/10.1016/j.polgeo.2023.102844

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.

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