

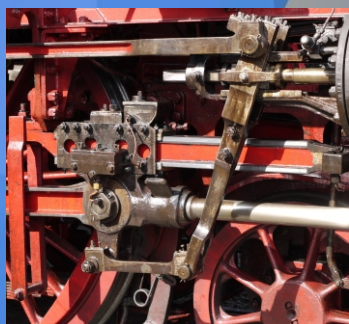
INTERNATIONAL SCIENTIFIC JOURNAL

INDUSTRY 4.0

ISSN (PRINT) 2534-8582

ISSN (WEB) 2534-997X

YEAR XI ISSUE 1/2026



①



②



③



④

PUBLISHED BY
SCIENTIFIC TECHNICAL UNION OF MECHANICAL ENGINEERING
“INDUSTRY 4.0”, BULGARIA

INTERNATIONAL SCIENTIFIC JOURNAL

INDUSTRY 4.0

YEAR XI, ISSUE 1 / 2026
ISSN (PRINT) 2534-8582, ISSN (WEB) 2534-997X

PUBLISHER
**SCIENTIFIC TECHNICAL UNION OF MECHANICAL
ENGINEERING “INDUSTRY 4.0”**

108, Rakovski Str., 1000 Sofia, Bulgaria
tel. (+359 2) 987 72 90,
tel./fax (+359 2) 986 22 40,
office@stumejournals.com
WWW.STUMEJOURNALS.COM

EDITOR IN CHIEF
Prof. D.Sc. Georgi Popov, DHC, Technical University of Sofia, BG

All articles are published after peer review by two independent reviewers.

EDITORIAL BOARD

Members:

Acad. Igor Bychkov, RU
Cor. member Alexey Beliy, BY
Cor. member Svetozar Margenov, BG
Prof. Alexander Afanasyev, RU
Prof. Alexander Guts, RU
Prof. Alshyn Altybayev, KZ
Prof. Andrzej Golabczak, PL
Prof. Bobek Shuklev, MK
Prof. Boris Gordon, EE
Prof. Branko Sirok, SI
Prof. Claudio Melchiorri, IT
Prof. Cveta Martinovska, MK
Prof. Dale Dzemydiene, LT
Prof. Dimitar Yonchev, BG
Prof. Dimitrios Vlachos, GR
Prof. Dobre Runchev, NM
Prof. Dragan Perakovic, HR
Prof. Galina Nikolcheva, BG
Prof. Galina Zhavoronkova, UA
Prof. Gerard Lyons, IE
Dr. Giovanni Pappaletta, IT
Prof. Henrik Carlsen, DK
Prof. Idilia Bachkova, BG
Prof. Idit Avrahami, IL
Prof. Inocentiu Maniu, RO
Prof. Iurii Bazhal, UA
Prof. Jürgen Köbler, DE
Prof. Jiri Maryska, CZ
Prof. Katia Vutova, BG
Prof. Lappalainen Kauko, FI
Dr. Liviu Jalba, RO

Prof. Luigi del Re, AT
Prof. Majid Zamani, DE
Dipl.-Kfm. Michael Grethler, DE
Prof. Michael Valasek, CZ
Prof. Milija Suknovic, RS
Prof. Miodrag Dashic, RS
Prof. Mohammed Taleb Obaidat, JO
Prof. Murat Alanyali, TR
Prof. Nadežda Jankelová, SK
Prof. Nafisa Yusupova, RU
Prof. Nina Bijedic, BA
Prof. Ninoslav Marina, MK
Prof. Olga Zaborovskaia, RU
Prof. Pavel Kovach, RS
Prof. Petar Kolev, BG
Prof. Peter Korondi, HU
Prof. Peter Sincak, SK
Prof. Petra Bittrich, DE
Prof. Predrag Dasic, RS
Prof. Radu Dogaru, RO
Prof. Raul Turmanidze, GE
Prof. René Beigang, DE
Prof. Rozeta Miho, AL
Prof. Salloom A. Al-Jaboori, JO
Prof. Seniye Ümit Oktay Firat, TR
Prof. Sreten Savicevic, ME
Prof. Sveto Svetkovski, MK
Prof. Tomislav Šarić, HR
Prof. Vasile Cartofeanu, MD
Prof. Vjaceslavs Bobrovs, LV
Prof. Vladyslav Alieksieiev, UA

TECHNICAL EDITORS:

M. Sc. eng. Radoslav Daskalov, Bulgaria
M. Sc. eng. Oleg Mihailov, Bulgaria

CONTENTS

TECHNOLOGICAL BASIS OF “INDUSTRY 4.0”

AI-Enhanced Cybersecurity in Critical Infrastructures: A TRITON Framework and Review

Valeri Mladenov, Gergana Vacheva, Sophia Karagiorgou, Mariza Konidi, Evangelos Kafantaris, Ioannis Pastellas, Theodora Anastasiou, Angie Carillo, Angeliki Vlachostergiou, Alberto Mozo, Esteban Martínez-Hoces, Thomas Boufikos, Panagiotis Radoglou- Grammatikis, Luca Demetrio 3

An AI-Assisted Digital Guide for Cultural and Religious Tourism Integrating Interactive Maps and Text-to-Speech Technologies: A Case Study of Svishtov

Daniela Pencheva 8

DOMINANT TECHNOLOGIES IN “INDUSTRY 4.0”

Machine Learning Prediction of Mechanical Properties for Al-Mg-Si Alloys Using a Hybrid Data Synthesis Approach

Mihail Kolev, Tatiana Simeonova 12

BUSINESS & “INDUSTRY 4.0”

A unified architectural framework for retail AI agents based on tiered functional decomposition

Ioannis Patias 15

Artificial intelligence in auditing and compliance in the telecommunications industry: a comparative empirical study of global operators

Ognen Firfov 20

SOCIETY & „INDUSTRY 4.0”

Improving the Living Environment in Evacuation Shelters Using Light-Emitting Diode Lanterns and Partitions

Mana Oobashira, Mayumi Tanaka, Keiko Ishihara, Satoru Tada, Masayuki Yamauchi, Koji Kakugawa, Takeshi Tanaka, Katia Vutova 24

Using Large Language Models for Emotional Support of Bulgarian Users: A Survey

Melania Berbatova 30

Artificial intelligence supporting human decision-making in technological crises

Zapryanova Tanya, Radeva, Tatyana 34

AI-based technologies in the authentication of fine art: toward a hybrid epistemology of cultural trust

Atanasova D. Y. 38

Artificial intelligence as a tool for systemic modelling and simulation of architectural strategies for a sustainable future

Stoyanka Ivanova 42

AI-Enhanced Cybersecurity in Critical Infrastructures: A TRITON Framework and Review

Valeri Mladenov^{1,*}, Gergana Vacheva¹, Sophia Karagiorgiou², Mariza Konidi², Evangelos Kafantaris², Ioannis Pastellas², Theodora Anastasiou², Angie Carillo³, Angeliki Vlachostergiou³, Alberto Mozo⁴, Esteban Martínez-Hoces⁴, Thomas Boufikos⁵, Panagiotis Radoglou-Grammatikis⁵, Luca Demetrio⁶

Technical University Sofia¹, UBITECH² EXUS³, Universidad

Politécnica de Madrid⁴, K3Y⁵, University of Genoa⁶

valerim@tu-sofia.bg; gergana_vacheva@tu-sofia.bg; mkonidi@ubitech.eu; a.carrillo@exus.ai; a.vlachostergiou@exus.ai; a.mozo@upm.es; esteban.martinez.hoces@upm.es; tboufikos@k3y.bg; pradoglou@k3y.bg

Abstract: Artificial intelligence (AI) is reshaping the field of penetration testing by enabling faster, more adaptive simulations of cyber threats. This paper explores how AI can be ethically integrated into penetration testing processes, focusing on the European Defence Fund-backed TRITON project. TRITON proposes a comprehensive AI-driven framework for testing the security of military and critical infrastructure systems, combining technologies like machine learning, generative models, and reinforcement learning. The TRITON project reviews recent advancements in AI-supported pentesting and discusses how these tools can improve vulnerability detection, attack simulations, and threat modeling. Alongside the technical discussion, which is examined in the TRITON project, ethical concerns—including transparency, human oversight, and dual-use risks—are of the essence and must be addressed to ensure responsible use. Comparisons with other EU cybersecurity initiatives, such as AI4CYBER and CyberSecDome, highlight TRITON's unique contributions and focus areas. Ultimately, it argues that AI-enhanced penetration testing can significantly strengthen cybersecurity defenses when implemented with appropriate safeguards and ethical oversight.

KEYWORDS: ARTIFICIAL INTELLIGENCE, PENETRATION TESTING, CYBERSECURITY, TRITON PROJECT

1. Introduction

The integration of Artificial Intelligence (AI) into cybersecurity has opened new possibilities for both defensive and offensive strategies. In particular, penetration testing—traditionally a manual and time-consuming process—has seen significant transformation through automation and AI-based tools. These technologies can simulate realistic attack scenarios, identify vulnerabilities more efficiently, and adapt to evolving threat landscapes [1-3].

However, the growing reliance on AI in this context brings with it several important considerations. Ethical questions surrounding transparency, accountability, and unintended consequences are becoming more pressing, especially when AI systems are used to probe sensitive or critical infrastructure. As AI takes on a larger role in security testing, it is vital to ensure its application aligns with ethical standards and human oversight.

This paper explores the ethical implications of AI-driven penetration testing, focusing on the European Defence Fund's TRITON project. TRITON is designed to develop advanced cybersecurity solutions for military and defense-related systems, incorporating cutting-edge AI methodologies. The project's goals include enhancing the effectiveness of penetration testing, improving situational awareness, and ensuring that AI tools operate within a framework of trust and responsibility.

Our objectives in this review are threefold:

- To examine the technological advancements driving AI-supported penetration testing – including the use of machine learning, generative models, and reinforcement learning in simulating cyberattacks and detecting vulnerabilities.
- To analyze the ethical challenges associated with deploying AI in cybersecurity, particularly in contexts involving national security, where the risks of misuse or unintended consequences are heightened.
- To position the TRITON project within the broader ecosystem of European cybersecurity initiatives, highlighting its distinct approach and the measures it takes to ensure ethical compliance.

To fully understand TRITON's contribution to AI-driven cybersecurity, it's helpful to compare it with other EU-funded initiatives operating in the same domain. While many of these projects aim to advance the state of cybersecurity using artificial intelligence, each has its own focus and intended application area.

For instance, CyberSecDome is a project that leverages AI for real-time monitoring and protection in cyber-physical systems, with an emphasis on cloud infrastructure. It uses AI techniques such as anomaly detection, predictive analytics, and autonomous threat response. Unlike TRITON, which targets secure environments in the defense and critical infrastructure sectors, CyberSecDome is more focused on enhancing situational awareness in cloud and hybrid networks.

Another initiative, AI4CYBER, works on improving cyber situational awareness and decision support through AI-based technologies. AI4CYBER focuses heavily on data fusion, threat intelligence, and human-machine collaboration. While both TRITON and AI4CYBER share the goal of trustworthy AI integration, TRITON distinguishes itself through its specialized penetration testing tools and ethical framework aligned with EU defense priorities.

What sets TRITON apart is its dedicated emphasis on AI-based automated penetration testing within regulated, high-risk environments. Unlike broader cybersecurity frameworks, TRITON is tailored specifically for secure-by-design systems, incorporating generative models, reinforcement learning, and ethical.

Moreover, TRITON integrates deeply with the European Defence Fund's strategic objectives, reflecting a dual commitment to technological innovation and ethical oversight. While other projects may emphasize real-time threat mitigation or network visibility, TRITON addresses the full lifecycle of penetration testing—starting from vulnerability identification to attack simulation and risk analysis—within a tightly governed framework.

In this way, TRITON complements the broader ecosystem of EU cybersecurity initiatives while offering a distinct, focused contribution: ethically guided, AI-driven pentesting designed for use in mission-critical infrastructures.

The following document is structured as follows: Section 2 addresses the State of the art of Penetration Testing and Automation, Section 3 addresses the TRITON Project Overview, Section 4 addresses the TRITON Focus Areas for AI-Driven Penetration Testing, Section 5 addresses the Applications of AI in Penetration Testing: State of the Art, Section 6 addresses the Challenges and Ethical Considerations in AI-Driven Pentesting, and the final Section 7 presents the Conclusion and Future Directions.

2. State of the art of Penetration Testing and Automation

Penetration testing, often referred to as ethical hacking, is a vital component of modern cybersecurity strategies. It involves simulating cyberattacks to assess the strength of an organization's defenses and identify potential vulnerabilities before malicious actors can exploit them. Traditionally, penetration testing requires skilled security professionals to manually probe systems, a process that can be time-intensive and limited in scope.

The demand for faster and more scalable testing methods has led to the increased use of automation. Tools like Nessus¹, Metasploit², and Wireshark³ have become standard in many security operations. These tools offer features such as automated vulnerability scanning, exploit development, and network analysis, making it easier to perform repetitive tasks and analyze large systems more efficiently. However, they still require human expertise to interpret results, fine-tune scans, and design attack scenarios.

Automation in penetration testing has significantly advanced with the integration of Artificial Intelligence (AI) and Machine Learning (ML). These technologies empower the creation of intelligent systems capable of learning from historical penetration data, adapting to novel threat environments, and autonomously generating sophisticated attack strategies. In particular, AI encompasses a broad spectrum of approaches, including Generative Adversarial Networks (GANs), which have been explored for generating realistic payloads or simulating adversarial traffic, and Reinforcement Learning (RL), which enables agents to iteratively learn optimal exploitation strategies by interacting with dynamic environments. These techniques complement traditional rule-based or signature-driven methods by enhancing adaptability and decision-making in complex, uncertain scenarios. AI-driven tools are especially valuable for behavioral analysis, anomaly detection in large-scale networks, and real-time vulnerability assessment, offering improved scalability, efficiency, and precision over classical automated testing frameworks. Despite these advances, full automation remains a challenge. Current AI systems can assist penetration testers—as seen with tools like PentestGPT, which provides interactive guidance to professionals—but they are not yet capable of replacing human intuition, contextual understanding, or ethical reasoning. Even though emerging research into frameworks like PentestAgent aims for a more fully autonomous approach, concerns remain about the reliability, transparency, and control of AI systems, especially in high-stakes environments such as government or critical infrastructure networks.

Several EU-backed projects, just like TRITON is addressing this gap by combining automation with human oversight to ensure safe and responsible AI deployment. These initiatives aim to balance the efficiency gains enabled by machine learning with the essential requirements of accountability, transparency, and stakeholder trust. In this context, the TRITON project stands out for its emphasis on ethically guided AI integration in penetration testing scenarios, particularly in defense-related domains. A key principle underpinning TRITON's approach is the adoption of human-in-the-loop (HITL) mechanisms, where human experts systematically review and validate AI-driven actions before execution—especially in high-risk environments. Additionally, the concept of human-on-the-loop is considered, whereby human operators monitor AI behavior and intervene when anomalies or unintended consequences arise. These human-centered control structures are increasingly recognized as best practices in the design of trustworthy and auditable AI systems. [4-6].

3. TRITON Project Overview

The TRITON project is a European Defence Fund initiative aimed at strengthening cybersecurity capabilities across Europe's defense landscape. Its central mission is to develop and implement advanced tools for AI-driven penetration testing that are not only effective but also ethically responsible and compliant with European regulations. TRITON stands for "TRusTworTHy AI for the auTomed identiFicatiON of cybersecurity threats", reflecting its focus on building trustworthy and intelligent systems.

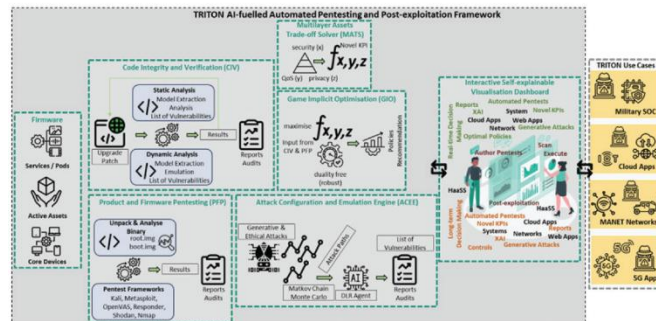


Fig.1. TRITON Conceptual Architecture[7]

Unlike traditional penetration testing tools, TRITON is designed to operate in sensitive environments such as military, defense, and other critical infrastructure systems. This raises the stakes considerably, requiring the tools to meet strict standards for safety, reliability, and ethical accountability.

At its core, TRITON seeks to automate the process of discovering and analyzing cybersecurity vulnerabilities using a combination of AI methods, including:

- Machine Learning (ML): For pattern recognition and behavioral analysis of networks and systems.
- Reinforcement Learning (RL): To train models that adapt and evolve based on simulated attack outcomes.
- Generative AI: To create synthetic attack and regular traffic scenarios and test how well systems respond to novel threats.

A distinguishing feature of TRITON is its emphasis on building "human-in-the-loop" systems, ensuring that AI tools are guided by expert judgment and ethical oversight. This approach helps mitigate the risks of over-automation and ensures that the results of AI-driven testing are interpretable, actionable, and aligned with human values.

The project also includes a modular platform architecture, allowing different tools, engines and AI modules to be integrated and customized depending on the operational context. This flexibility supports a wide range of testing environments, from enterprise IT systems to mission-critical defense infrastructure.

TRITON is part of a broader movement in the EU to establish "Trustworthy AI"—systems that are secure, fair, transparent, and aligned with fundamental rights. The project's outcomes are expected to contribute significantly to both the technological and ethical standards[10] for future AI applications in cybersecurity.

4. TRITON Focus Areas for AI-Driven Penetration Testing

The TRITON project introduces a structured, multi-phase areas for integrating AI into penetration testing, specifically used for high-security environments such as military and defense systems. This areas are designed to enhance the effectiveness of traditional testing approaches while addressing the ethical and operational challenges posed by autonomous systems.

¹ <https://www.tenable.com/products/nessus>

² <https://www.metasploit.com/>

³ <https://www.wireshark.org/>

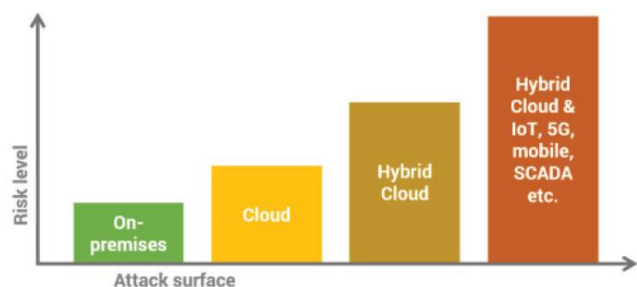


Fig.2. Cyberspace Trends in Networks, Systems, IoT-Cloud, and 5G infrastructures[7]

4.1 Automated Discovery of Vulnerabilities

At the initial stage, TRITON focuses on automatically identifying vulnerabilities within a system. This is achieved through advanced AI techniques like machine learning and pattern recognition. Unlike traditional scanners that rely on known vulnerabilities, TRITON's approach enables the system to detect novel or context-specific threats by analyzing traffic behavior, system configurations, and historical security data.

The platform combines signature-based detection with anomaly-based learning to improve accuracy. It adapts over time, using feedback from previous scans to fine-tune its models and reduce false positives. This capability is especially important in complex or evolving infrastructures, where static tools often fall short.

4.2 Simulation of Realistic Cyberattacks

Once vulnerabilities are identified, TRITON employs generative AI models to simulate realistic attack scenarios. Specifically, the system utilizes Large Language Models (LLMs) as a traffic generator, capable of creating distinct traffic types—such as realistic DDoS attacks—on demand. This allows the framework to synthesize network flows entirely from scratch without needing prior traffic samples, resulting in highly authentic connections. These simulations allow security teams to test system resilience under conditions that mirror actual cyber threats, including zero-day exploits and sophisticated multi-vector attacks.

The AI models are trained on a wide variety of attack techniques and behaviors, enabling them to generate dynamic and unpredictable test cases. This helps uncover weaknesses that might not be detected through conventional means. Importantly, these simulations are conducted in controlled environments to avoid unintended consequences or real-world disruptions.

4.3 Adaptive Learning and Reinforcement

A key innovation in TRITON's purposes is the use of reinforcement learning. Here, AI agents learn to navigate systems and develop attack strategies through trial and error, guided by reward signals based on their success in exploiting vulnerabilities. This allows for more intelligent and adaptive penetration tests that can mimic human decision-making.

However, reinforcement learning also introduces risks—such as overly aggressive or unsafe behaviors—if not carefully monitored. TRITON addresses this by integrating constraints and human oversight into the learning loop. Ethical safeguards ensure that AI agents operate within clearly defined limits, maintaining control and safety.

4.4 Reporting and Human Oversight

Throughout the process, TRITON emphasizes transparency and human involvement. Results are presented in a clear, interpretable format, allowing cybersecurity professionals to assess findings and

take corrective actions. Analysts remain in control of the decision-making process, using AI as a tool rather than a replacement.

This “human-in-the-loop” approach ensures accountability and prevents the misuse of autonomous systems. By keeping experts actively engaged, TRITON reinforces trust in AI-driven penetration testing while maintaining high ethical standards.

5. Applications of AI in Penetration Testing: State of the Art

AI has begun to reshape the landscape of penetration testing by introducing automation, adaptability, and advanced data analysis. While traditional pentesting depends heavily on manual processes, AI allows for a more scalable and dynamic approach to identifying and exploiting system vulnerabilities. As cyber threats become more complex, AI tools provide a way to respond quickly and intelligently.

One of the most significant developments in this area is the use of machine learning (ML) to analyze large volumes of network traffic, system logs, and behavioral data. By learning patterns of normal and abnormal activity, ML models can detect subtle anomalies that might indicate a security weakness or an active threat. This predictive capability improves the accuracy and efficiency of vulnerability assessments.

Reinforcement learning (RL) is another AI technique gaining traction in the pentesting field. It enables systems to learn through trial-and-error interactions, gradually improving their ability to simulate attacker behavior. RL agents can navigate complex environments, test different exploitation paths, and adapt strategies based on real-time feedback. This makes them particularly useful for discovering multi-step vulnerabilities or weaknesses in layered security defenses.

Generative AI models, such as Generative Adversarial Networks (GANs), offer a different kind of capability. These models can create realistic synthetic data, such as attack payloads or exploit scripts, which can be used to test system resilience under diverse scenarios. Generative models can also simulate variations of known attacks, helping organizations prepare for future threats that may not yet be documented.

Despite these advances, current AI systems still rely on human oversight. While they excel at processing data and running automated tasks, they often lack contextual understanding or the ability to make nuanced ethical decisions. For example, an AI system may identify a vulnerability but be unable to judge whether exploiting it in a particular context is safe or appropriate. Human expertise remains essential for interpreting results, validating findings, and making informed decisions.

A number of research initiatives [7-8] and tools have begun to integrate these AI techniques into operational workflows. Some focus on automating routine scanning and reconnaissance tasks, while others explore more advanced applications such as attack path generation, behavioral modeling, or proactive threat hunting. Collectively, these developments represent a shift toward more intelligent and autonomous penetration testing environments.

In sum, the state of the art in AI-driven penetration testing reflects a balance between automation and oversight. While AI offers significant gains in efficiency and coverage, the most effective systems still require collaboration between algorithms and skilled professionals. The TRITON project builds on these foundations, combining leading-edge AI methods with a strong commitment to ethical and secure deployment.

As AI continues to advance and integrate into cybersecurity operations, its role in penetration testing brings both powerful capabilities and serious ethical concerns. While AI can enhance efficiency and simulate complex attack behaviors, its deployment in sensitive domains—especially in military and critical infrastructure—demands careful scrutiny.

One of the central challenges is ensuring transparency. AI models, particularly those based on deep learning, often function as “black boxes,” making it difficult to understand how they arrive at certain conclusions or decisions. In the context of penetration testing, this lack of interpretability can pose risks—such as triggering unintended system disruptions or overlooking contextual nuances that a human tester might catch.

Accountability is another major concern. When an AI system autonomously performs penetration testing, it raises questions about who is responsible for any negative outcomes. Unlike human testers, AI tools may act unpredictably or make decisions that are difficult to trace back to specific design choices or user inputs. Assigning responsibility becomes even more complex in collaborative environments involving multiple stakeholders, such as defense contractors and government agencies.

There is also the issue of bias and fairness in AI algorithms. If training data used for AI-based pentesting does not represent a wide range of systems or threat models, the resulting tools may be less effective—or even harmful—when applied in unfamiliar contexts. For instance, a tool trained primarily on enterprise networks may fail to identify vulnerabilities in industrial control systems or embedded military platforms.

The dual-use nature of AI tools further complicates ethical deployment. Technologies developed for defensive penetration testing can potentially be misused for offensive purposes. Ensuring these tools are used responsibly—and not repurposed for unauthorized attacks—requires strict access control, legal oversight, and clear boundaries for acceptable use.

The TRITON project addresses these concerns by aligning its development with the European Union’s Ethics Guidelines for Trustworthy AI. These guidelines emphasize principles such as human oversight, robustness, privacy protection, and societal well-being. Within TRITON, ethical safeguards are embedded into both the design of AI tools and their operational use, ensuring that automation does not come at the expense of accountability or safety.

By maintaining a strong focus on “human-in-the-loop” governance, TRITON ensures that AI-enhanced testing remains interpretable and controllable. Human experts are kept involved in critical stages of analysis, decision-making, and validation. This balanced approach promotes both innovation and responsibility, allowing AI to enhance cybersecurity without undermining the ethical standards that are crucial in high-risk environments.

8. Conclusion and Future Directions

Artificial intelligence is rapidly transforming the field of penetration testing, offering new opportunities for automating threat discovery, simulating complex attack scenarios, and improving security analysis. The TRITON project demonstrates how AI can be thoughtfully integrated into cybersecurity practices, especially in environments where security and accountability are critical, such as defense and critical infrastructure.

In this manuscript, we have presented some conceptual architecture (Figure 1) behind TRITON, highlighting how it blends cutting-edge AI techniques—such as reinforcement learning and generative models—with a strong commitment to ethical design. Unlike many generic automation tools, TRITON is purpose-built to ensure human oversight, transparency, and alignment with EU principles for trustworthy AI.

We have also placed TRITON in the context of similar European initiatives, showing that while there is a growing ecosystem of AI-enabled cybersecurity projects, TRITON stands out for its specialized focus on automated penetration testing within ethically sensitive environments.

Looking ahead, the continued development of AI in cybersecurity must overcome several practical and strategic challenges. These include ensuring model interpretability, managing dual-use risks, preventing algorithmic bias, and maintaining meaningful human oversight. Within TRITON, these challenges are addressed through actionable mechanisms such as the use of AI Tool Registers, predefined Rules of Engagement, and ethics checkpoints throughout the testing pipeline.

Future directions are likely to involve the integration of explainable AI (XAI) techniques into AI-assisted pentesting tools, enabling human analysts to understand and audit the reasoning behind attack simulations and detection decisions. Additionally, more granular access controls, real-time risk scoring, and adaptive response strategies—driven by reinforcement learning—may be introduced to better contain potential misuse. Projects like TRITON also highlight the importance of developing sector-specific regulatory templates that align AI capabilities with GDPR, NIS2, and AI Act requirements.

Ultimately, while AI is not expected to replace human security professionals, it will play a transformative role in scaling threat detection, simulating novel attack paths, and augmenting human-led investigations. With properly enforced safeguards and governance models, AI-enhanced penetration testing can become an essential layer in defending critical infrastructure against advanced persistent threats.

Acknowledgements:

This work has received funding from the European Defence Fund 2022 Programme (EDF-2023-RA-SI) under Grant Agreement No. 101168103. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

9. References

1. Al-Sinani, H. S., & Mitchell, C. J. (2025). *PenTest++: Elevating Ethical Hacking with AI and Automation*. arXiv preprint arXiv:2502.09484. <https://arxiv.org/abs/2502.09484>
2. Ariadna Moreno, A., Sánchez, J., & Luna, J. (2025). *Analysis of Autonomous Penetration Testing Through Reinforcement Learning and Recommender Systems*. *Sensors*, 25(1), 211. <https://doi.org/10.3390/s25010211>
3. European Commission. (2022). *Trustworthy Artificial Intelligence for Cybersecurity Reinforcement and System Resilience (AI4CYBER)*. CORDIS Project ID 101070450. <https://cordis.europa.eu/project/id/101070450>
4. TRITON Consortium. (2025). *Generative Automation of Security Penetration Tests – Project Website*. <https://triton-edf.eu>
5. UBITECH. (2024, December 4). *UBITECH Hosts Kick-off of TRITON EDF Action Project to Revolutionize Automated Security Penetration Testing*. <https://ubitech.eu>
6. Jabir, R., Le, J., Nguyen, C., Öner, U., & Ashour, M. (2024). *Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Review of Human Factors*. *AI*, 6(8), 174. <https://doi.org/10.3390/ai6080174>
7. TRITON Project Consortium. *TRITON Project Partners Overview*. Available online: <https://triton-edf.eu/triton-consortium/> Ghanem, A., Ali, H., & Wang, Y. (2023). *Hierarchical Reinforcement Learning for AI-Driven Cybersecurity Attack Path Discovery*. *Computers & Security*, 130, 102868. <https://doi.org/10.1016/j.cose.2023.102868>
8. Fernandes, D. A., Rodrigues, J. J., & Kumar, N. (2024). *Using Reinforcement Learning to Simulate Advanced Persistent Threats*. *Journal of Network and Computer Applications*, 228, 103487. <https://doi.org/10.1016/j.jnca.2023.103487>
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). *Generative*

- Adversarial Nets*. Advances in Neural Information Processing Systems, 27, 2672–2680.
10. Doshi-Velez, F., & Kim, B. (2017). *Towards a Rigorous Science of Interpretable Machine Learning*. arXiv preprint arXiv:1702.08608. <https://arxiv.org/abs/1702.08608>
 11. ENISA. (2021). *Artificial Intelligence Threat Landscape*. European Union Agency for Cybersecurity. <https://www.enisa.europa.eu/publications/artificial-intelligence-threat-landscape>
 12. High-Level Expert Group on AI. (2019). *Ethics Guidelines for Trustworthy AI*. European Commission. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
 13. Shrestha, R., Mahmood, A., & Hu, J. (2021). *Review of Deep Learning Algorithms and Architectures for Cybersecurity Applications*. IEEE Access, 9, 29684–29713. <https://doi.org/10.1109/ACCESS.2021.3058794>
 14. Sommer, R., & Paxson, V. (2010). *Outside the Closed World: On Using Machine Learning for Network Intrusion Detection*. IEEE Symposium on Security and Privacy, 305–316. <https://doi.org/10.1109/SP.2010.25>
 15. CyberSecDome Consortium. (2025). *CyberSecDome Project Overview and News Release*. EIT Digital. <https://28digital.eu/news/cybersecdome-eit-cybersecurity-platform>
 16. Deng, G., Liu, Y., Mayoral-Vilches, V., Liu, P., Li, Y., Xu, Y., Zhang, T., Liu, Y., Pinzger, M., & Rass, S. (2024). *PentestGPT: An LLM-empowered Automatic Penetration Testing Tool*. <https://arxiv.org/abs/2308.06782>
 17. Shen, X., Wang, L., Li, Z., Chen, Y., Zhao, W., Sun, D., Wang, J., & Ruan, W. (2025). *PentestAgent: Incorporating LLM Agents to Automated Penetration Testing*. <https://arxiv.org/abs/2411.05185>

An AI-Assisted Digital Guide for Cultural and Religious Tourism Integrating Interactive Maps and Text-to-Speech Technologies: A Case Study of Svishtov

Daniela Pencheva¹
University of Economics - Varna, Bulgaria¹
Science and Research Institute¹
d.pencheva@ue-varna.bg¹

Abstract: This study presents the development of an AI-assisted digital guide designed to support cultural and religious tourism through the integration of interactive mapping and text-to-speech technologies. The project focuses on the cultural heritage of the city of Svishtov, Bulgaria, where several historically significant temples and monasteries are organized into a structured digital cultural route. Historical and descriptive information about each site was collected from local historiography and field observations and subsequently transformed into a digital format suitable for web presentation. An interactive map created in Google Maps allows users to visualize the spatial distribution of the sites and navigate the cultural route. In addition, an AI-generated audio guide based on text-to-speech technologies provides accessible audio narration for each location, enhancing the interpretive experience for visitors. The platform integrates visual content, geographic navigation, and automated voice narration within a single digital environment. The proposed approach demonstrates how artificial intelligence tools and interactive mapping technologies can support the interpretation of cultural heritage, improve the accessibility of historical information, and create engaging digital experiences for visitors. The study highlights the potential of combining digital humanities approaches with modern AI technologies for the promotion of local cultural heritage and the development of innovative tourism applications.

Keywords: ARTIFICIAL INTELLIGENCE, CULTURAL HERITAGE, DIGITAL TOURISM, INTERACTIVE MAPS, TEXT-TO-SPEECH, AUDIO GUIDE

1. Introduction

Digital technologies increasingly play an important role in the preservation, interpretation, and promotion of cultural heritage. In recent years, interactive platforms, geographic information tools, and artificial intelligence applications have created new opportunities for presenting historical and cultural sites in accessible and engaging ways [1]. These technologies are particularly relevant in the field of cultural and religious tourism, where visitors often seek structured information, spatial orientation, and interpretive narratives about heritage locations. Recent studies highlight the growing role of artificial intelligence and digital technologies in cultural heritage interpretation and tourism applications [2, 3]. In this context, the present study proposes the development of an AI-assisted digital guide designed to support cultural and religious tourism in the city of Svishtov.

This study explores the development of a digital cultural route focused on several historically significant temples and monasteries located in the city of Svishtov, Bulgaria. By collecting historical and descriptive information from local historiography and field observations, the project transforms traditional textual descriptions into a structured format suitable for web-based presentation. The platform integrates an interactive Google Maps route with an AI-generated audio guide based on text-to-speech technology, enabling users to explore the sites through both spatial navigation and audio interpretation. This approach demonstrates how modern digital tools can support the interpretation of cultural heritage and enhance visitor engagement. The resulting platform provides an accessible and intuitive environment that combines geographic visualization, historical information, and automated voice narration. The proposed model illustrates how artificial intelligence technologies and interactive mapping can be applied in the development of digital guides for cultural and religious tourism. The main contribution of this study lies in the development of an accessible digital platform that integrates interactive mapping, AI-assisted content structuring, and text-to-speech technologies for the interpretation and promotion of cultural and religious heritage.

The study is guided by the following research question:

How can interactive mapping and AI-based text-to-speech technologies be integrated into a digital platform to support cultural and religious tourism and improve the accessibility and interpretation of cultural heritage sites?

2. Problem discussion

Information about cultural and religious heritage sites is often dispersed across different sources, including historical publications, local historiography, archival materials, and institutional websites. In many cases, this information is presented in static textual form, making it difficult for visitors to obtain a structured and accessible overview of heritage locations within a specific region. As a result, valuable cultural sites remain insufficiently visible to potential visitors and are rarely integrated into coherent thematic routes that could support cultural and religious tourism.

In the case of Svishtov, the city possesses a rich historical and spiritual heritage, including several churches, monasteries, and sacred locations that have developed over centuries. Despite their historical significance, information about these sites is often fragmented and scattered across local publications, museum descriptions, and separate online resources. Such materials provide valuable historical insight but are not always easily accessible to visitors, particularly those who rely on digital navigation tools and mobile devices when exploring new destinations.

Another limitation is the lack of integrated digital platforms that combine geographic visualization, historical context, and interpretive narration within a single environment. Many cultural heritage descriptions remain text-based and do not incorporate modern interactive tools such as digital maps or audio interpretation. This limits the ability of visitors to engage with cultural heritage sites in a more intuitive and immersive way.

However, an integrated digital platform that combines interactive mapping, structured historical information, and AI-generated audio narration for the cultural and religious heritage sites of Svishtov has not yet been developed. This creates an opportunity to explore how modern digital technologies and artificial intelligence can support the presentation and interpretation of cultural heritage while improving accessibility for visitors and promoting local tourism.

3. Objective and research methodologies

The main objective of this study is to develop an AI-assisted digital platform designed to support cultural and religious tourism through the integration of interactive mapping and text-to-speech technologies. The project focuses on organizing and presenting the cultural heritage of the city of Svishtov, Bulgaria, by structuring historical information about local temples and monasteries into a digital cultural route. The goal is to transform traditional textual

descriptions of heritage sites into an accessible digital format that combines geographic visualization, multimedia content, and AI-generated audio narration.

The research methodology follows a structured, multi-step process (see Fig. 1). The proposed methodological framework integrates cultural heritage documentation, digital platform development, and artificial intelligence technologies in order to create an accessible digital guide for visitors. Each stage of the process contributes to transforming traditional historical information into an interactive digital environment suitable for cultural and religious tourism.

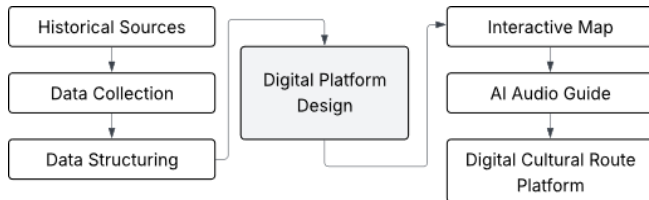


Fig. 1 Methodological framework for developing the AI-assisted digital guide

Data Collection: Historical and descriptive information about the religious heritage sites of Svishtov was collected from local historiography, cultural heritage publications, and field observations. These sources provided the historical background, architectural descriptions, and contextual information necessary for presenting each location within the digital platform.

Data Structuring: The collected information was organized into a structured dataset describing each cultural heritage site, including historical context, location, and visual documentation. This structured content served as the foundation for the digital presentation of the cultural route.

Digital Platform Design: A web-based platform was developed using Canva, allowing the integration of textual descriptions, images, interactive navigation, and geographic visualization. The platform was designed to provide a clear and intuitive interface suitable for both desktop and mobile users.

Interactive Mapping: An interactive cultural route was created using Google Maps, enabling users to visualize the spatial distribution of the heritage sites and navigate between them. Each mapped location contains descriptive information, images, and navigation support.

AI Audio Guide Generation: Text-to-speech technology was used to generate AI-based audio narration for selected cultural heritage sites. The audio guide provides an accessible interpretive layer, allowing visitors to listen to historical information while exploring the locations.

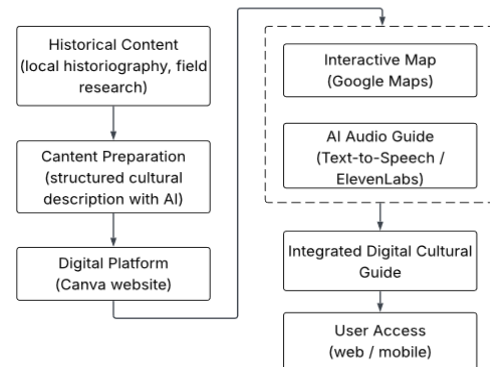
This methodological framework combines cultural heritage documentation with modern digital tools and artificial intelligence technologies. The approach demonstrates how interactive mapping, web-based platforms, and AI-generated audio narration can be integrated to create engaging and accessible guides for cultural tourism.

4. Results and visualization

The developed solution represents an integrated digital platform designed to present cultural and religious heritage sites through interactive visualization and AI-supported interpretation. The platform combines structured cultural content, geographic

navigation, and automated audio narration in order to create an accessible digital guide for visitors.

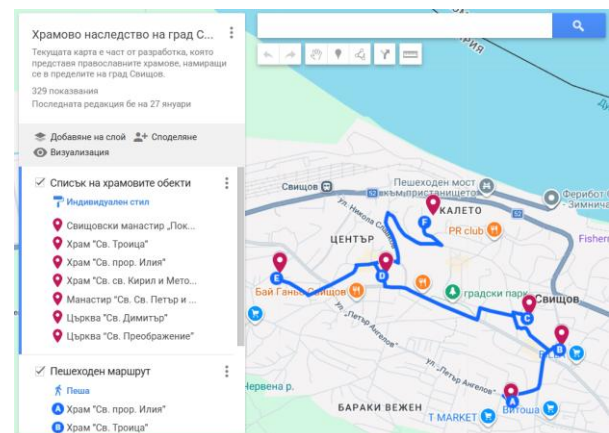
The overall system architecture integrates several technological components, including a web-based platform developed in Canva, interactive mapping through Google Maps, and AI-generated audio narration based on text-to-speech technologies (see Fig. 2) [4]. This architecture allows the platform to combine visual, spatial, and



auditory information within a single digital environment, facilitating a more engaging and intuitive user experience.

Fig. 2 System architecture of the AI-assisted digital cultural guide platform

The core element of the digital guide is an interactive cultural route created using Google Maps. The map visualizes the spatial distribution of the selected religious heritage sites in the city of Svishtov and allows users to navigate between them through a structured route. Each location is marked on the map and connected to descriptive information and visual documentation of the site, supporting intuitive geographic orientation for visitors (see Fig. 3). The map-based navigation also enables users to plan their visits



more effectively by understanding the spatial relationships between the different cultural sites within the urban environment.

Fig. 3 Interactive cultural route map of religious heritage sites in Svishtov

The digital platform interface was designed using the Canva website builder, allowing the integration of textual descriptions, images, and navigation elements within a unified visual environment [5]. The platform presents the cultural heritage objects through dedicated pages that combine historical context, visual documentation, and access to additional interpretive resources (see Fig. 4). The use of a visual website builder enabled rapid prototyping and flexible content organization, making it possible to structure cultural information in a user-friendly and easily navigable format.



Fig. 4 Web interface of the digital cultural guide platform

An additional interpretive layer is provided through an AI-generated audio guide created using text-to-speech technologies [6]. The audio narration transforms written historical descriptions into accessible spoken content that visitors can listen to while exploring the cultural sites. This functionality enhances user engagement and provides an alternative mode of accessing historical information, particularly suitable for mobile users and self-guided cultural routes (see Fig. 5). Audio narration also improves accessibility for users who prefer auditory learning or who access the platform while physically visiting the sites.

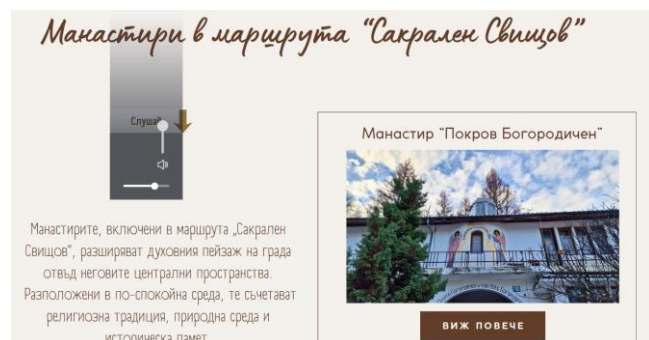


Fig. 5 AI-generated audio guide interface based on text-to-speech technology

The audio narration was generated with the ElevenLabs text-to-speech platform, which converts written descriptions into natural-sounding speech. During the development process, ChatGPT (OpenAI) was used as an assistive tool for structuring the website content, refining textual descriptions, and supporting the conceptual organization of the digital platform [7]. These AI tools contributed to the efficient transformation of historical information into accessible digital content suitable for web presentation and audio narration. The use of generative AI technologies demonstrates how contemporary digital tools can support the interpretation and dissemination of cultural heritage.

Together, these technological components form a unified digital environment that integrates interactive mapping, web-based presentation, and AI-supported audio interpretation. The resulting platform demonstrates how modern digital tools can be applied to support cultural heritage interpretation and improve accessibility to local historical information. Such approaches can contribute to the digital promotion of regional cultural resources and provide new opportunities for developing innovative tourism experiences.

5. Key findings and discussion

Figure 6 presents the final output of this study – the operational version of the developed digital platform, accessible through a QR code and web link. The platform represents an AI-assisted digital guide that integrates interactive mapping, structured cultural information, and automated audio narration into a single digital environment. Through this interface, users can access a curated

cultural route presenting religious heritage sites in the city of Svishtov, including churches and monasteries that form an important part of the local historical landscape.

The digital platform allows visitors to explore the selected heritage sites through multiple complementary layers of information. Users can navigate the route using the interactive Google Maps interface, view visual documentation of the sites, read historical descriptions, and listen to AI-generated audio narration created through text-to-speech technologies. This combination of spatial navigation, visual representation, and audio interpretation provides a richer and more engaging experience compared to traditional static information sources.

One of the key advantages of the proposed solution is the accessibility of cultural heritage information through modern digital technologies. By integrating interactive maps with web-based presentation and AI-generated audio guides, the platform supports self-guided exploration of the city's religious heritage. Visitors can access the information directly from mobile devices while physically visiting the sites, which enhances the interpretive experience and facilitates independent cultural tourism.

The developed digital guide also demonstrates the potential of low-cost and widely available digital tools for cultural heritage interpretation. The combination of Canva for website development, Google Maps for spatial visualization, and AI tools for text processing and audio generation illustrates how accessible technologies can be effectively used to build digital cultural platforms. This approach enables rapid development of digital heritage applications without the need for complex programming infrastructure.

Beyond its immediate application as a digital tourism guide, the platform also provides a methodological model for integrating artificial intelligence tools into cultural heritage presentation. The use of AI-assisted content structuring and automated voice narration contributes to the efficient transformation of historical knowledge into interactive digital formats. Such approaches can support the wider promotion of regional cultural heritage and encourage the development of innovative digital tourism initiatives.



Fig. 6 QR code providing access to the AI-assisted digital cultural guide

The QR code presented in Fig. 6 provides direct access to the developed platform, allowing users to explore the digital cultural route and experience the AI-assisted guide in practice. This approach allows users to access the digital guide instantly using mobile devices without the need for prior installation of dedicated applications. The full digital guide is also available online at:

<https://shevitsa.my.canva.site/svishtovchurcheaiguide/>.

6. Conclusion

The presented study demonstrates the development of an AI-assisted digital platform designed to support the interpretation and promotion of cultural and religious heritage through modern digital technologies. By integrating interactive mapping, web-based presentation, and AI-generated audio narration, the project illustrates how different digital tools can be combined into a unified environment for presenting historical information in an accessible and engaging way. The resulting digital guide provides users with

the opportunity to explore the religious heritage sites of Svishtov through geographic navigation, visual documentation, and audio interpretation, offering an alternative approach to traditional forms of cultural information dissemination.

The platform highlights the potential of combining interactive maps with text-to-speech technologies in order to enhance the visitor experience and improve accessibility to cultural heritage resources. Through the use of an interactive Google Maps route and structured web pages dedicated to individual sites, users are able to navigate between different cultural locations and access contextual historical information. The integration of AI-generated audio narration further expands the interpretive dimension of the platform by transforming written descriptions into accessible spoken content, allowing visitors to engage with the cultural environment while physically exploring the sites.

An important aspect of the presented approach is the use of accessible and widely available digital tools for the development of the platform. The combination of Canva for web interface design, Google Maps for spatial visualization, and AI-based text and audio technologies demonstrates how digital cultural applications can be created without the need for complex technical infrastructure. This methodological approach may serve as a practical model for similar initiatives aimed at the digital promotion of regional cultural heritage.

Beyond its immediate application as a digital tourism guide, the proposed platform contributes to the broader field of digital humanities by illustrating how artificial intelligence tools can assist in structuring, interpreting, and presenting cultural knowledge. The integration of AI-supported content generation and automated narration supports the transformation of traditional historical materials into interactive digital formats that are more accessible to contemporary audiences.

Future developments of the platform could include the expansion of the cultural route with additional heritage sites, the integration of multilingual audio guides, and the implementation of more advanced interactive functionalities. Such developments could further enhance the potential of the platform as a tool for cultural tourism, education, and public engagement with local heritage. Overall, the study demonstrates how the combination of digital mapping, web technologies, and artificial intelligence can support innovative approaches to the interpretation and promotion of cultural heritage in the digital age.

7. References

1. P. Champion, Critical gaming: Interactive history and virtual heritage, *Journal on Computing and Cultural Heritage*, vol. 14, no. 2, pp. 1–19, 2021.
2. D. Buhalis and A. Amaranggana, Smart tourism destinations enhancing tourism experience through ICT, in *Information and Communication Technologies in Tourism*, Springer, pp. 553–564, 2015.
3. M. Li, L. Xu, and L. Liu, Artificial intelligence in cultural heritage: A review, *Journal of Cultural Heritage*, vol. 49, pp. 1–12, 2021.
4. Google, Google Maps Platform Documentation, (accessed 09.03.2026) Available at: <https://developers.google.com/maps>
5. Canva Pty Ltd., Canva: Visual communication platform, (accessed 09.03.2026) Available at: <https://www.canva.com>
6. ElevenLabs, AI Voice Generator and Text-to-Speech Platform, (accessed 09.03.2026) Available at: <https://elevenlabs.io>
7. OpenAI, ChatGPT: Large language models for conversational AI, (accessed 09.03.2026) Available at: <https://openai.com>

Machine Learning Prediction of Mechanical Properties for Al-Mg-Si Alloys Using a Hybrid Data Synthesis Approach

Mihail Kolev¹, Tatiana Simeonova^{12*}

¹ Institute of Metal Science, Equipment and Technologies with Hydro- and Aerodynamics Centre "Acad. A. Balevski", Bulgarian Academy of Sciences, Sofia, 1574, Bulgaria

² Institute of Mechanics, Bulgarian Academy of Sciences, Sofia, 1113, Bulgaria

* Corresponding author: tsimeonova@imbm.bas.bg

Abstract: This study develops a machine learning framework for predicting the mechanical properties of 6xxx series Al-Mg-Si alloys (6061, 6063, 6082) across seven temper conditions. A hybrid dataset of 860 real measurement-based and 4,200 Monte Carlo augmented samples was generated from a literature-mined dataset. Random Forest (RF), Gradient Boosting (GBR), and Multilayer Perceptron (MLP) models were evaluated via 5-fold cross-validation (CV). RF achieved the best or comparable accuracy: coefficient of determination $R^2 = 0.80$ (ultimate tensile strength), 0.92 (yield strength), and 0.83 (elongation). Feature importance analysis showed that alloy type and temper encodings dominated predictions (>90% combined), while individual compositional features contributed <3% a result partly attributable to the augmentation strategy, which decoupled measured compositions from their corresponding property values. Learning curve analysis confirmed model convergence above 2,000 samples.

Keywords: Al-Mg-Si ALLOYS; MACHINE LEARNING; HYBRID DATA AUGMENTATION; MECHANICAL PROPERTIES; RANDOM FOREST; MLP NEURAL NETWORK

1. Introduction

The 6xxx series Al-Mg-Si alloys are among the most versatile aluminum alloys in automotive, construction, and general engineering applications, owing to their favorable combination of medium strength, excellent formability, good corrosion resistance, and weldability [1–3]. Alloys 6061, 6063, and 6082 represent a broad strength spectrum: 6063 is a lower-strength extrusion alloy [4], 6061 is a general-purpose structural alloy [5], and 6082 offers higher strength [6]. Their mechanical properties are governed by Mg_2Si precipitation, controlled by the temper condition [7]. Machine learning (ML) methods have shown promise for materials property prediction [8–10], but 6xxx alloys present particular challenges because alloy-to-alloy property variations are large and available data are sparse for certain conditions.

In this work, we apply three ML regression models — Random Forest, Gradient Boosting, and MLP Neural Network — to a hybrid dataset combining real measurements with Monte Carlo augmented data for predicting ultimate tensile strength (UTS), yield strength (YS), and elongation of 6061, 6063, and 6082 alloys across seven temper conditions. We evaluate model accuracy, analyze feature importance, and assess data sufficiency through learning curve analysis.

2. Materials and Methods

Composition and property data were sourced from the publicly available dataset of Pfeiffer et al. [11], which contains aluminum alloy data extracted from scientific manuscripts and US patents. Three 6xxx alloys were selected covering seven temper conditions: 6061 (T4, T6, T651), 6063 (T4, T5, T6), and 6082 (T6), consistent with their broad use in structural and extrusion applications [12,13]. The compositional features comprised 9 elements (Al, Cu, Mn, Si, Mg, Zn, Cr, Fe, Ti), with Al calculated as the balance. Figure 1 presents radar chart profiles of the normalized mechanical properties for each alloy-temper combination, highlighting the wide property envelope spanning UTS from 150 MPa (6063-T4) to 340.5 MPa (6061-T651) and YS from 90 MPa (6063-T4) to 248.0 MPa (6061-T651).

A hybrid dataset was constructed in two parts: (i) real property measurements from the corpus were each paired with 20 random composition realizations sampled from specification ranges, yielding 860 measurement-based samples; It should be noted that the 43 original measurements were unevenly distributed across conditions: 6061-T6 accounted for 33 measurements (77%), while the remaining six conditions (6082-T6, 6061-T651, 6063-T6, 6061-T4, 6063-T4, 6063-T5) each contained 1–3 measurements. In the augmentation step, each real property measurement was paired with randomly sampled compositions drawn uniformly from the alloy

specification ranges, rather than retaining the original measured composition. This approach captures the allowable compositional variability but decouples the composition–property linkage inherent in each measurement; (ii) for each of the 7 conditions, 600 additional Monte Carlo samples were generated with 3% Gaussian noise on mean property values, yielding 4,200 augmented samples (5,060 total). The feature space comprised 17 input variables: 9 elemental compositions, 3 alloy-type one-hot encodings, 4 temper one-hot encodings, and 1 ordinal temper code. Three models were trained: RF (200 trees, depth 12), GBR (200 trees, depth 6, $lr = 0.1$), and MLP (three hidden layers: 64-32-16 neurons, Rectified Linear Unit (ReLU) activation, Adam optimizer, adaptive learning rate, $\alpha = 0.001$, early stopping). Features were standardized for MLP. Performance was assessed by 5-fold CV using coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

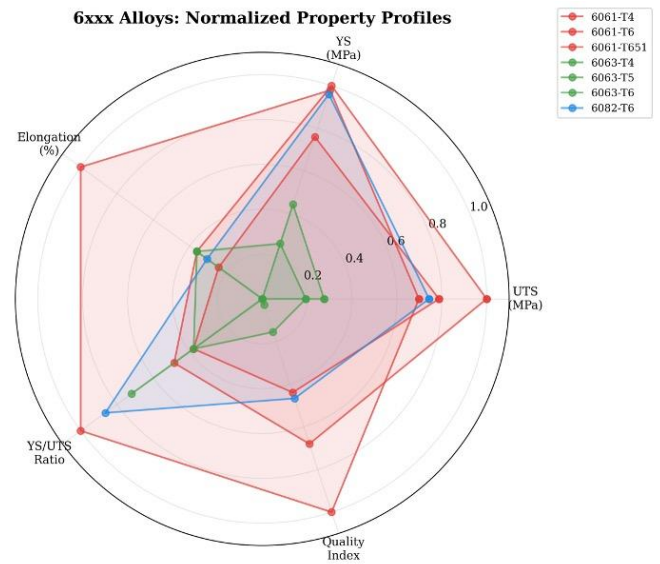


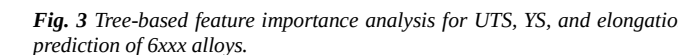
Fig. 1 Normalized radar chart profiles of mechanical properties for 6xxx alloy-temper combinations.

3. Results and Discussion

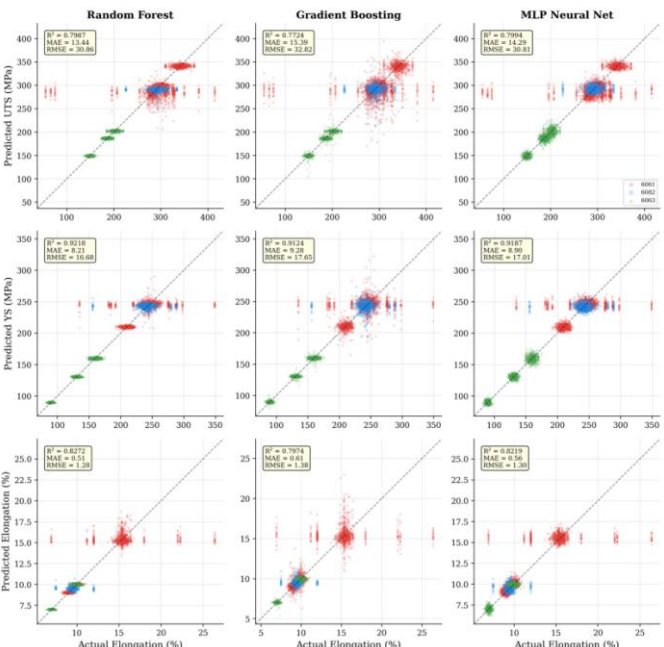
Table 1 summarizes the cross-validation metrics. RF achieved the best overall performance: $R^2 = 0.799$ (UTS), 0.922 (YS), and 0.827 (elongation), with MAE of 13.44, 8.21, and 0.51 respectively. MLP performed virtually identically for UTS ($R^2 = 0.799$) and comparably for YS ($R^2 = 0.919$). The moderate R^2 values (0.772–0.922) reflect the intrinsically wider property scatter within the 6xxx series, where UTS spans 150–340.5 MPa across alloys with

Table 1: Cross-validation performance metrics for 6xxx alloy ML models.

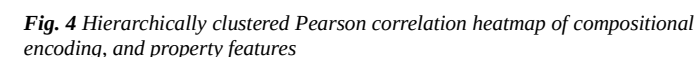
Model	Target	R ²	MAE	RMSE
Random Forest	UTS (MPa)	0.7987	13.44	30.86
	YS (MPa)	0.9218	8.21	16.68
	Elongation (%)	0.8272	0.51	1.28
Gradient Boosting	UTS (MPa)	0.7724	15.39	32.82
	YS (MPa)	0.9124	9.28	17.65
	Elongation (%)	0.7974	0.61	1.38
MLP Neural Net	UTS (MPa)	0.7994	14.29	30.81
	YS (MPa)	0.9187	8.90	17.01
	Elongation (%)	0.8219	0.56	1.30



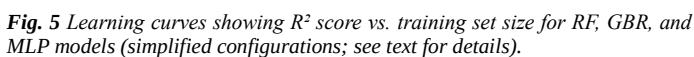
6xxx Alloys: ML Model Predictions vs. Actual Values (5-Fold CV)



Feature importance analysis (Fig. 3), assessed as the average of Random Forest and Gradient Boosting tree-based importances, revealed clear target-specific patterns. For UTS, the 6063 alloy indicator was the dominant feature (~76%), with the temper condition contributing ~5% and the T651 indicator ~3%. For YS, the 6063 indicator was overwhelmingly dominant (~83%), reflecting the large strength gap between 6063 and the other alloys. For elongation, the T6 temper (50%) and 6061 alloy (31%) indicators dominated. Individual compositional features each contributed <3%, indicating that alloy designation and temper condition encodings dominate predictions. However, this low compositional contribution is partly a consequence of the augmentation strategy, in which property measurements were paired with randomly sampled compositions from specification ranges rather than the actual measured compositions. Because the composition–property linkage was decoupled during augmentation, the models cannot learn genuine composition–property relationships within each alloy–temper class, and the low importance of individual elements should therefore be interpreted as a methodological limitation rather than a purely physical finding.



<3%, indicating that alloy designation and temper condition encodings dominate predictions. However, this low compositional contribution is partly a consequence of the augmentation strategy, in which property measurements were paired with randomly sampled compositions from specification ranges rather than the actual measured compositions. Because the composition–property linkage was decoupled during augmentation, the models cannot learn genuine composition–property relationships within each alloy-



4. Conclusions

A hybrid ML framework combining real measurements with Monte Carlo augmentation was developed for predicting UTS, YS, and elongation of 6061, 6063, and 6082 Al-Mg-Si alloys across seven temper conditions. RF provided the best or comparable performance across all three targets (UTS: $R^2 = 0.80$, MAE = 13.44 MPa; YS: $R^2 = 0.92$, MAE = 8.21 MPa; elongation: $R^2 = 0.83$, MAE = 0.51 percentage points). Alloy type and temper condition encodings dominated predictions (>90% combined importance), while compositional features contributed <3% each; the low compositional contribution is partly attributable to the augmentation strategy, in which random compositions from specification ranges were paired with property measurements, thereby decoupling the composition–property linkage. Learning curve analysis confirmed convergence above ~2,000 samples. The moderate R^2 values reflect the wide property envelope of the 6xxx series. The framework demonstrates that corpus-mined data with hybrid augmentation provides a viable route for rapid property screening of engineering aluminum alloys, although future work should explore augmentation strategies that preserve composition–property correlations within each alloy–temper class.

5. Funding

This research was funded by the Bulgarian National Science Fund, Project **KII-06-H97/10** “Fabrication and Investigation of Antifriction Hybrid Nanocomposites Based on Aluminum Alloys Using an Integrated Approach Combining Experimental Methods and Machine Learning”.

6. References

1. Chauhan, K.P.S.; FUTURE INSTITUTE OF ENGINEERING AND TECHNOLOGY, BAREILLY (U.P.) Influence of Heat Treatment on the Mechanical Properties of Aluminium Alloys (6xxx Series): A Literature Review. *Int. J. Eng. Res. Technol.* (Ahmedabad) **2017**, *V6*, doi:10.17577/ijertv6is030301.
2. Gao, Y.-H.; Zhang, X.-X.; Zhou, J.-Z.; Xu, P.; Xin, Z.-J.; Gao, J.-H.; Shi, Y.-H.; Lü, Y.; Zhao, Y.-F.; Li, J.-Y. Engineering Heterogeneous Microstructures for Enhanced Strength and Ductility in Air-Cooled Al-Mg-Si Alloys. *Trans. Nonferrous Met. Soc. China* **2025**, *35*, 1017–1031, doi:10.1016/s1003-6326(24)66731-2.
3. Hagen, A.B.; Wenner, S.; Bjørge, R.; Wan, D.; Marioara, C.; Holmestad, R.; Ringdalen, I.G. Local Mechanical Properties and Precipitation Inhomogeneity in Large-Grained Al-Mg-Si Alloys. *SSRN Electron. J.* **2022**, doi:10.2139/ssrn.4216356.
4. Effect of Copper/Alumina Hybrid Reinforcement on the Microstructure and Mechanical Properties of Stir Cast Aluminum Alloy AA6063;
5. Wang, H.; Zheng, H.; Hu, M.; Ma, Z.; Liu, H. Synergistic Effect of Al₂O₃-Decorated Reduced Graphene Oxide on Microstructure and Mechanical Properties of 6061 Aluminium Alloy. *Sci. Rep.* **2024**, *14*, 16213, doi:10.1038/s41598-024-67004-x.
6. Jiang, W.; Zhu, J.; Li, G.; Guan, F.; Yu, Y.; Fan, Z. Enhanced Mechanical Properties of 6082 Aluminum Alloy via SiC Addition Combined with Squeeze Casting. *J. Mater. Sci. Technol.* **2021**, *88*, 119–131, doi:10.1016/j.jmst.2021.01.077.
7. Qian, X.; Parson, N.; Chen, X.-G. Effect of Post-Homogenisation Cooling Rate and Mn Addition on Mg₂Si Precipitation and Hot Workability of AA6060 Alloys. *Can. Metall. Q.* **2020**, *59*, 189–200, doi:10.1080/00084433.2020.1719332.
8. Shawon, A.R.; Ghosh, R.; Islam, M.A. Analysis of Thermal Conductivity of Aluminum Alloys by Compositions and Tempering Process Using Machine Learning. *Sci. Rep.* **2025**, *15*, 33352, doi:10.1038/s41598-025-15868-y.
9. Li, J.; Zhang, Y.; Cao, X.; Zeng, Q.; Zhuang, Y.; Qian, X.; Chen, H. Accelerated Discovery of High-Strength Aluminum Alloys by Machine Learning. *Commun. Mater.* **2020**, *1*, doi:10.1038/s43246-020-00074-2.

10. J. Soofi, Y.; Rahman, M.A.; Gu, Y.; Liu, J. A Feasibility Study of Machine Learning-Assisted Alloy Design Using Wrought Aluminum Alloys as an Example. *Comput. Mater. Sci.* **2022**, *215*, 111783, doi:10.1016/j.commatsci.2022.111783.
11. Pfeiffer, O.P.; Liu, H.; Montanelli, L.; Latypov, M.I.; Sen, F.G.; Hegadekatte, V.; Olivetti, E.A.; Homer, E.R. Aluminum Alloy Compositions and Properties Extracted from a Corpus of Scientific Manuscripts and US Patents 2021.
12. Silva, M.S.; Barbosa, C.; Acselelad, O.; Pereira, L.C. Effect of Chemical Composition Variation on Microstructure and Mechanical Properties of a 6060 Aluminum Alloy. *J. Mater. Eng. Perform.* **2004**, *13*, 129–134, doi:10.1361/10599490418307.
13. Trindade, W.G.; Padovezzi, R.O.; Pinto, E.P.G.; Marques, G.A.; Amorese, R.A.; Silva, M.C.; Damasio, T.S.; Andrade, G.O. de; Luca, R. de; Nascimento, P.H.T. do Influence of Alloying Elements on the Mechanical Properties and Electrical Conductivity of 6000 Series Aluminum Alloy. A Review. *CONTRIBUCIONES A LAS CIENCIAS SOCIALES* **2025**, *18*, e18763, doi:10.55905/revconv.18n.6-193.

A unified architectural framework for retail AI agents based on tiered functional decomposition

Ioannis Patias

Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Sofia - Bulgaria¹
patias@fmi.uni-sofia.bg

Abstract: The rapid emergence of autonomous AI agents in commerce is reshaping how consumers interact with retailers, how decisions are made, and how value is exchanged across digital ecosystems. Yet despite accelerating innovation, the retail sector lacks a unified architectural model that defines how such agents should ingest information, reason about complex commercial contexts, and execute actions in a transparent, interoperable, and trustworthy manner. This paper introduces a unified architectural framework for retail AI agents based on tiered functional decomposition, offering a structured, three-tier model (Input, Model, Output) that standardizes agent behavior across diverse retail environments.

The *Input Tier* formalizes how agents acquire and normalize information, including user intent, preference profiles, product catalogs, pricing signals, inventory data, logistics constraints, and retailer policies. By defining a consistent intake layer, the framework ensures that agents operate on coherent, comparable, and semantically aligned data structures.

The *Model Tier* provides the cognitive core of the agent. It encompasses goal decomposition, relevance scoring, multi-objective optimization, negotiation logic, risk assessment, and alignment safeguards. This tier transforms raw inputs into structured decisions, enabling agents to act as rational, preference-aligned economic participants capable of navigating complex retail ecosystems.

The *Output Tier* defines how agents act on decisions and how they learn from outcomes. It includes action execution (e.g., purchasing, reserving, scheduling), transparent justification, user-agent dialogue, cross-agent communication, feedback ingestion, and adaptive preference updating. This tier closes the loop between perception, reasoning, and action, enabling continuous improvement and long-term alignment with user goals.

Together, these tiers form a modular, interoperable architecture that can be adopted by retailers, marketplaces, and technology providers to ensure predictable, explainable, and scalable agent behavior. The framework supports progressive capability levels, enabling retailers to evolve from informational agents to fully autonomous purchasing systems. By establishing a shared architectural foundation, this work aims to accelerate the development of agentic commerce and foster a more efficient, transparent, and user-aligned retail ecosystem.

KEYWORDS: RETAIL AI AGENTS, TIERED FUNCTIONAL ARCHITECTURE, AGENTIC COMMERCE, INTEROPERABILITY FRAMEWORKS

1. Introduction

This section introduces the concept of retail AI agents, positioning them as emerging autonomous systems that mediate interactions between consumers, retailers, and digital commerce infrastructures. It outlines the motivations driving their development, the limitations of current retail technologies, and the need for a unified architectural framework capable of supporting scalable, interoperable, and trustworthy agent behavior. The section also situates retail AI agents within broader trends in agentic commerce and multi-agent systems, establishing the conceptual foundations for the tiered functional decomposition proposed in this paper. By clarifying the problem space and the research gap, it prepares the ground for the architectural model that follows.

Retail is undergoing a profound transformation as artificial intelligence shifts from passive recommendation engines to **autonomous agents capable of perceiving, reasoning, and acting within complex commercial environments** [1]. These retail AI agents represent a new class of intelligent systems designed to interpret user intent, evaluate product options, negotiate with retailer or marketplace agents, and execute transactions with minimal human intervention [2]. Their emergence reflects broader advances in agentic AI, multi-agent coordination, and preference-aligned decision systems, all converging to redefine how value is created and exchanged in digital commerce [3].

Despite rapid progress, the deployment of retail AI agents remains fragmented. Existing systems are typically proprietary, narrowly scoped, and optimized for isolated tasks such as product ranking, conversational assistance, or dynamic pricing [4]. They lack a **shared architectural foundation** that defines how agents should ingest information, structure reasoning, and produce actions in a transparent and interoperable manner. This absence of standardization limits scalability, inhibits cross-platform collaboration, and increases the risk of misalignment between user goals, retailer policies, and agent behavior.

The need for a unified framework is amplified by the increasing complexity of retail ecosystems. Modern commerce involves heterogeneous data sources, dynamic pricing mechanisms, intricate logistics networks, and diverse regulatory constraints. Retail AI agents must navigate these environments while maintaining alignment with user preferences, ensuring fairness, and preserving trust. Without a structured architectural model, agents risk becoming opaque, unpredictable, or incompatible across platforms.

This paper addresses this gap by introducing a **tiered functional decomposition** that organizes retail AI agent capabilities into three coherent layers: **Input, Model, and Output**. This decomposition provides conceptual clarity and operational modularity. The **Input Tier** defines how agents acquire and normalize information from users, retailers, and external ecosystems. The **Model Tier** captures the reasoning processes that transform inputs into decisions, including optimization, negotiation, and alignment safeguards. The **Output Tier** governs how agents act on decisions, communicate with stakeholders, and learn from feedback.

By formalizing these tiers, the proposed framework establishes a foundation for **interoperability, explainability, and progressive capability development**. It enables retailers, marketplaces, and technology providers to adopt agentic systems in a structured manner while ensuring consistent behavior across diverse environments. Ultimately, this architecture aims to accelerate the evolution of agentic commerce and support the creation of intelligent, trustworthy, and user-aligned retail ecosystems.

2. Background and related work

This section situates retail AI agents within the broader landscape of agentic commerce, multi-agent systems, and retail AI research. It reviews foundational concepts, highlights limitations of existing architectures, and identifies gaps that motivate the need for a unified tiered framework. By examining prior work in

autonomous agents, retail decision systems, and interoperability standards, the section establishes the conceptual and technical context for the proposed three-tier functional architecture.

2.2. Formal definitions

We start with the foundational terminology necessary for understanding the conceptual and technical framing of agent-driven retail systems. A retail AI agent can be understood as an autonomous computational system designed to interpret user intent, analyze retail-specific information, and carry out commercial actions in a manner that remains aligned with both user preferences and the operational policies of the retailer [2]. Such an agent does more than simply process commands. It engages in contextual reasoning, evaluates available options, and selects actions that optimize outcomes within the constraints of the retail environment. This definition emphasizes autonomy, interpretive capability, and policy-aligned execution as core attributes.

Building on this idea, the notion of agentic commerce describes a broader economic paradigm in which autonomous agents mediate interactions, negotiations, and transactions on behalf of individuals or organizations [5]. In this paradigm, agents are not passive intermediaries but active participants capable of evaluating offers, comparing alternatives, and making decisions that reflect the goals of the entity they represent. Agentic commerce therefore shifts traditional commercial processes toward a model where software agents assume responsibilities that previously required direct human involvement, enabling more efficient, scalable, and personalized economic exchanges.

To support such sophisticated behavior, functional decomposition provides a methodological framework for structuring an agent's internal architecture. Functional decomposition involves dividing the agent's capabilities into modular layers that distinguish between information intake, reasoning processes, and action execution [6]. By separating these components, designers can create systems that are more interpretable, maintainable, and adaptable. The intake layer focuses on perception and data acquisition, the reasoning layer handles inference and decision-making, and the execution layer translates decisions into concrete actions. This structured approach ensures that complex agent behaviors emerge from well-defined, interoperable modules rather than monolithic designs, ultimately enabling more robust and scalable retail AI systems.

2.2. Related work in multi-agent systems

Multi-agent system research offers a rich foundation for understanding how autonomous entities can coordinate, negotiate, and make decisions within distributed environments. Over several decades, this field has produced a variety of theoretical models and interaction protocols that describe how agents form beliefs, pursue goals, and resolve conflicts in shared computational spaces. Prominent among these are BDI (Belief-Desire-Intention) architectures [7], which formalize how agents represent mental states and translate them into rational action. Also applicable are contract-net protocols [8], which provide structured mechanisms for task allocation and cooperative problem-solving. Finally, widely used, for intermodal behaviors, we see the utility-based agent models [9], which frame decision-making as the optimization of quantifiable preferences or outcomes. These frameworks supply essential conceptual tools for reasoning about autonomy, coordination, and strategic behavior.

However, despite their generality and expressive power, these classical models do not directly confront the domain-specific challenges inherent to modern retail ecosystems. Retail environments introduce constraints that extend beyond abstract coordination problems, including extreme catalog heterogeneity, rapidly changing product attributes, and complex fulfillment logistics that span inventory management, delivery routing, and real-time availability [4]. Additionally, retail agents must operate

within strict regulatory and compliance boundaries, such as consumer protection rules, data-handling requirements, and marketplace fairness policies. These factors impose practical limitations and operational nuances that traditional multi-agent frameworks were not designed to capture. As a result, while multi-agent system research provides valuable building blocks, it requires significant adaptation and augmentation to support the demands of agentic commerce in large-scale, real-world retail settings.

2.3. Retail AI and decision systems

Retail AI has historically centered on predictive modeling techniques designed to optimize specific operational tasks such as product recommendations, demand forecasting, and dynamic pricing [10]. These systems have delivered substantial commercial value, but they are typically engineered as narrowly scoped, monolithic pipelines that excel at a single function while lacking the adaptability required for broader autonomous behavior [11]. Their architectures are often inflexible, with tightly coupled components that make it difficult to extend capabilities, incorporate new data modalities, or support interactive decision-making. As a result, traditional retail AI solutions are effective within predefined boundaries but are not well suited for the open-ended reasoning and multi-step action planning expected of autonomous agents.

The emergence of LLMs has introduced a new layer of flexibility, enabling systems that can interpret natural language, generalize across tasks, and perform higher-order reasoning. However, despite their impressive capabilities, LLMs alone do not provide the structural discipline needed for reliable agentic behavior. Without a well-defined architectural framework, their decision processes remain opaque, inconsistent, and difficult to audit or standardize across retail workflows. This gap highlights the need for modular, interpretable agent architectures that can harness LLM reasoning while ensuring predictable, policy-aligned execution in complex retail environments.

2.5. Interoperability standards

Efforts such as schema.org, GS1 standards [12], and OpenAPI specifications [13] have contributed important building blocks for interoperability across digital retail ecosystems. These initiatives establish shared vocabularies, product identifiers, and interface descriptions that allow systems to exchange structured information with greater consistency and reliability. They help reduce fragmentation in how product attributes are represented, how services expose their capabilities, and how platforms communicate basic operational details. Despite their value, these standards remain focused on data formats and service schemas rather than on the behavioral logic required for autonomous agents. None of them articulate how an agent should reason about goals, evaluate constraints, or select actions within a retail workflow. They also do not specify protocols for multi-step decision-making or coordinated execution. This absence creates a critical gap that the proposed three-tier architecture seeks to address by defining the behavioral and procedural foundations necessary for agentic retail systems.

2.5. Conceptual diagram: positioning retail AI agents

The schema (see figure 1) situates retail AI agents within the continuum of retail technologies, illustrating how they extend beyond traditional systems such as search engines, recommendation models, and pricing algorithms. While existing tools provide valuable but isolated capabilities, they lack autonomy, structured reasoning, and cross-system interoperability. Retail AI agents occupy a higher conceptual layer, integrating perception, decision-making, and action within a unified architecture.

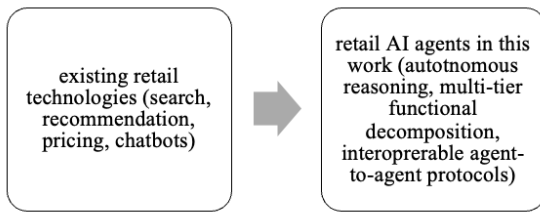


Fig. 1 Retail AI agents within the continuum of retail technologies.

This section emphasizes that agents are not incremental improvements but a categorical shift toward systems capable of negotiating, coordinating, and transacting on behalf of users. By positioning agents above legacy components, the diagram highlights the need for a tiered functional framework that can support their expanded role in agentic commerce.

3. Conceptual foundations

The conceptual foundations of the proposed framework establish the principles, requirements, and design goals that guide the development of retail AI agents. As retail ecosystems evolve toward agent-mediated interactions, it becomes essential to articulate the underlying assumptions and constraints that shape agent behavior. This section defines the theoretical basis for tiered functional decomposition, clarifies the operational environment in which retail AI agents must operate, and identifies the architectural qualities necessary for scalable, trustworthy, and interoperable agentic systems.

Retail AI agents operate at the intersection of user intent, commercial data, and autonomous decision-making. Their design must therefore reflect three core principles. First, **modularity**, which ensures that perception, reasoning, and action remain separable yet interoperable components. Modularity enables independent evolution of capabilities, simplifies debugging, and supports heterogeneous implementations across retailers. Second, **alignment**, which requires agents to act consistently with user preferences, retailer policies, and regulatory constraints. Alignment mechanisms must be embedded throughout the architecture, ensuring that autonomy does not compromise trust or compliance. Third, **interoperability**, which allows agents to communicate, negotiate, and coordinate across platforms, retailers, and service providers. Without interoperability, agentic commerce risks becoming fragmented and inefficient.

Dynamic pricing, variable inventory, complex logistics, and diverse product taxonomies characterize the operational environment of retail AI agents. Agents must therefore handle uncertainty, incomplete information, and conflicting objectives. This necessitates a structured approach to functional decomposition, where each tier (Input, Model, and Output) addresses a distinct aspect of the agent's lifecycle.

4. Three-tier functional architecture

This section introduces the central contribution of the paper: a unified three- tier functional architecture designed specifically for retail AI agents (see figure 2). The framework articulates the structure, responsibilities, and coordinated interactions of the Input, Model, and Output tiers, establishing a clear separation of concerns that supports modularity and interpretability. The Input tier governs data acquisition and intent interpretation, the Model tier handles reasoning and decision- making, and the Output tier manages action execution across retail systems. Together, these components form a coherent architectural blueprint, illustrated through formal descriptions and conceptual diagrams that clarify their roles within agentic retail environments.

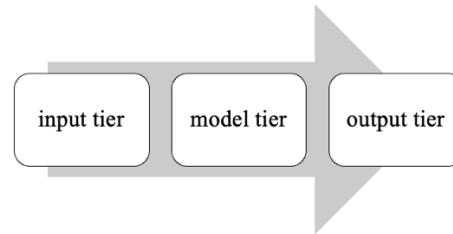


Fig. 2 A unified three- tier functional architecture designed for retail AI agents.

4.1. Input tier: perception and context acquisition

The **input tier** is responsible for acquiring, normalizing, and structuring all information required for downstream reasoning. It integrates signals from users, retailers, and external ecosystems, ensuring that the agent operates on coherent and semantically aligned data.

The key components of the input tier include:

- user intent and preferences: captures goals, constraints, and long-term behavioral patterns that guide decision-making.
- product catalog and attributes: normalizes retailer catalog data into a consistent schema for comparison and reasoning.
- pricing, promotions and loyalty: incorporates dynamic prices, discounts, and user-specific benefits.
- inventory and logistics data: provides stock levels, delivery windows, and fulfillment constraints.
- retailer policies: includes return rules, payment options, and compliance requirements.
- external ecosystem data: integrates reviews, sustainability scores, competitor prices, and third-party metadata.

The diagram (see figure 3) illustrates how diverse data sources flow into the input tier and are transformed through a normalization layer into a unified retail data model.

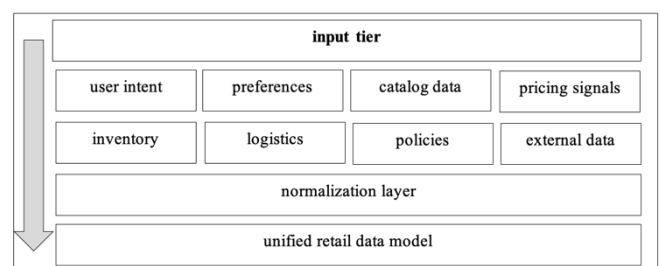


Fig. 3 Diverse data sources flow into the input tier and are transformed through a normalization layer into a unified retail data model.

4.2. Model tier: reasoning, optimization, and decisioning

The **model tier** forms the cognitive core of the retail AI agent. It transforms normalized inputs into actionable decisions through structured reasoning, optimization, and alignment.

The key components of the model tier include:

- goal decomposition and task planning: breaks user intent into actionable sub-tasks and defines the reasoning workflow.

- relevance scoring and ranking: evaluates product suitability based on attributes, preferences, and contextual signals.
- multi-objective optimization: balances competing factors such as price, quality, delivery time, and sustainability.
- negotiation and offer evaluation: assesses bundles, promotions, and dynamic pricing opportunities.
- risk and trustworthiness assessment: identifies unreliable sellers, low-quality products, or fulfillment risks.
- alignment and safety constraints: ensures decisions remain consistent with user preferences and retailer policies.
- explainability and traceability: generates interpretable reasoning paths for transparency and auditability.

The diagram (see figure 4) shows how the model tier integrates multiple reasoning modules (ranking, optimization, negotiation, and alignment) into a unified decision synthesis process. The output is a structured set of recommended or selected actions, ready for execution by the output tier.

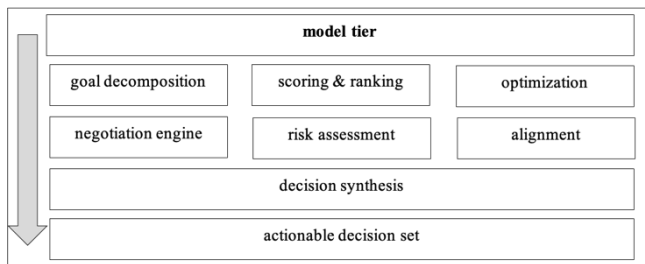


Fig. 4 The model tier integrates multiple reasoning modules into a unified decision synthesis process.

4.3. Output tier: action execution and adaptive learning

The **output tier** operationalizes decisions and closes the perception-reasoning-action loop. It executes commercial actions, communicates with users and other agents, and integrates feedback to refine future behavior.

The key components of the model tier include:

- action execution: performs tasks such as adding items to cart, purchasing, reserving, or scheduling delivery.
- transparent justification: provides human-readable explanations for decisions and trade-offs.
- cross-agent communication: interacts with retailer, logistics, and payment agents to complete workflows.
- feedback ingestion: captures user satisfaction, corrections, and post-purchase signals.
- adaptive preference updating: refines user models based on behavior and feedback loops.
- lifecycle automation: supports replenishment, warranty reminders, and subscription management.

The diagram (see figure 5) highlights how the output tier not only executes actions but also feeds insights back into the agent's preference model. This creates a continuous learning loop, enabling the agent to improve alignment and performance over time.

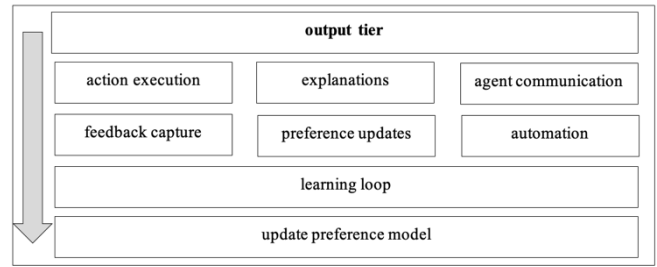


Fig. 5 The output tier not only executes actions but also feeds insights back into the agent's preference model.

The three- tier functional architecture outlined in this section establishes a clear and modular foundation for retail- focused agentic systems. By separating perception, reasoning, and action into distinct yet interoperable tiers, the framework supports scalable development, transparent behavior, and seamless integration across heterogeneous retail environments.

Conclusions

The transition toward agentic commerce marks one of the most significant paradigm shifts in the history of retail technology. As AI agents increasingly mediate interactions between consumers, retailers, and logistics networks, the need for a coherent architectural foundation becomes critical. This paper has introduced a unified framework based on tiered functional decomposition, offering a structured approach to designing, evaluating, and standardizing retail AI agents.

By decomposing agent functionality into input, model, and output tiers, the framework provides clarity, modularity, and interoperability. The input tier ensures that agents operate on consistent, semantically aligned data, enabling fair comparisons and reliable decision- making across heterogeneous retail environments. The model tier establishes a principled foundation for reasoning, optimization, and alignment, ensuring that agents behave as rational, preference- aligned economic actors. The output tier operationalizes decisions, enabling transparent action execution and continuous learning through feedback loops.

This architecture not only supports current retail use cases (such as product discovery, personalized recommendations, and dynamic pricing) but also anticipates future capabilities, including autonomous negotiation, multi- agent collaboration, and fully automated purchasing workflows. By defining clear interfaces and responsibilities for each tier, the framework enables retailers to adopt agentic technologies progressively, reducing integration complexity while maintaining alignment with user expectations and regulatory requirements.

Moreover, the proposed architecture lays the groundwork for industry-wide standards. As retailers, marketplaces, and technology providers converge on shared protocols and functional definitions, the ecosystem can evolve toward greater transparency, efficiency, and user empowerment. The framework thus serves as both a conceptual foundation and a practical blueprint for the next generation of retail AI systems.

Ultimately, this work aims to catalyze a shift toward more interoperable, trustworthy, and intelligent retail ecosystems, where autonomous agents enhance decision- making, streamline operations, and deliver meaningful value to consumers and businesses alike.

References

1. Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey. IEEe Access. Accessed Jan 2026, at: <https://ieeexplore.ieee.org/abstract/document/10849561>
2. Mishra, L. N., & Senapati, B. (2025). Retail Resilience Engine: An Agentic AI Framework for Building Reliable Retail

- Systems With Test-Driven Development Approach. IEEE Access. Accessed Jan 2026, at: <https://ieeexplore.ieee.org/abstract/document/10930951>
3. Ntumba, C., Aguayo, S., & Maina, K. (2023). Revolutionizing retail: a mini review of e-commerce evolution. *Journal of Digital Marketing and Communication*, 3(2), 100-110. Accessed Jan 2026, at: <https://tecnoscientifica.com/journal/jdmc/article/view/365/184>
 4. Hunt, W., & Rolf, S. (2022). Artificial intelligence and automation in retail. Friedrich Ebert Stiftung. Accessed Jan 2026, at: https://uniglobalunion.org/wp-content/uploads/Artificial-Intelligence-and-Automation-in-Retail_EN.pdf
 5. Allouah, A., Besbes, O., Figueroa, J., Kanoria, Y., & Kumar, A. (2025). What is your ai agent buying? evaluation, implications, and emerging questions for agentic e-commerce. *Evaluation, Implications, and Emerging Questions for Agentic E-Commerce* (August 04, 2025).
 6. Hernandez, I., Watson, B. C., Weissburg, M. J., & Bras, B. (2024). Using functional decomposition to bridge the design gap between desired emergent multi- agent- system resilience and individual agent design. *Systems Engineering*, 27(5), 911-930.
 7. Agiollo, A., & Omicini, A. (2025). Integrating Machine Learning into Belief-Desire-Intention Agents: Current Advances and Open Challenges. *arXiv preprint arXiv:2510.20641*.
 8. Kang, J., Ren, S., & Wang, C. (2022, August). A Situated Contract Net Protocol Realization in Command and Control. In *Proceedings of the 5th International Conference on Information Science and Systems* (pp. 49-56).
 9. Diallo, A. O., Lozenguez, G., Doniec, A., & Mandiau, R. (2025). Utility-based agent model for intermodal behaviors: a case study for urban toll in Lille. *Applied Intelligence*, 55(4), 282.
 10. Muthukalyani, A. R. (2023). Unlocking accurate demand forecasting in retail supply chains with AI-driven predictive analytics. *Information Technology and Management*, 14(2), 48-57.
 11. Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025). Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges. *arXiv preprint arXiv:2505.10468*.
 12. Alhava, O., Arola, T., Torro, O., Järvenpää, M., Järvinen, T., & Ruottinen, B. (2025). AI-Agent Application for Semantic Data Enrichment in Ventilation Systems Using National Nomenclature for IFC and GS1-Based Product Information.
 13. Ocansey, J. T. (2024, September). Enhanced Interoperability and Consistency in Heterogeneous Systems with CorrLang and OpenAPI. In *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems* (pp. 200-203).

Artificial intelligence in auditing and compliance in the telecommunications industry: a comparative empirical study of global operators

Ogden Firfov

University American College Skopje, North Macedonia; e-mail: ogden.firfov@uacs.edu.mk

Abstract: Telecommunications operators usually operate in highly competitive and highly regulated, data-intensive environments characterized by complex technological IT and NT infrastructure, high level of transaction volumes, and cross-border regulatory obligations. Traditional audit and compliance models — usually and largely based on sample-driven reviews in different time periods — are increasingly insufficient to provide timely and overall assurance. This article analyses the application of Artificial Intelligence (AI) tools in auditing and compliance within six global telecom operators and discusses adoption models, outcomes, and governance challenges.

KEYWORDS: INTERNAL AUDIT; ARTIFICIAL INTELLIGENCE; AI GOVERNANCE; COMPLIANCE; TELECOMMUNICATIONS; CONTINUOUS AUDITING

1. Introduction

The telecommunications industry is (aside from the banking industry) among the most heavily regulated and technologically complex sectors of the global business and economy. Telco operators manage enormous volumes of commercial transactions, customer data, financial transactions and network events while complying with sector-specific regulation, financial reporting standards, cybersecurity obligations, and data protection regimes such as the General Data Protection Regulation (GDPR). As digitalization accelerates through 5G (and expected 6G in mid-term future), cloud computing, and Internet of Things (IoT) deployments, the scale and complexity of risks confronting telecom operators continue to grow [1]; [2].

Previous research shows that traditional audit and compliance approaches—typically retrospective and sample-based—are increasingly insufficient to address the pace and complexity of digital developments in telecommunications [3]. In response and completely logically, organizations are increasingly exploring Artificial Intelligence (AI) as a means to transform auditing and compliance into continuous, data-driven functions capable of detecting anomalies, predicting risks, supporting proposed improvements and supporting proactive decision-making [4].

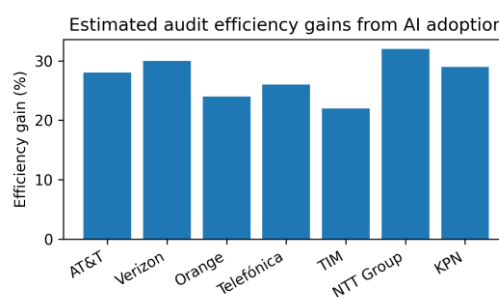
2. Literature review and theoretical background

Different authors have an opinion that the auditing landscape is changing with AI, to continuous auditing and complete data populations. The application of AI in auditing has been framed as a transition from conventional assurance models toward continuous auditing and continuous controls monitoring according to [5]. Machine learning enables auditors to analyse complete data populations rather than relying on samples, thereby improving audit coverage and detection capability according to [6]; [7]. Process mining further enhances audit transparency by reconstructing real business processes from event logs, revealing deviations from expected controls [8]; [9]; [10]

Professional bodies acknowledge the transformative potential of AI. The International Auditing and Assurance Standards Board emphasizes that advanced analytics and automation can enhance audit quality, provided that professional judgment and accountability are preserved [4]. Rikhardsson and Yigitbasoglu highlight AI's role in strengthening audit assurance in complex, automated environments [11], while cautioning against overreliance on opaque models and underscoring the importance of explainability and human oversight [12]. It is interesting to note that beside enhanced audit quality, there is also expectation of

increasing audit efficiency per different telco operators due to AI adoption as shown on Figure 1.

Fig. 1 Estimated audit productivity improvements from AI adoption (AT&T, Verizon, Orange, Telefónica, TIM, NTT Group, KPN)



Explications: Estimated reductions in manual audit work achieved through AI-enabled auditing across the 6 analysed operators.

Beyond audit, artificial intelligence is increasingly deployed in compliance and risk management to monitor regulatory obligations, detect policy breaches, and automate reporting processes [13]. Natural Language Processing (NLP) techniques enable the analysis of large volumes of unstructured textual data, supporting policy and regulatory analysis according to [14]. However, governance and explainability are essential in regulated industries [15]; [16]; [12].

Telecommunications operators are usually a critical environment due to regulatory exposure, complex IT infrastructures, and sensitivity of customer data.

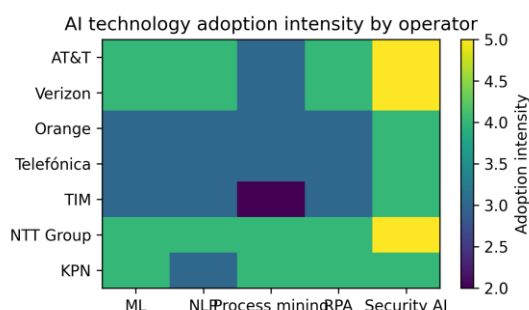
According to European Agency for Cyber Security [17], the deployment of AI in telecommunications can strengthen cybersecurity and fraud detection mechanisms, but simultaneously expands the attack surface and introduces new governance challenges, which aligns with the risk and governance considerations outlined in the NIST AI Risk Management Framework [1].

3. Research methodology

This study adopts a qualitative comparative case study methodology suitable for analysing complex socio-technical systems across organizations [18]. Six multinational telecommunications operators with different initial origins were selected: AT&T and Verizon, Orange, Telefónica, Telecom Italia (TIM), NTT Group, and KPN.

Data were collected from publicly available sources including annual reports, sustainability and governance disclosures, regulatory filings, and authoritative industry studies. A structured empirical dataset was developed covering: (1) AI technologies employed; (2) implementation and governance frameworks; and (3) observed outcomes [16].

Fig. 2 AI technology adoption intensity by operator (including AT&T and Verizon)



Explications: Heatmap of adoption intensity across ML, NLP, process mining, RPA, and security AI (coded low–high) [10].

4. Solution of the examined problem

In order to propose a viable solution, we tried to synthesize the comparative evidence and translate it into an actionable assurance model for large operators. Across the firms which were subject of this paper, the standard pattern is the move from periodic sampling toward population-level testing and continuous monitoring. In telecom, this shift is particularly valuable because critical risk signals (e.g., revenue leakage, roaming anomalies, privileged access misuse, data leakage suspicions) are often rare events embedded in high-volume streams. Machine-learning anomaly detection can detect and surface these outliers earlier than traditional controls testing, provided that data quality and model governance are addressed [4].

From a process perspective, AI-enabled assurance usually follows a three-layer architecture. First, data ingestion consolidates financial, operational, and security telemetry from billing, CRM, network OSS/BSS, identity, and SIEM platforms. Second, analytics and automation implement tests such as: (i) statistical and ML anomaly detection on transactions; (ii) process mining to reconstruct the 'as-is' control flow and detect deviations from the intended process; and (iii) intelligent automation to execute repetitive compliance checks and evidence collection. These approaches are consistent with contemporary continuous auditing models, in which controls are assessed through ongoing evidence streams rather than ad-hoc sampling [5]; [9].

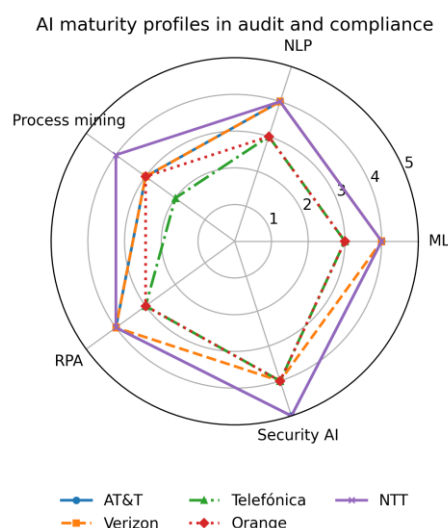
Third, an assurance governance layer operationalises 'human-in-the-loop' decision making. Practically, this means auditors and compliance staff review exceptions flagged by models, validate and confirm root causes, and decide remediation priorities. This is essential because explainability and accountability remain core requirements in regulated environments, especially when automated systems influence compliance decisions [15]; [1].

In comparative terms, the study highlights a clear EU–US contrast, which is somehow expected. EU operators tend to emphasise transparency and regulator-facing traceability, driven by data-

protection enforcement and cross-border supervisory expectations. US operators AT&T and Verizon, operating at extreme scale on a very big single market, typically prioritise operational resilience and security analytics integration, where AI is embedded in monitoring processes and incident response. However, the same governance principles are valid: model controls, audit trails, and clear ownership are required to avoid 'automation bias' and to ensure that AI outputs remain auditable evidence rather than opaque assertions [19].

Figures 1–3 provide a compact representation of these findings. Figure 1 covers outcome-level impact (productivity improvements) which was slightly commented before in section 1. Figure 2 maps technology breadth coverage and general adoption intensity. Figure 3 demonstrates maturity patterns across key AI dimensions in the analysed operators. This three figures together support the conclusion that maturity is not only a function of tools deployed, but of integration into audit workflows and processes and governance controls that make AI outputs reliable for assurance.

Fig. 3 AI maturity profiles in audit and compliance



Explications: Radar of maturity patterns across Key AI dimensions in the analysed telco companies

It is worth explaining that Figure 3 complements Figures 1 and 2 by illustrating the relative maturity of key AI capabilities across AT&T, Verizon, Telefónica, Orange, and NTT. While efficiency gains (Figure. 1) and adoption breadth (Figure. 2) indicate quantitative and measurable progress, the radar representation highlights qualitative differences in how deeply AI is embedded in audit operations and compliance processes. Operators with higher and more balanced profiles across machine learning, process mining, and security-focused AI can demonstrate greater readiness for continuous auditing and real-time compliance monitoring, provided that governance and explainability controls are in place [5].

5. Results and discussion

In order to interpret the empirical patterns and derive implications for audit and compliance operating models, we came to five conclusions. First and foremost, AI adoption changes the unit of assurance: instead of selecting samples, teams can test entire

populations for defined control assertions (completeness, accuracy, authorization, timeliness). This is particularly relevant for telecom billing, interconnect settlements, roaming traffic settlements, and high-frequency access events (like customer related data issues or transactions). The practical consequence is a reallocation of audit effort—from manual testing to exception investigation, root-cause analysis, and remediation validation [4].

Second and also very important, AI makes stronger the connection between compliance monitoring and security operations. Telecom compliance risks increasingly manifest through cyber issues (compromise of user accounts and passwords, insider misuse, misconfiguration, third-party access). Therefore, audit functions should benefit from integrating with security monitoring and adopting assurance views over SOC (Security Operations Center) processes. Sector guidance emphasises that AI in cybersecurity can improve detection but also introduces new risks (model inversion, data poisoning, automation of attacks), which must be reflected in assurance plans [17]; [20].

Third, the results reinforce that explainability is not a 'nice to have' thing, but it is more and more a pure compliance requirement. When AI supports decisions that affect customers, employees, or regulated reporting, organizations must demonstrate traceability: what data and when they were used, how the model was validated, and how decisions were reviewed and cross checked. Regulatory and policy guidance for trustworthy AI converges on similar control themes: governance, transparency, accountability, and robustness [15]; [1].

Fourth, the study suggests a maturity roadmap for telecom audit functions. Stage 1 focuses on descriptive analytics and automated evidence capture (intelligent RPA). Stage 2 adds anomaly detection and process mining to detect control breakdowns early. Stage 3 integrates predictive risk scoring for audit planning and continuous controls monitoring at scale. Progression toward effective AI-enabled assurance requires sustained investment in data engineering and robust model risk management, rather than a narrow focus on analytics tools alone [21, 22].

Finally, the comparative outcomes should be interpreted quite cautiously. Productivity improvements (Fig. 1) can be achieved quickly in evidence-heavy audits with huge populations, but sustained value depends on how exceptions are handled (or if they are handled at all). If alert volumes are high and false positives are not controlled, AI can shift effort without improving assurance and not much positive gains at the end. Accordingly, operators should track not only efficiency metrics but also quality metrics: precision/recall of alerts, time-to-remediate, recurrence rates, and control failure severity.

6. Conclusion

AI-enabled assurance is becoming a necessity in telecommunications (not only in telecommunications but also in other areas of businesses), but real value is realised only when IT systems & technology, governance of processes, workflows and audit and compliance staff skills evolve together. The evidence indicates that technology-enabled continuous auditing can shorten audit cycles, improve anomaly detection, and strengthen compliance responsiveness, particularly when continuous monitoring is integrated into the audit lifecycle [5].

From a corporate governance standpoint, telecom operators should embed and institutionalise a model risk management approach: documented model purpose, training data lineage, validation results, periodic drift monitoring, and clear accountability for overrides. These controls align with emerging international expectations for AI risk management [1].

Future research should mostly focus on the three gaps. (1) Quantitative measurement: longitudinal datasets linking AI adoption to audit outcomes (cycle time, findings the intensity and severity, remediation speed) and to compliance outcomes (incident rates, regulatory findings). (2) Regulatory acceptance: empirical study of how regulators in the telecom environments evaluate AI-generated evidence and what assurance proofs increase trust. (3) Human-AI collaboration: the organizational structure of audit teams, skill profiles of auditors, and training models needed to operationalise AI while preserving professional judgment.

In practical sense, a telecom-ready research design could be a mix: process mining of billing and access-control workflows; supervised and assisted learning for labelled control failures; and qualitative validation through audit committee and regulator interviews afterwards. Such a proposed design would move beyond descriptive adoption claims and provide causal evidence for what works, under which standard and non standard conditions, and at what cost for the telecom companies.

References

- [1] NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0), Gaithersburg, NIST, 2023.
- [2] Voigt, P. The EU General Data Protection Regulation (GDPR): A Practical Guide, Springer International Publishing, 2017 (Voigt, P., A. Von dem Bussche).
- [3] Vasarhelyi, M.A. Big data in accounting: An overview – Accounting Horizons, vol. 36, 2022, p. 1–22 (Vasarhelyi, M.A., A. Kogan, B.M. Tuttle).
- [4] IAASB The Use of Technology in the Audit of Financial Statements, New York, IFAC, 2020.
- [5] Alles M.G. Drivers of the use and facilitators and obstacles of the evolution of continuous auditing – Accounting Horizons, vol. 29, 2015, p. 439–448.
- [6] Issa, H. Research ideas for artificial intelligence in auditing – Journal of Emerging Technologies in Accounting, vol. 18, 2021, p. 1–20 (Issa, H., T. Sun, M.A. Vasarhelyi).
- [7] Munoko, I. Innovating in Times of Change: The Place for Artificial Intelligence in Auditing – ISACA Journal, vol. 2, 2021.
- [8] van der Aalst, W.M.P. Process Mining: Data Science in Action, Berlin, Springer, 2016.
- [9] Vasarhelyi, M.A. Big data in accounting: An overview – Accounting Horizons, vol. 29, 2015, p. 381–396.
- [10] van der Aalst, W.M.P. Process Mining Put into Context – IEEE Internet Computing, vol. 16, 2012, p. 82–86 (van der Aalst, W.M.P., S. Dustdar).
- [11] Rikhardsson, P. Business intelligence and analytics in management accounting and control – International Journal of Accounting Information Systems, vol. 48, 2023, p. 100598 (Rikhardsson, P., O. Yigitbasioglu).
- [12] Gunning, D. XAI: Explainable artificial intelligence – Science Robotics, vol. 4, 2019, p. 37 (Gunning, D., M. Stefik, J. Choi, T. Miller, S. Stumpf, G-Z. Yang).
- [13] Bai, C. Industry 4.0 technologies assessment – International Journal of Production Economics, vol. 247, 2022, p. 108479 (Bai, C., P. Dallasega, G. Orzes, J. Sarkis).
- [14] Jadeau, J. Enhancing the Availability of Quality Data for Policymaking Using AI – PARIS21 Discussion Paper 19, 2025 (Jadeau, J., M. Fogarassy).

-
- [15] European Commission Ethics Guidelines for Trustworthy AI, Brussels, European Commission, 2019.
- [16] World Economic Forum The Global Risks Report 2021, Cologne, World Economic Forum - WEF, 2021.
- [17] ENISA Threat Landscape – European Union Agency for Cybersecurity, 2023.
- [18] Yin, R.K. Case Study Research and Applications, Thousand Oaks, Sage, 2018.
- [19] Gunning, D. Explainable artificial intelligence – AI Magazine, vol. 40, 2019, p. 44–58
- [20] ISO/IEC Artificial Intelligence – Risk Management, Geneva, ISO, 2023.
- [21] Grover, V. Creating strategic business value from big data analytics – Journal of Management Information Systems, vol. 35, 2018, p. 388–423 (Grover, V., R. Chiang, T.-P. Liang, D. Zhang).
- [22] Ransbotham, S. Reshaping business with artificial intelligence – MIT Sloan Management Review, vol. 62, 2021, p. 1–17 (Ransbotham, S., D. Kiron, P. Gerbert, M. Reeves).

Improving the Living Environment in Evacuation Shelters Using Light-Emitting Diode Lanterns and Partitions

Mana Oobashira¹, Mayumi Tanaka¹, Keiko Ishihara², Satoru Tada³, Masayuki Yamauchi³, Koji Kakugawa³, Takeshi Tanaka³, and Katia Vutova⁴

Hiroshima Prefectural Hiroshima Minami High School, Hiroshima, Japan¹

Hiroshima International University, Kure, Japan²

Hiroshima Institute of Technology, Hiroshima, Japan³

Institute of Electronics, Bulgarian Academy of Sciences, Sofia, Bulgaria⁴

tanaka@cc.it-hiroshima.ac.jp

Abstract: This study examines the effects of light-emitting diode (LED) lanterns and partition colors on the psychological condition of individuals in evacuation shelters. A previous questionnaire survey involving 193 participants evaluated emotional responses to four LED colors (yellow, green, blue, and red). Yellow light was most frequently perceived as calming and comfortable, whereas red light was strongly associated with tension and anxiety. Building on these emotional findings, we investigated environmental factors through an experiment involving blue, pink, and green partitions. The participants stayed inside each partition for 25 min, and their vital signs, including body temperature, pulse rate, and blood pressure, were measured every 5 min. The results indicate that blue partitions produced the most stable physiological responses, including lower systolic blood pressure and steady pulse rates, reflecting parasympathetic nervous system activation. In contrast, the pink and green partitions increased pulse rates and led to greater fluctuations, indicating higher stress. These findings demonstrate that appropriate lighting and partition colors can promote psychological stability and improve shelter environments. This study also highlights the importance of simple, low-cost environmental design strategies for stress reduction during prolonged evacuations.

KEYWORDS: EVACUATION SHELTERS, LIGHT-EMITTING DIODE LANTERN COLOR, PARTITION COLOR, PSYCHOLOGICAL STRESS, AUTONOMIC NERVOUS SYSTEM, ENVIRONMENTAL DESIGN

1. Introduction

As third-year high school students aspiring to pursue careers in nursing, we are concerned about the increasing frequency and severity of natural disasters, particularly earthquakes and heavy rainfall, which have caused substantial damage to communities in recent years. According to the Ministry of Land, Infrastructure, Transport and Tourism's "2020 White Paper on Land, Infrastructure, Transport and Tourism"[1], the probability of a major Nankai Trough earthquake occurring within the next 30 years is estimated to be approximately 80%. This probability has gradually increased since official assessments began in 2013. Given this situation, anticipating the potential occurrence of large-scale disasters and enhancing disaster preparedness and awareness at an early stage are considered essential. To explore practical actions that can be undertaken at present, volunteer activities were conducted in March of this year in disaster-affected regions, including Sendai, Miyagi Prefecture, and the Noto Peninsula in Ishikawa Prefecture. During these activities, evacuation facilities that were actively used were visited and systematically observed.

2. Background

(1) Participation in the Miyagi Fieldwork

Participation in the Sendai Disaster-Prevention Future Forum, held at the Sendai International Center on March 8, 2025, where we provided valuable insight into disaster preparedness. This year marked the 10th anniversary of the Great East Japan Earthquake. To link experiences and lessons derived from past disasters to future prevention efforts, the event featured lectures on disaster prevention, mitigation, and first aid, along with presentations, booth exhibitions, and workshops. During the first aid lecture, participants learned techniques for immobilizing injured areas using magazines, umbrellas, and string, as well as methods for constructing a simple stretcher from a blanket to reduce mortality caused by traumatic shock during disasters. Participation in the "Sendai Framework for Disaster Risk Reduction/5 Years Until the SDGs Target Year" session enabled engagement with five organizational representatives and university students, followed by discussions on disaster prevention and mitigation. These discussions highlighted the importance of preserving disaster memory and the difficulty of increasing awareness among individuals who have not directly experienced disasters.

In addition to these presentations, the event included a tour of Arahama Elementary School, which now serves as a museum dedicated to conveying the magnitude and severity of disasters.

Inside the building, the tsunami had left empty cans lodged in the ceiling of a first-floor classroom used by students, a space where such objects would not normally be trapped. The window frames bent by the tsunami, as shown in Figure 2, further demonstrate the destructive force of the disaster. These observations made disasters feel more immediate and increased the awareness that similar events could occur in our region.



Fig. 1 Arahama Elementary School



Fig. 2 Exterior wall of Arahama Elementary School

(2) Participation in the Oku-Noto Support Tour

Field observations conducted during the Oku-Noto Support Tour from March 28 to 31, 2025, provided direct experience in disaster recovery activities. In the Noto Peninsula, which was severely affected by the earthquake, participation in debris removal and community meal services (Figure 3) revealed that despite a year passing since the disaster, recovery remained stalled due to a shortage of volunteers and the extensive scale of damage (Figure 4). Restoration of lifeline services had also progressed slowly. Through

conversations with affected residents, we learned about the severe living conditions following the disaster, including the absence of heating, lack of electricity, and limited access to drinking water due to water supply disruptions. These conditions were further worsened during the January snowfall. We also stayed overnight in a temporary tent inside a supermarket used as an evacuation shelter. The space lacked heating, and despite the use of sleeping bags and blankets, it remained cold and uncomfortable, while noise from nearby occupants disrupted sleep.

At that time, we realized that prolonged exposure to such conditions would have a significant impact on both physical and psychological health, and we recognized the need to develop strategies to improve the living environment.



Fig. 3 Cooking okonomiyaki for the soup kitchen.



Fig. 4 Removal of soil and debris from inside a house.

(3) Issues Revealed Through Volunteering

Life in evacuation shelters significantly differs from daily life. In addition to major changes in the living environment, disaster survivors experience the loss of loved ones, homes, and employment, leading to anxiety and stress. These conditions are likely to have adverse health effects. Nurses are expected to improve the living environment in shelters and support evacuees in maintaining their health. Based on these observations, we focused on the shelter environment and disaster-prevention supplies to identify strategies for reducing psychological stress. We observed that partitions and tents used in shelters are available in various colors. During the Oku-Noto Support Tour, staying in a tent revealed physical and psychological stress from noise, such as neighboring snoring and conversation, as well as inadequate protection against cold weather. These experiences also revealed that living in a shared environment can be a significant source of stress during evacuation.

Light-emitting diode (LED) lighting color significantly influences psychological and physiological responses, especially in emergency environments such as evacuation shelters. A questionnaire survey involving 193 participants evaluated the effects of four LED colors (yellow, green, blue, and red) using handmade plastic (polyethylene terephthalate) bottle lanterns. The results demonstrated clear differences in emotional responses to light color. Yellow light was most frequently selected as the most calming and comfortable, indicating that warm tones promote relaxation and a sense of security. Green light was also highly rated and associated with nature, freshness, and visual comfort. In contrast, blue light produced mixed reactions; while some

participants reported calmness, others perceived it as cold or tense. The red light was identified as the most stressful, increasing both tension and anxiety. These findings indicate that appropriate LED color selection is essential for creating comfortable lighting environments. For practical application, yellow lighting is recommended for nighttime use, green for orientation and exit guidance, blue for task-oriented areas, and red for emergency warnings. Therefore, careful color design may improve visibility and psychological well-being in disaster situations.

In addition, heartwarming illustrations[2] (Figure 5) have been applied to white LED lanterns to promote emotional comfort among disaster-affected individuals.



Fig. 5 Lantern with a picture painted on the side.

Therefore, this study investigates the psychological effects of evacuation shelter partition colors and identifies those most suitable for promoting psychological stability among evacuees.

3. Research Objective

This study investigates whether different partition colors in evacuation shelters produce varying physical and psychological effects. Based on these results, this study aims to identify partition colors that minimize stress among evacuees.

4. Research Method

(1) Research Content

First, to examine the psychological effects of partition colors, participants stayed inside each colored partition for 25 min, and their vital signs were measured at designated times. Vital signs were assessed to observe changes in autonomic nervous system activity as indicators of psychological responses to partition color. The autonomic nervous system comprises sympathetic and parasympathetic components distributed throughout the body. Dominance of either system affects physiological functions and causes fluctuation of measured values. When the sympathetic nervous system is dominant, catecholamines are released, resulting in increases in body temperature, heart rate, blood pressure, and skin temperature. In contrast, when the parasympathetic nervous system becomes dominant, body temperature, heart rate, blood pressure, and skin temperature decrease. Increased stress is generally associated with an increase in body temperature, pulse rate, and blood pressure, whereas relaxation is associated with a reduction in these parameters[3,4,5].

During the experiment, partitions were placed around the desk to ensure that the selected color remained within the participants' field of view. Participants read inside each partition, and vital signs were measured every 5 min to observe fluctuations over 25 min. The experiment was conducted using three partition colors on separate days. Because partition colors can be broadly classified as warm, cool, and neutral, blue, pink, and green colors were selected. The following conditions were considered during the experiment.

1. To enable comparison, the experiment was conducted with two participants.
2. In evacuation shelters, partitions are used to separate families;

therefore, this situation was simulated in the experiment. Furthermore, given that the dominance of either the parasympathetic or sympathetic nervous system generally affects the body for approximately 20 min, the participants were stationed within the partitions for a 25-minute reading period.

3. Reading was selected because the restoration of lifeline services is often slow, making the use of electronic devices difficult. Therefore, reading was considered a simple and practical form of entertainment in such environments.



Fig. 6 Partition installation



Fig. 7 Experiment in progress

(2) Condition Setting

The experiment was conducted in a classroom to maintain stable temperature, humidity, and airflow. Initially, a gymnasium used as an evacuation shelter was considered. However, due to the mid-July timing and the physical condition of the student researchers, the classroom was selected as it better satisfies these requirements.

(3) Experimental Materials and Equipment

Blue, pink, and green partitions; an electronic blood pressure monitor; a pulse oximeter; a thermometer; a desk; a chair; a tape measure; and books.

5. Results

(1) Results for Participant A

Table 1: Changes in Vital Signs Under the Blue Partition (26.1°C and 50% Humidity)

Curtain Color ●Blue	Body Temperature (°C)	Pulse Rate (beats/min)	Systolic Blood Pressure (mmHg)	Diastolic Blood Pressure (mmHg)
Immediately before the start	35.9	65	99	59
5 min after the start	36.4	66	96	55

10 min after the start	36.6	69	93	51
15 min after the start	36.3	68	99	49
20 min after the start	36.9	67	85	56
25 min after the start	35.9	65	85	56

Participant A's impression: No notable changes were reported from the beginning to the end, and a calm state was maintained throughout.

Table 2: Changes in Vital Signs Under the Pink Partition (26.8°C and 50% Humidity)

Curtain Color ●Pink	Body Temperature (°C)	Pulse Rate (beats/min)	Systolic Blood Pressure (mmHg)	Diastolic Blood Pressure (mmHg)
Immediately before the start	36.6	81	98	67
5 min after the start	36.1	79	98	64
10 min after the start	35.9	74	96	55
15 min after the start	36.3	72	92	59
20 min after the start	36.8	80	94	47
25 min after the start	36.5	83	90	50

Participant A's impression: A calm state was maintained during the first 5–10 min; however, tension increased at the end.

Table 3: Changes in Vital Signs Under the Green Partition (25.1°C and 52% Humidity)

Participant A's impression: There was no particular change from the start to the end, and he remained calm throughout.

Curtain Color ●Green	Body Temperature (°C)	Pulse Rate (beats/min)	Systolic Blood Pressure (mmHg)	Diastolic Blood Pressure (mmHg)
Immediately before the start	36.5	87	96	64
5 min after the start	36.4	76	93	51
10 min after the start	36.1	83	91	62
15 min after the start	37	82	97	60
20 min after the start	36.9	76	94	57
25 min after the start	36.9	91	87	58

(2) Results for Participant B

Table 4: Changes in Vital Signs Under the Blue Partition (26.1°C and 50% Humidity)

Curtain Color ●Blue	Body Temperature (°C)	Pulse Rate (beats/min)	Systolic Blood Pressure (mmHg)	Diastolic Blood Pressure (mmHg)
Immediately before the start	36.3	79	98	60
5 min after the start	36.7	68	104	62
10 min after the start	36.6	69	97	65
15 min after the start	36.7	65	95	62
20 min after the start	36.7	71	102	63
25 min after the start	36.5	79	94	68

Participant B's impression: A calm state was maintained throughout the experiment, with no notable psychological changes reported.

Table 5: Changes in Vital Signs Under the Pink Partition (26.7°C and 50 Humidity)

Curtain Color ●Pink	Body Temperature (°C)	Pulse Rate (beats/min)	Systolic Blood Pressure (mmHg)	Diastolic Blood Pressure (mmHg)
Immediately before the start	36.7	75	115	71
5 min after the start	36.8	74	102	61
10 min after the start	36.8	74	101	64
15 min after the start	36.5	80	109	64
20 min after the start	36.9	74	102	62
25 min after the start	36.6	85	115	67

Participant B's impression: Approximately 15 min after the start, the surrounding noise became bothersome, and a desire for the experiment to end quickly was reported.

Participant B's impression: After 20 min, the surrounding noise became bothersome, and the bell increased anxiety. The concern that the experiment had not ended led to restlessness.

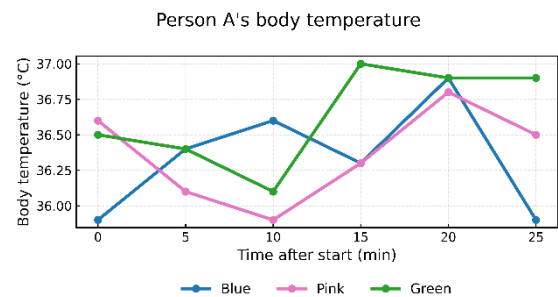
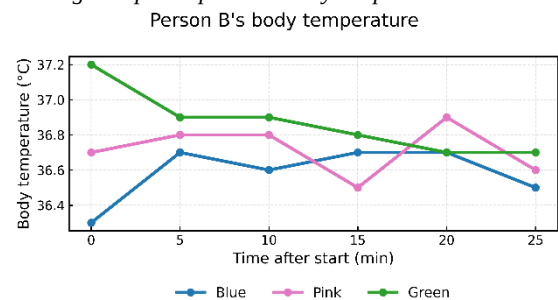
(3) Changes in Body Temperature, Pulse Rate, and Blood Pressure

Table 6: Changes in Vital Signs Under the Green Partition (25.8°C and 53% Humidity)

Curtain Color ●Green	Body Temperature (°C)	Pulse Rate (beats/min)	Systolic Blood Pressure (mmHg)	Diastolic Blood Pressure (mmHg)
Immediately before the start	37.2	79	100	57
5 min after the start	36.9	73	103	66
10 min after the start	36.9	82	108	64
15 min after the start	36.8	79	99	63
20 min after the start	36.7	92	103	60
25 min after the start	36.7	88	113	70

1. Body Temperature

The body temperature of participant A decreased when the partition was pink, whereas that of participant B decreased when the partitions were pink and green. With the other partition colors, body temperatures increased between the beginning and the end of the experiment.

**Fig. 8** Changes in participant A's body temperature.**Fig. 9** Changes in participant B's body temperature.

2. Pulse Rate

Figures 9 and 10 show that the pulse rates of participants A and B experienced minimal changes under the blue partition; however, they increased under the pink or green partitions.

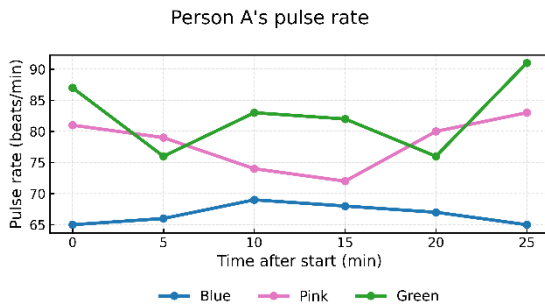


Fig. 10 Changes in participant A's pulse rate.

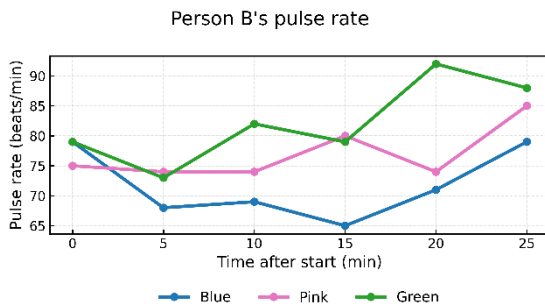


Fig. 11 Changes in participant B's pulse rate.

Figures 10 and 11 show the change in the pulse rate immediately before the start and 25 min after the start.

*Values in parentheses indicate the magnitude of change, with the value immediately before the start set as 0 and compared with the value after 25 min.

Participant A

Blue: 65 beats/min → 64 beats/min (−1 beat/min)
 Pink: 81 beats/min → 83 beats/min (+2 beats/min)
 Green: 87 beats/min → 91 beats/min (+4 beats/min)

Participant B

Blue: 79 beats/min → 79 beats/min (0 beats/min)
 Pink: 75 beats/min → 85 beats/min (+10 beats/min)
 Green: 79 beats/min → 88 beats/min (+9 beats/min)

3. Blood Pressure

Figures 12 and 13 show that the systolic blood pressure of both participants A and B was lowest under the blue partition.

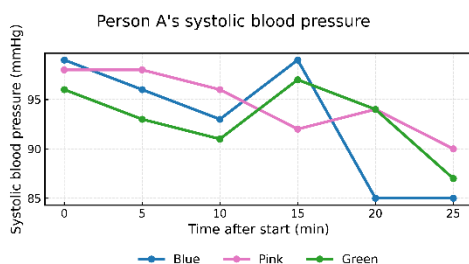


Fig. 12 Changes in participant A's systolic blood pressure

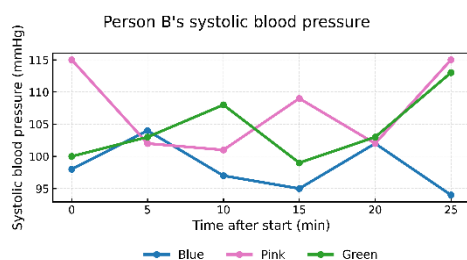


Fig. 13 Changes in Participant B's Systolic Blood Pressure

• Changes in values immediately before and 25 mins after the start are shown in Figures 12 and 13

*Values in parentheses indicate the magnitude of change, with the value immediately before the start set as 0 and compared with the value after 25 min.

Participant A

Blue: 99mmHg → 85mmHg (−14mmHg)
 Pink: 98mmHg → 90mmHg (−8mmHg)
 Green: 96mmHg → 87mmHg (−9mmHg)

Participant B

Blue: 98mmHg → 94mmHg (−4mmHg)
 Pink: 115mmHg → 115mmHg (0)
 Green: 100mmHg → 113mmHg (+13mmHg)

6. Discussion

The autonomic nervous system, a component of the peripheral nervous system that regulates the internal organs, comprises two divisions with opposing functions.

The sympathetic nervous system prepares organs and tissues for increased physical activity and is particularly active during physical exercise, mental arousal, or anxiety. When activated, the heart rate and respiration increase, the bronchi and pupils dilate, and digestion is suppressed. Blood vessels also constrict, leading to increased blood pressure and increased sweating.

In contrast, the parasympathetic nervous system is active during relaxation. Activation reduces heart rate and respiration while constricting the bronchi and pupils. Gastrointestinal motility and digestive secretion are promoted, supporting digestion and absorption.

Therefore, when the sympathetic nervous system is dominant, catecholamines are released, leading to increases in body temperature, heart rate, blood pressure, and skin temperature. In contrast, when the parasympathetic nervous system is dominant, the body temperature, heart rate, blood pressure, and skin temperature decrease.

Based on the body temperature results, both participants A and B showed decreases under the pink partition. Under the blue partition, increases were observed in both participants A and B. Under the green partition, participant A's body temperature increased, whereas that of participant B decreased. Body temperature generally decreases when the parasympathetic nervous system is dominant; thus, the results indicate that parasympathetic activity was enhanced under the pink partition. However, the increase under the blue partition, which has been associated with parasympathetic dominance in other studies, may be explained by air retention within the partition, leading to increased humidity during breathing.

The pulse rate results indicated increases for both participants A and B under the pink and green partitions, indicating sympathetic dominance.

Blood pressure results show that systolic blood pressure for both A and B was lowest under the blue partition, indicating that this condition may most strongly promote parasympathetic activity.

Based on subjective impressions, participant A remained calm throughout the experiment regardless of the partition color, whereas participant B became impatient under the pink and green partitions, resulting in greater physiological fluctuations.

7. Conclusion

This study demonstrates that both LED lighting and partition colors influence the psychological condition of individuals living in evacuation shelters. An evaluation using blue, pink, and green partitions revealed that the blue partition produced the most stable physiological responses, including lower systolic blood pressure and stable pulse rates, indicating parasympathetic activity and a relaxed state. These findings indicate that appropriately selected lighting and partition colors can reduce psychological stress and improve shelter environments. Although the study was conducted on a limited scale, it highlights the potential of simple, low-cost environmental design

strategies to support evacuees' mental well-being. Further studies under realistic shelter conditions are required to establish practical guidelines for disaster preparedness and shelter management.

Acknowledgments

We extend our heartfelt gratitude to the people of Wajima City and Suzu City in Ishikawa Prefecture, SAVE THE HIROSHIMA, the students of the Hygiene and Nursing Department and Specialized Department of Hiroshima Prefectural Hiroshima Minami High School, and the students of Hiroshima Prefectural Miyajima Technical High School for their invaluable cooperation for this study.

We also thank Enago (www.enago.jp) for their English language review.

References

- [1] White Paper on Land, Infrastructure, Transport and Tourism in Japan, 2020. (in Japanese)
- [2] Onuma Hiroaki: 6F Artist, ART BORN (Hiroshima). [https://www.artborn-hiroshima.net/news/artists/%e5%a4%a7%e6%b2%bc-%e5%af%9b%e6%98%8e/\(February 13, 2026\)](https://www.artborn-hiroshima.net/news/artists/%e5%a4%a7%e6%b2%bc-%e5%af%9b%e6%98%8e/(February%2013,%202026).). (in Japanese)
- [3] Hashimoto H, ed. Kango-gaku nyumon 1: Jintai no shikumi to hataraki [Introduction to nursing 1: Structure and function of the human body]. 5th ed. Tokyo: Medical Friend; 2021. p.309. (in Japanese)
- [4] Ninomiya I, Suga H, editors. Cardiovascular Physiology: Function and Pathophysiology. Tokyo: Nankodo Co., Ltd.; 1993.p. 161. (in Japanese)
- [5] Sakai T, Okada T. Kaibo Seirigaku (Anatomy and Physiology). 9th ed. Tokyo: Igaku-Shoin; 2014. (Systematic Nursing Course: Basic Field. Structure and Function of the Human Body; Vol.1), p. 23, 243. (in Japanese)

Using Large Language Models for Emotional Support of Bulgarian Users: A Survey

Melania Berbatova¹

¹Faculty of Mathematics and Informatics - Sofia University "St. Kliment Ohridski", Bulgaria

msberbatova@fmi.uni-sofia.bg

Abstract: The use of large language models (LLMs) for psychological and emotional support (ES) has rapidly evolved, becoming the most widely used application of generative artificial intelligence among consumers by 2025. This paper presents the results of an anonymous survey of 100 Bulgarian users, primarily high school, university, and doctoral students, to explore their attitudes toward and usage of chatbots for emotional support. Findings indicate that approximately one-half of the surveyed population utilizes chatbots for ES, with ChatGPT being the most dominant platform. Users primarily seek support for coping with stress in interpersonal relationships and work or study-related environments. While 71% of users perceive the technology as effective, non-users remain sceptical. Despite the growing adoption, significant concerns persist regarding data security, technology reliability, and the tendency of chatbots to provide excessive affirmation.

Keywords: LARGE LANGUAGE MODELS, EMOTIONAL SUPPORT, BULGARIAN USERS

1. Introduction

The field of using large language models for psychological and emotional help has rapidly evolved in recent months, with a large volume of published scientific and journalistic studies. According to a study published in Harvard Business Review [1], emotional support is the most widely used application of generative artificial intelligence among consumers in 2025. In this paper, we present the results of a survey of 100 Bulgarians on their use of large language models for emotional and psychological support.

1.1. Emotional support

Emotional support (ES) in theory is a communication style which aims at reducing individuals' emotional distress and helping them understand and work through the challenges that they face. Research shows that recipients of sensitive emotional support feel better and cope with problems more effectively and may even be healthier. Contrary, recipients of insensitive emotional support may feel worse than ever, and even suffer from stress-related illness [2].

According to Lehman and Hampill [7], the characteristics of effective emotional support include:

- expressing love, concern, or understanding;
- encouragement;
- listening;
- praising abilities;
- involvement in social activities;
- emotional presence ("being there").

In essence, emotional support, unlike psychotherapy, does not involve long-term problem solving, but rather short-term coping with emotional stress. However, ES can help individuals develop coping mechanisms for stressful situations, thus leading to the development of a more resilient psyche in those receiving support.

1.2. Challenges of traditional methods

Today, traditional therapy faces several challenges. Therapy sessions are expensive and often hard to organize, and in many cultures, there is a social stigma around participating in them [3]. On the other hand, there is a global shortage of mental health workers – a survey conducted by the World Health Organizations (WHO) in 2022 [4] showed that on average, there are 13 mental health workers for average 100,000 people.

Furthermore, events such as pandemics, wars, and natural disasters lead to circumstances in which millions of people simultaneously need emotional and mental support. It is estimated that only the Covid-19 pandemic triggered an additional 53 million cases of depression and 76 million cases of anxiety globally [5]. In such situations it is impossible for everyone affected to get personal support and automatic systems may be of great help in dealing with massive emotional distress.

1.3. Risks and benefits of AI assistants

In the context of emotional support, there are many benefits of using personal AI assistants. First, they can be available anytime, day or night. They are significantly cheaper than traditional therapy, if not completely free. It is also important to note that studies show that some people feel more comfortable confessing their feelings to a robot rather than a person [6].

Despite their indisputable benefits, the use of AI for emotional support poses some significant risks. Some of the most challenging and dangerous risks noted in literature [8] include:

- **Misinformation:** Fluency is frequently conflated with factuality in LLMs, which generate credible answers that may not align with real-life evidence. Hallucinated responses, especially when presented in a confident tone, can dangerously mislead users [9].
- **Persuasion and manipulation:** LLM-based chatbots can generate highly persuasive messages, raising concerns in mental health contexts where users are frequently in a vulnerable state.
- **Over-reliance:** Users and clinicians may overtrust such AI assistants, treating their outputs as undoubtedly truthful and reliable. Fluency and confidence in responses can mislead users into accepting poor advice [9]. Repeated use of chatbots may also lead to emotional dependency and reduced engagement with human care [10].

2. Survey of Bulgarian users

In order to gain a better insight into the attitudes of the Bulgarian population towards such a technology, we conducted an anonymous survey on the usage of chatbots for emotional support among Bulgarian-speaking internet users. We obtained 100 survey answers from respondents from different age groups and sociological and educational backgrounds.

2.1 Demographics

The survey was completed by 100 people online, mainly high school, university and doctoral students from various disciplines, as well as a community of young professionals in Bulgaria. Therefore, while the distribution of the surveyed population is not representative of the overall Bulgarian population, we were able to obtain answers from participants of various ages and educational levels.

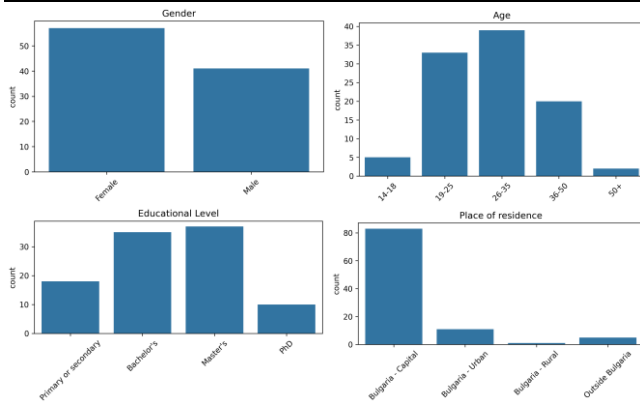


Fig. 1 Demographics of the surveyed population.

2.2 Usage of chatbots for ES

Roughly one-half of the surveyed population (Figure 2) uses chatbots for emotional support. As shown in Figure 3, the most used chatbot for emotional support is, without competition, ChatGPT. The survey reveals only isolated instances of usage of other language models, such as Gemini, Claude, Grok3, and Gemma. When it comes to preferred language for communication, the majority of users state that they mix Bulgarian and English, followed by English-only usage and Bulgarian-only usage.

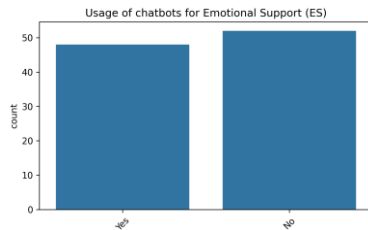


Fig. 2 Usage of chatbots for Emotional Support (ES)

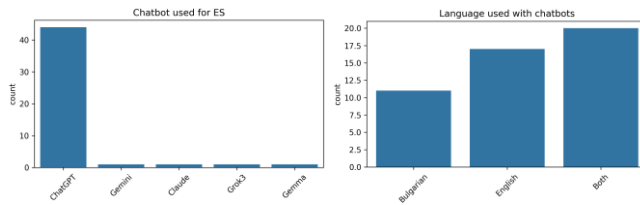


Fig. 3 Characteristics of usage

2.3 Purposes, concerns and perceived effectiveness

The most frequently stated purpose of using chatbots for ES is coping with stress in interpersonal relationships (37 respondents), followed by stress related to education or work (27 respondents). 13 people indicate that they use chatbots to cope with the symptoms of more prolonged mental health issues, such as depression, bipolar disorder, and others. 3 people add other purposes, such as psychoeducation or personal and spiritual development. Figure 4 illustrates the distribution.

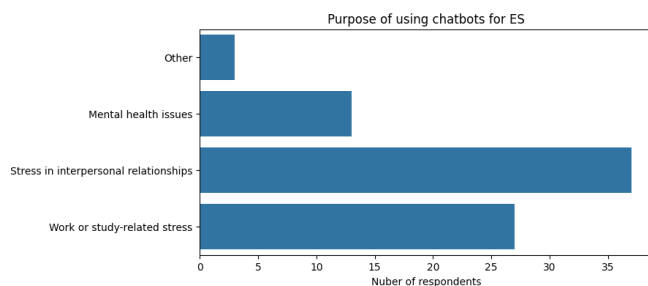


Fig. 4 Purpose of using chatbots for ES

The most frequently stated concerns (Figure 5) are data confidentiality and technology reliability (with 35% of users citing these as their concerns), followed by concerns about long-term negative effects on individuals and interpersonal relationships (21%). Twenty-nine percent of users say they have no concerns about the technology. The most frequently listed reason for concern outside of the predefined possible answers is the tendency of chatbots to always agree and affirm the user's point of view (8%).

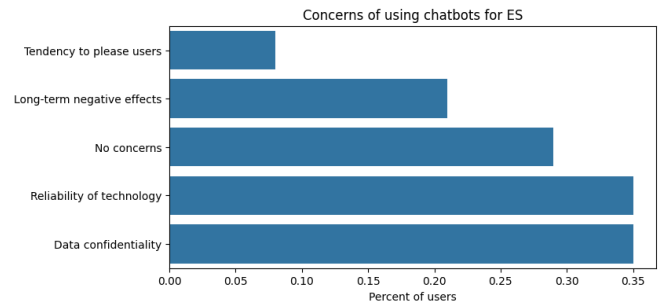


Fig. 5 Concerns of using chatbots for ES

As shown in Figure 6, 71% of users believe that such technology is effective, 19% cannot decide, and 10% do not believe in its effectiveness. By comparison, among respondents who do not use the technology, 55% believe it is not effective, only 6% believe it is effective, and 37% cannot decide.

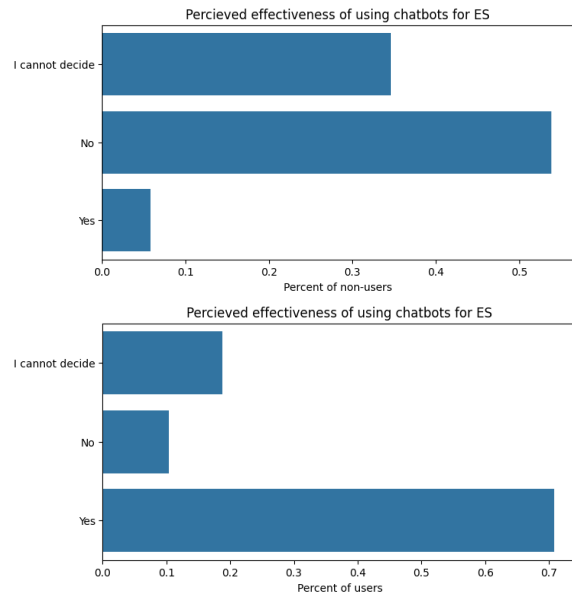


Fig. 6 Perceived effectiveness among users and non-users

2.4 Demographics by usage

Figure 7 shows the joint demographic distributions in response to the question "Do you use AI-based chatbots as a means of providing emotional support?".

The statistical analysis of the distributions showed the following correlations:

- **Gender:** no statistically significant correlation ($p = 0.344$).
- **Age:** a tendency towards a correlation, but not statistically significant ($p = 0.069$).
- **Level of education:** no significant correlation ($p = 0.423$).
- **Place of residence:** no significant correlation ($p = 0.443$).

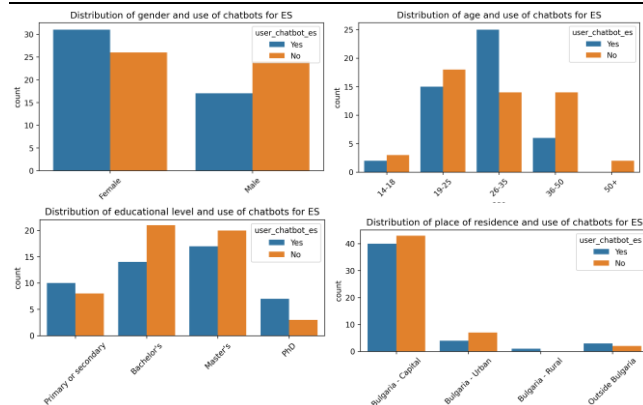


Fig. 7 Demographics by usage

Overall, no significant correlation between demographic characteristics and usage of AI-based chatbots for emotional support was found, meaning that both adopters and skeptics of such a technology occur across different age ranges, genders, educational levels and places of residence. The “age” category was the only one in which a tendency for correlation was found, showing that such use of the technology is mostly adopted by the 26–35 years’ old (people born 1990–1999, often informally referred to as “Zillennials” and “90’s kids”).

3. Qualitative research: Perspectives of users and specialists

In order to further get insights from the quantitative data from the survey, qualitative research was conducted through semi-structured interviews with users (non-specialists) and experts in the field of mental health. The results reveal a nuanced understanding of the role of AI as a tool for emotional support.

3.1. The chatbot as a tool for self-reflection and “venting”

The interviewed users identified several key functions of the LLM as an emotional supporter:

Guided journaling: Users describe the interaction as a form of “guided journaling” that helps structure thoughts and needs. “We often say, ‘Sit down and write it out’ or ‘Keep a journal’. Here, however, there is something a little different – because you get a response. Often, this response is filled with empathy and can encourage you to continue sharing more and more.”

Safe sharing: The chatbot is perceived as a “friend” to whom you can “vent” without the risk of embarrassing yourself or other people in front of loved ones or family.

Rationalization and cognitive reframing: The models help change your perspective (reframing) by asking clarifying questions. A respondent shares: “It helps you change your perspective if you ask good questions. In my opinion, the benefit depends on how well you use these models in general. As questions, I ask: ‘Ask me several clarifying questions to help me decide what to do.’ or ‘Is there anything I am overlooking in this situation?’”

No social pressure: For some users, AI is preferable to a psychologist because it eliminates the need to carefully consider their words and allows them to dive straight into the desired topic, without the need to agree with the psychologist’s point of view.

Encouragement and validation: An interesting case from the qualitative interviews involves a medical professional using AI to manage postpartum stress. Despite her professional background, she noted a distinct need: “I need to communicate with someone who is as knowledgeable as I am and can give me adequate advice.” She shares that her primary motive was not information retrieval, but emotional support and the need for a non-judgmental, yet knowledgeable “listener” to validate her maternal decisions – “And

even though I ask about medical conditions I have read about, I realize the responsibility of the decisions I make as a mother, and I need someone to encourage me.”

3.2. The sceptics’ point of view

The majority of non-users (55%) believe that technology is ineffective for emotional support purposes. Sceptics believe that human connection and personal relationships in the realm of mental health and emotional well-being cannot be replaced by technology. They prefer to receive understanding and attention from a real person (psychologist or psychotherapist) rather than talking to a chatbot.

Some users prefer to cope on their own or talk to friends in difficult moments – “I see no point in using artificial intelligence in the near future.”

3.2. The professional perspective

Interviews with specialists highlight the critical difference between “emotional relief” and “psychological change”. The experts emphasize on the lack of a biological basis for empathy: AI does not have “mirror neurons” and cannot establish real live contact. While therapists use emotional reactions (transference) to analyze clients’ behavior patterns, AI is not emotionally intelligent and cannot sense repressed emotions. Further, experts warn on the risk of superficial validation, as chatbots often give the user exactly what they want and ask for, rather than working with their psychological defenses.

4. Discussion

The findings of this research suggest a nuanced landscape of AI adoption for emotional support among Bulgarian users, characterized by a preference for multilingual support. The linguistic switching between English and Bulgarian may be due to the need of expressing both culturally specific emotional states and terminology, typical for the English language.

A central demographic trend identified in this study is the high adoption rate among individuals born between 1990 and 1999, often referred to as “Zillennials” [11]. Although the statistical correlation for age showed only a tendency rather than full significance ($p=0.069$), this group emerges as the primary user base. The group is characterized as an intersection of high digital literacy and significant life stressors, such as early-career pressures, early parenthood and evolving interpersonal relationships. For these users, the chatbot often serves as a “guided journal” or a “rational friend” that provides a non-judgmental space to “blow off steam” without the social risks.

However, the perceived effectiveness of these models – reported at 71% among users – must be weighed against the risk of “social sycophancy” [12], or the tendency of LLMs to provide excessive affirmation. This “flattery” concern was explicitly noted by 8% of respondents who worried that chatbots consistently take the user’s side. As suggested in the qualitative interviews, this tendency can be counterproductive, as the AI may simply validate the user’s current state rather than challenging their views in the discussed situation. This poses the risk of over-reliance: users may stop seeking human intervention if the AI always agrees with them. While this kind technology is excellent at providing immediate support, it lacks genuine empathy and the ability to work with complex phenomena like transference and countertransference [12].

5. Conclusion

The results of the presented study suggest that there is no consensus among the Bulgarian users on the extent to which chatbots based on LLMs are a useful technology for providing emotional support or not. Based on the conducted survey, the proportion of people who use them for this purpose is approximately equal to those who do not. Among the stated purposes for use, coping with stress in interpersonal relationships stands out, which certainly affects the

way people communicate with each other – now and in the near future. No statistically significant correlation was found between attitudes and use of the technology and demographic characteristics such as gender, level of education, age, and place of residence. The only notable exception is the group of people born between 1990 and 1999, which emerge as the predominant adopter group of the technology within this sample. This trend demonstrates a shift towards searching proactive, digital and conversational self-help, especially on the topics of parenthood, family life, and career development.

The biggest concerns stated by the surveyees are related to the reliability of the technology and data privacy, but tendency of models to provide excessive affirmation or “flattery” also stands out. This tendency can create emotional echo chambers that prevents psychological growth and encourage dangerous behavior or thought patterns. A primary recommendation for developers of such technology is a movement toward “constructive challenging” – designing models that can identify and gently confront a user’s harmful thought patterns rather than simply validating them.

In summary, while LLMs offer a solution to some of the shortcomings of traditional therapy by providing an always-available “guided journal”, they must be viewed as a supplemental technology for resilience rather than a replacement for the genuine human empathy and the deep psychological work inherent in human-to-human therapy.

Acknowledgements

The author gratefully acknowledge the support provided by the project UNiTe BG16RFPR002-1.014-0004 funded by PRIDST.

References

- [1] M. Zao-Sanders. How People Are Really Using Gen AI in 2025. Apr. 9, 2025. url: <https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025> (visited on 05/03/2026).
- [2] B. R. Burleson. Emotional support skills. Lawrence Erlbaum Associates Publishers, 2003.
- [3] L. Brocki, G. C. Dyer, A. G. ladka, and N. C. Chung. “Deep Learning Mental Health Dialogue System.” In: 2023 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE. 2023, pp. 395–398.
- [4] World Health Organization. Mental Health Atlas 2020. Geneva, October 8, 2021. url: <https://www.who.int/publications/i/item/9789240036703> (visited on 05/03/2026).
- [5] D. F. Santomauro, A. M. M. Herrera, J. Shadid, P. Zheng, C. Ashbaugh, D. M. Pigott, C. Abbafati, C. Adolph, J. O. Amlag, A. Y. Aravkin, et al. “Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic.” In: The Lancet 398.10312 (2021), pp. 1700–1712.
- [6] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency. “It’s only a computer: Virtual humans increase willingness to disclose.” In: Computers in Human Behavior 37 (2014), pp. 94–100.
- [7] D. R. Lehman and K. J. Hemphill. “Recipients’ perceptions of support attempts and attributions for support attempts that fail”. In: Journal of Social and Personal Relationships 7.4 (1990), pp. 563–574.
- [8] M. Baidal, E. Derner, and N. Oliver. “Guardians of Trust: Risks and Opportunities for LLMs in Mental Health.” In: Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI). Vienna, Austria: Association for Computational Linguistics, July

2025, pp. 11–22. isbn: 978-1-959429-19-7. url: <https://aclanthology.org/2025.nlp4pi-1.2/>.

[9] N. Obradovich, S. S. Khalsa, W. U. Khan, J. Suh, R. H. Perlis, O. Ajilore, and M. P. Paulus. “Opportunities and risks of large language models in psychiatry.” In: NPP—Digital Psychiatry and Neuroscience 2.1 (2024), p. 8.

[10] A. Babu and A. P. Joseph. “Artificial intelligence in mental healthcare: transformative potential vs. the necessity of human interaction.” In: Frontiers in Psychology 15 (2024), p. 1378904.

[11] Carganilla, Marielle, and John Rey Pelila. "ZILLENIAL MICROGENERATION: HYBRID TRAITS, DIGITAL BEHAVIOR, AND GENERATIONAL BOUNDARIES." *Lingue: Jurnal Bahasa, Budaya, dan Sastra* 7.2 (2025): 218-227.

[12] Cheng, Myra, et al. "Social sycophancy: A broader understanding of llm sycophancy." *arXiv preprint arXiv:2505.13995* (2025).

[13] Balint, Alice, and Michael Balint. "On transference and counter-transference." *The International Journal of Psycho-Analysis* 20 (1939): 223.

Artificial intelligence supporting human decision-making in technological crises

Zapryanova Tanya., Eng Radeva, Tatyana.

Bulgarian Academy of Sciences, Sofia, Bulgaria, Institute of Information and Communication Technologies
tanya.zapryanova@iict.bas.bg, tatyana.radeva@iict.bas.bg

Abstract: *Technological crises refer to critical events—system failures and collapses, cybersecurity and data breaches, software meltdowns, critical infrastructure disruptions, etc.—in which the failure, malfunction, or disruption of technological systems creates significant risk to safety, operations, or the environment. These crises typically arise unexpectedly and require rapid, well-informed decision-making under conditions of uncertainty, time pressure, and missing or incomplete information, to prevent escalation. Artificial intelligence (AI) systems have emerged as powerful tools capable of augmenting human judgment in these contexts. This article examines the role of AI in supporting human decision-making during technological crises, analyzing its capabilities, limitations, and implications for organizational resilience. It emphasized that AI is most effective when used as a collaborative partner rather than a replacement for human expertise, and it outlines design principles for trustworthy, human-centered AI systems.*

Keywords: ARTIFICIAL INTELLIGENCE, DECISION-MAKING, TECHNOLOGICAL CRISES, CRISIS MANAGEMENT, HUMAN-AI COLLABORATION, SITUATIONAL AWARENESS

1. Introduction

Modern societies depend on complex technical systems—energy grids, transportation networks, industrial plants, and digital infrastructures—that are vulnerable to unexpected failures. When crises occur, decision makers must interpret rapidly evolving data, coordinate responses, and mitigate cascading consequences. Yet traditional decision support tools often struggle to keep pace with the scale and speed of contemporary challenges. In this context, artificial intelligence (AI) has become a foundational component of modern decision making, enabling organizations to extract meaningful insights from high-dimensional datasets.

The rapid expansion of data in contemporary organizational environments has transformed the nature of decision-making, requiring advanced analytical tools capable of identifying trends, patterns, and actionable insights. Artificial intelligence has emerged as a critical enabler of data-driven decision processes, offering capabilities that surpass traditional analytical methods in scale, speed, and complexity.

This article examines the role of AI tools—including machine learning, natural language processing, deep learning, and business intelligence systems—in enhancing organizational decision-making. Particular emphasis is placed on the application of AI in technical crisis situations, where human decision-makers face extreme time pressure, uncertainty, and information overload. In such contexts, AI supports human cognition by improving situational awareness, detecting early warning signals, accelerating root-cause analysis, and recommending evidence-based response actions. AI systems also reduce cognitive load by filtering noise, prioritizing alerts, and automating routine tasks, thereby minimizing the risk of human error. Furthermore, AI-driven communication tools enhance coordination across teams, while post-crisis analytical capabilities contribute to organizational learning and resilience. Despite these advantages, the article highlights the importance of maintaining a balanced human-AI partnership, as human judgment remains essential for ethical reasoning, contextual interpretation, and final decision authority. In doing so, the article underscores that sustainable performance emerges not from replacing humans, but from integrating AI in ways that elevate human insight. Overall, the integration of AI into decision-making processes—especially during technical crises—strengthens organizational responsiveness, accuracy, and strategic foresight, positioning AI as a foundational component of modern decision support systems.

2. AI in High-Pressure Decision Environment

The integration of artificial intelligence into organizational decision processes has been widely examined across domains such as emergency management, cybersecurity, industrial operations, and critical infrastructure protection. Existing research highlights that

AI excels in environments characterized by high data volume, rapid change, and the need for predictive insight—conditions that closely mirror technological crisis scenarios. In such contexts, decision makers must process heterogeneous information streams, ranging from sensor data and system logs to human reports and external intelligence. Traditional analytical approaches often fail to synthesize these inputs quickly enough to support timely intervention.

AI systems, particularly those based on machine learning and deep learning, offer capabilities that enhance human cognition by identifying latent patterns, forecasting system behaviour, and detecting anomalies that may precede failure. Natural language processing (NLP) tools further expand these capabilities by extracting meaning from unstructured text sources such as incident reports, operator notes, and real-time communications. Together, these technologies form the backbone of modern decision support systems designed for crisis environments.

However, the literature also emphasizes that AI's effectiveness depends on its alignment with human cognitive processes [11], [15], [22]. Research in human-machine teaming suggests that AI should not replace human judgment but rather complement it by providing clarity, reducing uncertainty, and enabling faster comprehension of complex situations. This perspective underscores the need for human-centered AI design principles that prioritize transparency, interpretability, and trust. Scholars further argue that when AI systems are designed with an awareness of human cognitive limits—such as attention constraints and bias tendencies—they become far more reliable partners in high-pressure environments. As a result, effective integration requires not only technical sophistication but also a deep understanding of how humans perceive, reason, and make decisions under stress.

2.1. AI Capabilities Relevant to Technological Crisis Management

Technological crises are characterized by complexity, uncertainty, and rapid escalation, often involving tightly coupled systems where failures propagate unpredictably [18]. Human decision makers in such environments face cognitive overload, time pressure, and incomplete information, which can impair judgment and increase the likelihood of error [10], [11].

AI systems have emerged as critical tools for augmenting human cognition in these contexts. Machine learning, natural language processing, and predictive analytics enable organizations to process large volumes of heterogeneous data and detect patterns that would be difficult for humans to identify unaided [9]. From a socio-technical perspective, AI functions as a cognitive extension of human operators, supporting perception, comprehension, and projection—the three pillars of situational awareness [7], [8].

Human decision making under stress and uncertainty: In crisis conditions, human cognition is constrained by stress, fatigue, and information overload. Heuristics and intuitive judgments, while sometimes adaptive, can lead to systematic errors such as confirmation bias, tunnel vision, or overconfidence. Traditional decision support tools, which rely on static rules or manual data analysis, are often too slow or rigid to accommodate rapidly evolving situations. This creates a gap between the demands of the environment and the capabilities of human decision makers.

Early Detection and Anomaly Identification: AI-driven monitoring systems can detect deviations from normal operational patterns long before they escalate into full-scale crises. Machine learning models trained on historical system behaviour can identify subtle anomalies—such as irregular network traffic, abnormal sensor readings, or unusual equipment vibrations—that may indicate emerging threats. [4]. These early warning signals allow organizations to intervene proactively, reducing the likelihood of catastrophic failure.

Enhancing Situational Awareness: During crises, situational awareness deteriorates rapidly as information becomes fragmented and overwhelming. AI systems help reconstruct a coherent operational picture by aggregating data from multiple sources, filtering irrelevant information, and highlighting critical developments. Visualization tools powered by AI can dynamically map system interdependencies, enabling decision makers to understand cascading effects and prioritize response actions. AI-powered visualization systems assist crisis leaders by revealing interdependencies, vulnerabilities, and potential cascading failures across complex infrastructures. [6], [7], [14], [22].

Accelerating Root Cause Analysis: Identifying the underlying cause of a technological failure is often time-consuming, especially when multiple subsystems interact in complex ways. AI accelerates this process by analyzing logs, correlating events, and generating hypotheses about potential failure points. In cybersecurity incidents, for example, AI can trace intrusion paths, identify compromised assets, and recommend containment strategies [20].

Reducing Cognitive Load: Crisis environments impose significant cognitive strain on human operators. AI mitigates this burden by automating routine tasks, prioritizing alerts, and suppressing false positives. By reducing noise and focusing attention on the most critical issues, AI minimizes the risk of human error and supports more deliberate, informed decision-making. AI reduces cognitive load by filtering noise, clustering alerts, and automating routine tasks—critical in environments where information overload is common [5], [17].

Decision Recommendation and Scenario Simulation: Advanced AI models support decision makers by simulating potential response strategies and forecasting their outcomes. These tools evaluate trade-offs, estimate risks, and propose optimal courses of action based on available data. While humans retain final decision authority, AI-generated scenarios help reduce uncertainty and improve the quality of strategic choices. Simulation models and optimization algorithms help evaluate alternative response strategies and forecast outcomes under uncertainty [2], [13].

Data Integration and Real-Time Monitoring: AI-enabled monitoring systems can fuse sensor data, logs, and external intelligence into a unified operational picture, improving situational awareness and enabling early detection of anomalies [3], [22].

3. Human–AI Collaboration Models in Crisis Management

Effective crisis management requires a clear understanding of how responsibility, autonomy, and oversight are distributed between human decision makers and AI systems. The literature identifies three foundational models—Human-In-the-Loop (HITL), Human-On-the-Loop (HOTL), and Human-Out-of-the-Loop (HOOTL)—which describe different configurations of human–AI interaction (see Fig.1) [5], [21]. These models are particularly relevant in technological crises, where the balance between speed, accuracy, and accountability is critical. Selecting the appropriate model is not merely a technical choice but a strategic one, shaped by the nature of the crisis, the reliability of available data, and the ethical implications of delegating authority to automated systems. Understanding these models helps organizations calibrate AI autonomy in ways that enhance performance without compromising safety or human oversight.

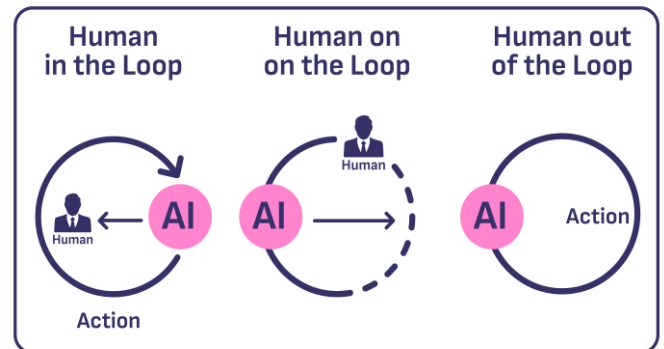


Fig. 1 Three roles of humans in AI systems

3.1 Human-In-the-Loop (HITL): Human as Primary Decision Authority

In the HITL model, humans remain central to all critical decisions. AI systems provide analytical support—such as anomaly detection, risk assessment, or scenario generation—but no action is executed without explicit human approval. This model aligns with research showing that human judgment is essential for contextual interpretation, ethical reasoning, and handling uncertainty [12], [16]. HITL is particularly valuable in ambiguous or ethically sensitive situations where human intuition, moral reasoning, and situational awareness cannot be fully replicated by algorithms. However, this model can introduce delays in fast-moving crises, making it most suitable for environments where accuracy and accountability outweigh the need for immediate action.

3.2 Human-On-the-Loop (HOTL): Human as Supervisor and Intervener

The HOTL model grants AI systems a degree of autonomy to act within predefined boundaries, while humans maintain supervisory control and can intervene when necessary. This model is widely used in high-reliability domains such as aviation, cybersecurity, and industrial automation [5], [7]. HOTL is particularly effective when the speed or volume of data exceeds human cognitive capacity, but oversight remains essential to prevent automation surprises. By allowing AI to manage routine or time-critical tasks, HOTL reduces cognitive load and frees human operators to focus on strategic decision making. Yet, this model also requires robust interface design and clear alerting mechanisms to ensure that humans can re-enter the decision loop quickly and effectively when anomalies occur.

3.3 Human-Out-of-the-Loop (HOOTL): AI as Autonomous Decision Maker

In the HOOTL model, AI systems operate independently without real-time human involvement. This model is used in environments requiring millisecond-level responses, such as autonomous shutdown systems, intrusion prevention systems, or industrial safety controls [1]. However, HOOTL raises concerns about accountability, transparency, and the potential for cascading failures if the AI misinterprets a situation [19]. While full autonomy can dramatically increase operational speed and reduce human error in highly structured environments, it also concentrates decision authority in systems that may lack contextual understanding or the ability to adapt to novel conditions. Consequently, HOOTL is most appropriate only when system behaviour is well-defined, risks are tightly controlled, and fail-safe mechanisms are rigorously tested.

Table 1: AI–Human Collaboration in High-Pressure Decision Environments

Dimension	Human Strengths	AI Strengths	Collaborative Advantage
Information Processing	Interprets ambiguity; applies contextual judgement	Processes large, heterogeneous data streams rapidly	Improved situational awareness under time pressure
Pattern Recognition	Experience-based intuition	Detects latent patterns, anomalies, correlations	Earlier detection of failures or emerging threats
Decision Speed & Accuracy	Ethical reasoning; ensures accountability	Provides rapid computation, forecasting, and scenario evaluations	Timely, well-informed decisions in crisis conditions
Adaptability	Creative problem-solving; improvisation	Consistent monitoring; scalable analysis	Resilient response among dynamic and evolving crises
Communication & Interpretation	Negotiations; stakeholder coordination	NLP-based extraction from unstructured reports and logs	Reduced cognitive load; clearer information flow
Autonomy Models (HITL/HOTL/HOOTL)	Oversight; value-based judgment	Automation of routine or time-critical tasks	Balanced autonomy, safety, and accountability

Table 1 outlines the complementary roles of humans and artificial intelligence in high-pressure decision environments, emphasizing how their distinct strengths can be integrated to enhance overall performance. Humans contribute contextual judgment, ethical reasoning, improvisation, and interpersonal communication—capabilities that remain essential when dealing with ambiguity, value-laden choices, and complex stakeholder dynamics. In contrast, AI systems excel at rapid data processing, large-scale pattern detection, continuous monitoring, and automated execution of routine or time-critical tasks.

When combined, these strengths create a collaborative advantage across several dimensions. Joint information processing improves situational awareness by merging human contextual insight with AI’s ability to analyze heterogeneous data streams. Pattern recognition benefits from both human intuition and AI’s capacity to detect latent anomalies, enabling earlier identification of emerging

risks. Decision-making becomes faster and more accurate when human accountability and ethical oversight are paired with AI-driven forecasting and scenario evaluation. Collaboration also enhances adaptability: human creativity complements AI’s consistent, scalable analysis to support resilient responses in dynamic crises.

Finally, communication and autonomy models benefit from this partnership. Human negotiation and coordination skills are reinforced by AI’s natural-language processing capabilities, reducing cognitive load and improving information flow. Across different configurations—human-in-the-loop, human-on-the-loop, and human-out-of-the-loop—balanced integration ensures safety, accountability, and effective task allocation. Overall, the table highlights that optimal performance in high-pressure environments emerges not from replacing human decision-makers, but from designing systems that strategically combine human judgment with AI’s computational strengths.

4. Conclusion

Technological crises expose the limits of human cognition and traditional decision support tools, demanding rapid, informed action under uncertainty and time pressure. Artificial intelligence offers a powerful set of capabilities that can substantially enhance crisis management: augmenting situational awareness, identifying weak signals and emergent anomalies, accelerating diagnostic processes, and structuring high-dimensional decision spaces in ways that exceed unaided human capacity. Yet its greatest value emerges not when it replaces human judgment, but when it functions as a collaborative partner embedded within socio-technical systems.

Realizing this potential requires organizations to confront the limitations and risks inherent in AI systems. Issues such as data bias, model opacity, overreliance on automated outputs, and susceptibility to adversarial manipulation can undermine trust and compromise crisis response. Addressing these vulnerabilities demands sustained investment in human-centered design, transparent model development, rigorous validation, and governance structures that promote calibrated trust rather than blind dependence. Equally important is cultivating a workforce capable of understanding AI’s strengths and constraints, ensuring that human operators remain empowered to question, interpret, and override automated recommendations when necessary.

Future research should advance theoretical and empirical understanding of adaptive human–AI collaboration models capable of evolving dynamically as crises unfold. Promising directions include the development of AI systems that support cross-domain transfer of crisis knowledge, the integration of AI with emerging infrastructures such as digital twins, distributed sensing networks, and edge computing architectures, and the exploration of hybrid intelligence frameworks that combine machine learning with human expertise, institutional memory, and contextual reasoning. Such efforts must be accompanied by interdisciplinary inquiry into the organizational, legal, and ethical implications of AI-mediated crisis response.

By aligning AI capabilities with human expertise, institutional processes, and robust governance mechanisms, organizations can enhance systemic resilience, mitigate the impacts of technological disruptions, and navigate an increasingly complex risk landscape with greater foresight, adaptability, and confidence. Ultimately, the future of crisis management will depend not on the ascendancy of autonomous systems, but on the deliberate design of socio-technical ecosystems in which humans and AI function as mutually reinforcing partners.

5. Literature

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. Concrete problems in AI safety, (2016). arXiv:1606.06565.
2. Bankes, S. Exploratory modeling for policy analysis. *Operations Research*, 41(3), (1993), 435–449.
3. Batareseh, F. A., & Yang, R. AI for data-driven crisis management. Springer. (2021).
4. Chandola, V., Banerjee, A., & Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), (2009), 1–58.
5. Cummings, M. LMan versus machine or man + machine? *IEEE Intelligent Systems*, 29(5), (2014), 62–69.
6. Endsley, M. R. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), (1995), 32–64.
7. Endsley, M. R. Automation and situation awareness. In *Handbook of human factors and ergonomics* (2018), pp. 533–546
8. Hollnagel, E., *Safety-II in practice: Developing the resilience potentials*. Routledge, (2017)
9. Jordan, M. I., & Mitchell, T. M., *Machine learning: Trends, perspectives, and prospects*. *Science*, 349(6245), (2015), 255–260.
10. Kahneman, D., *Thinking, fast and slow*. Farrar, Straus and Giroux, (2011)
11. Klein, G., *Naturalistic decision making*. *Human Factors*, 50(3), (2008), 456–460.
12. Klein, G., *Sources of power: How people make decisions*. MIT Press., (2017)
13. Miller, C. A., & Parasuraman, R., *Designing for flexible interaction between humans and automation*. *Human Factors*, 49(1), (2007), 57–71.
14. Ouyang, M., *Review on modeling and simulation of interdependent critical infrastructure systems*. *Reliability Engineering & System Safety*, 121, (2014), 43–60.
15. Panteli, M., & Mancarella, P. Influence of extreme weather and climate change on the resilience of power systems. *IEEE Transactions on Smart Grid*, 7(3), (2015), 1–10.
16. Parasuraman, R., & Riley, V., *Humans and automation: Use, misuse, disuse, abuse*. *Human Factors*, 39(2), (1997), 230–253.
17. Parasuraman, R., Sheridan, T. B., & Wickens, C. D., *A model for types and levels of human interaction with automation*. *IEEE Transactions on Systems, Man, and Cybernetics*, 30(3), (2000), 286–297.
18. Perrow, C., *Normal accidents: Living with high-risk technologies*. Princeton University Press. (1999)
19. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Wellman, M., *Machine behaviour*. *Nature*, 568(7753), (2019), 477–486.
20. Salfner, F., Lenk, M., & Malek, M. A survey of online failure prediction methods. *ACM Computing Surveys*, 42(3), (2010), 1–42.
21. Sheridan, T. B., *Human–robot interaction: Status and challenges*. *Human Factors*, 58(4), (2016), 525–532.
22. Shneiderman, B. *Human Centered AI*. Oxford University Press, Oxford, UK, (2022), 111–137; 145–179

AI-based technologies in the authentication of fine art: toward a hybrid epistemology of cultural trust

Atanasova D. Y.
Sofia, Bulgaria
E-mail: yankovad@gmail.com

Abstract: *The authentication of fine art has customarily relied on expert connoisseurship, material analysis, and provenance research. In recent years, artificial intelligence (AI) and AI-based technologies have appeared as significant tools in this domain, enabling new forms of algorithmic evidence, probabilistic reasoning, and large-scale pattern recognition. This paper examines how AI-based systems support museums, galleries, collectors, and private institutions in authenticating fine art paintings. It argues that AI does not replace human expertise but establishes a hybrid epistemic framework in which algorithmic forensics and art-historical knowledge co-produce authenticity. The study analyses key technological approaches, institutional applications, epistemological implications, and structural drawbacks, positioning AI as a catalytic agent in reconfiguring trust, authority, and knowledge production within the contemporary art ecosystem.*

KEYWORDS: ARTIFICIAL INTELLIGENCE, ART AUTHENTICATION, HYBRID EPISTEMOLOGY, PROVENANCE ANALYSIS, CULTURAL ANALYTICS

1. Introduction

Authentication constitutes the epistemic and economic foundation of the fine art ecosystem, determining not only cultural value but also legal status, market price, and institutional legitimacy [10]. Traditional authentication methods—connoisseurship, material analysis, and provenance research—now face growing challenges from increasingly sophisticated forgeries and the global expansion of art markets, which exceed human analytical capacity.

Recent advances in artificial intelligence have introduced computational paradigms capable of analysing visual, material, and historical characteristics at unprecedented scale and precision [13]. However, the assimilation of algorithmic systems into cultural heritage raises fundamental epistemological questions about the nature of authenticity itself.

This paper contends that AI-based technologies not just supplement but fundamentally reshape the process of authenticating fine art. We advance a “hybrid epistemology of cultural trust”, which systematically integrates computational forensics, expert art-historical analysis, and institutional authority to establish authenticity. Rather than relying on a single methodological approach, this model coordinates algorithmic and human assessment across technical, scholarly, and organisational domains. While it enhances rigour and transparency, it also introduces possible challenges, including diminished expert authority, excessive dependence on opaque probabilistic methods, and reinforcement of organisational biases. Therefore, this system highlights the importance of continuous critical assessment of the limitations of merging AI with established authentication practices.

To operationalise this hybrid model, museums and universities can implement concrete measures at multiple levels. Staff training programs should be updated to include basic AI literacy, ensuring that curators, conservators, and registrars are familiar with the principles and limitations of machine learning tools. Institutions can revise authentication workflows to incorporate AI analyses as an evidentiary layer that complements traditional review panels, encouraging multidisciplinary collaboration. Policy updates might require transparent documentation of AI-supported decisions and regular audits of algorithmic outputs to promote accountability. Furthermore, ethical guidelines should be introduced to guard against bias, mandate data diversity, and clarify the interpretive role of human experts alongside automated assessments. By systematically embedding these changes, institutions will be better equipped to integrate AI into everyday authentication practice while upholding scholarly rigour and cultural trust.

2. Results and Discussions

2.1. Technological Foundations and Methodologies

- *Computer Vision and Deep Learning.* Modern AI authentication systems employ convolutional neural networks (CNNs) and visual transformers to analyse micro-level visual features, including brushstroke dynamics, colour distributions, compositional geometry, and textual patterns [5]. Recent studies show that hybrid CNN-transformer models attain improved performance by extracting both local brushstroke features and global compositional structures, with classification accuracies frequently ranging from 80% to 95% on controlled datasets [13]

The breakthrough lies in converting subjective stylistic judgement into assessable computational data, producing probabilistic outputs such as “92% probability of Rembrandt authorship” rather than binary determinations [8].

- *Multispectral and Hyperspectral Analysis.* AI-enhanced spectral imaging enables non-invasive examination through X-ray fluorescence (XRF), infrared reflectography (IRR), and hyperspectral imaging. Machine learning methods can detect anachronistic pigments, modern binders in historical works, hidden underdrawings, and restoration layers with reliability exceeding 85% under regulated conditions [3].

ASML's Impasto Project illustrates this approach, creating 3D digital twins of Vincent van Gogh's artworks with 100 gigabytes of data per scan, utilising nanometre-capable chipmaking technology to generate “Google Maps for paintings” [12].

- *Provenance and Network Analysis.* Graph-based AI systems construct knowledge graphs from auction records, exhibition catalogues, museum databases, and archival photographs. These systems can automatically detect missing provenance periods, suspicious ownership patterns, and stylistic clusters associated with known forgers, thereby raising anomaly-detection efficiency by approximately 30-40% across big datasets [9].

2.2. Institutional Applications and Case Studies

- *Museums.* Major museums use AI for broad collection analysis, but implementation reveals varying institutional priorities that shape the scope and impact of technological adoption. The Rijksmuseum applies residual neural networks to identify large-scale patterns across entire collections, emphasising a macro-level understanding of organisational trends and misattributions. In contrast, the Metropolitan Museum of Art incorporates AI more narrowly into conservation workflows for individual objects,

focusing on micro-level technical analysis and object-specific decision-making [6]. The distinction between the Rijksmuseum's collection-level approach and the Metropolitan Museum's object-level focus highlights how institutional aims shape both the strategic aims and the epistemic contributions of AI integration. Despite this contrast, both institutions face comparable challenges: ensuring unbiased training data, preventing cataloguing errors, and managing the interpretability of algorithm-generated outputs. These cases underscore the significance of aligning innovative AI implementation with established scholarly practices and shared curatorial authority to promote institutional effectiveness and credibility.

- *Auction Houses and Galleries.* Sotheby's and Christie's have incorporated AI-driven technical imaging and machine learning into their due diligence procedures, primarily aiming to enhance pre-acquisition risk assessment and market intelligence. While these integrations enable auction houses to detect potential forgeries with increased accuracy and efficiency, they also introduce new complexities into the authentication process. For instance, the reliance on algorithmic outputs may shift the locus of expertise from traditional specialists to technology providers, potentially redefining what is considered authoritative authentication. Moreover, there is a risk that overreliance on AI-generated evidence could obscure subtle interpretive or contextual nuances that human experts might discern. Despite these concerns, auction houses have leveraged these technologies to reduce reputational risk and bolster buyer confidence, directly benefiting stakeholders such as collectors, consignors, and institutional clients by providing greater assurance in the authenticity and market value of works. However, these developments also necessitate ongoing scrutiny regarding how algorithmic opacity, data representativeness, and potential biases might impact the broader art market. Thus, although AI-based implementations have the potential to increase trust and inform more rational economic decision-making in speculative markets, they simultaneously require critical evaluation to ensure that enhanced efficiency does not compromise the epistemic integrity of authentication practices [2].

Table 1: Comparative Performance of Authentication Approaches

Connoisseurship	70-85%	Contextual judgment	Subjectivity
Material Analysis	80-95%	Scientific rigour	Expensive, slow
AI Image Analysis	80-92%	Pattern detection at scale	Data dependence
Hybrid Approach	90-97%	Complementary validation	Integration complexity

- *Private Collections and Commercial Platforms.* Commercial platforms such as Art Recognition illustrate the market demand for scalable computational attribution. In 2023, Art Recognition was commissioned by a private Swiss collector to assess the authenticity of a portrait attributed to Amedeo Modigliani. Utilising their proprietary AI system, trained on verified Modigliani works, the platform analysed high-definition images and identified differences in brushwork and design features, resulting in a revised attribution [4]. The AI's rapid visual analysis detected subtle variations in brushstroke rhythm and pigment layering that typically require extensive microscopic examination by human experts. The probability analysis assigned a significantly lower likelihood of

Modigliani authorship, casting doubt on the painting's authenticity and prompting further scholarly debate regarding the evidentiary weight of AI results compared to provenance documentation and stylistic evaluation. A critical synthesis reveals that AI provided a novel form of evidence—quantifiable and replicable—that augmented but did not replace traditional authentication methods. Rather than serving as a final arbiter, the AI's findings highlight the probabilistic and provisional nature of computational analysis, emphasising the need for art historians and domain experts to scrutinise both the methodology and implications of algorithmic judgments. This case exposes the epistemological complexity of integrating AI with human expertise. While AI enabled efficient triage and revealed previously undetected features, its output depended on the completeness of reference datasets and required contextual validation within disciplinary frameworks. Ultimately, the Modigliani attribution demonstrates that AI-generated reports should function not as unilateral determinants but as catalysts for multidisciplinary dialogue, reinforcing the necessity of critical engagement and iterative review within evolving art authentication practices.

To address the broader international scope of art authentication, some platforms have begun integrating non-Western and underrepresented art traditions into their datasets, broadening institutional relevance beyond Western-centric collections. The recent pilot project involving Ming dynasty porcelain highlights this shift: AI algorithms, trained on digitised records and multispectral imagery sourced from Chinese museums and private collections, were used to authenticate ceramic works. The project's successful differentiation between authentic Ming pieces and later reproductions—based on subtle kiln signatures and glaze compositions unique to specific regions—demonstrates that AI methods can be tailored to the material and stylistic complexities of globally diverse artefacts, not solely Western paintings. Nevertheless, persistent limitations, such as the scarcity of high-quality reference images and the underrepresentation of localised styles, challenge the inclusivity and generalizability of these systems, revealing a fundamental constraint of the broader hybrid epistemological model. The Ming porcelain pilot, therefore, not only illustrates the expanded potential for AI-supported authentication in non-Western contexts but also exemplifies how the diversity and representativeness of underlying datasets ultimately limit the effectiveness of hybrid approaches. This case demonstrates that without deliberate efforts to ensure inclusivity at the data-collection and methodological levels, the promise of a truly global hybrid framework remains constrained, reinforcing the need for institutions to address structural data biases when integrating algorithmic evidence with human expertise on a worldwide scale.

Table 2: Institutional AI Applications in Art Authentication

Rijksmuseum	Computer vision	Attribution research	Collection-wide analysis
Sotheby's	Technical imaging + ML	Due diligence	Risk reduction
Christie's	Data analytics	Market intelligence	Pricing confidence
Getty Research Institute	Provenance data transformation	21st-century research	Enhanced accessibility
Art Recognition	CNN authentication	Collector support	Pre-purchase screening

2.3. The Hybrid Epistemological Framework

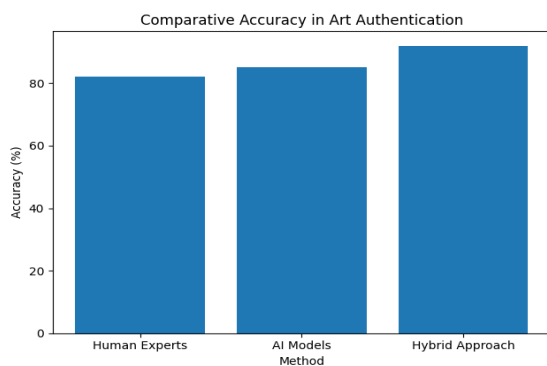
- *From Authority to Statistical Coherence.* Artificial intelligence shifts the concept of authenticity from reliance on expert authority to recognition of statistical patterns. Rather than asserting absolute truth, AI generates probabilistic knowledge derived from empirical data, thereby framing authenticity as a system-based process rather than solely expert opinion.

- *Augmented Connoisseurship Model.* "Augmented connoisseurship" refers to a hybrid epistemic system in which:

- Within this framework, AI provides computational proof by generating quantitative evidence through the analysis of large datasets and the application of algorithmic models.
- Experts provide contextual interpretation.
- Institutions provide legitimacy validation.

Authenticity emerges from interactions among datasets, algorithms, and institutional authentication procedures rather than from any single authority [10].

Fig. 1 (Hybrid Epistemology Diagram)



- *Cultural Trust Networks.* The Getty Research Institute's provenance data initiative embodies the hybrid model by directly combining computational innovation with established scholarly protocols. By digitising and consolidating provenance records into an accessible database, the project not only increases data transparency but also enables cross-institutional research without jeopardising curatorial oversight or scholarly review standards [7]. This concentrated approach illustrates how digital technology can expand research capacity and accessibility while retaining the strict evaluation required for data dependability and interpretive validity. In parallel, the Centre for Art Law's Framework for Responsible Use of AI in Art Authentication provides structured guidelines that integrate practical, legal, and ethical considerations, ensuring that new AI tools are introduced without undermining institutional confidence and stewardship in authentication processes [14].

2.4. Constraints and Difficulties

- *Structural Boundaries.*

- Despite impressive technical progress, AI confronts major limitations:
 - Dataset Bias: Systems trained on historically attributed works may reproduce and amplify existing misattributions
 - The Perfect Forgery Problem: Skilled forgeries may statistically outperform genuine experimental works which diverge from established stylistic norms
 - Ontological Limits: AI cannot access artistic intention, symbolic meaning, or cultural context

- *Ethical and Methodological Concerns.* These challenges emphasise the vital need for a concise set of moral guidelines for the use of AI in art authentication. Primary recommendations include ensuring data representativeness to avoid marginalising non-Western artistic traditions, increasing transparency in algorithmic processes to address the opacity of deep learning, and implementing routine multidisciplinary oversight.

To implement these recommendations, institutions can take several concrete actions. First, museums and galleries should establish dedicated ethics committees to oversee the integration of AI tools, review algorithms for fairness, and evaluate their impacts on established authentication procedures. Additionally, acquisition protocols and cataloguing standards should be updated to require that provenance documentation, artists' records, and technical analyses include data fields relevant to AI-assisted evaluation and are maintained in formats compatible with computational analysis.

Institutions ought to establish clear ethical codes, apply mechanisms to correct past biases inside datasets, and avoid treating probabilistic AI outputs as definitive judgments. Routine training programs can be introduced to increase AI literacy among curators, conservators, and administrative staff, ensuring informed interpretation of AI-generated results. Regular audits and transparent reporting of AI-supported decisions further promote accountability. These specific measures, when adopted systematically, jointly aim to promote fairness, accountability, and equity, safeguarding the ethical duties of stakeholders across the global art ecosystem.

2.5. Upcoming Directions and Implications

The future trajectory of AI in art authentication is expected to involve multimodal models that analyse high-resolution visual imagery, material composition data from multispectral scanning, and digitised archival or textual documentation in combination. Implementing these innovations will require streamlined protocols for data capture, standardised metadata annotation, and systematic cross-linking of diverse sources. Advances in explainable AI, including saliency attribution maps and transparent rule-based decision trees, will enhance methodological transparency by making algorithmic reasoning accessible for review by both art historians and non-specialist stakeholders. This transparency is crucial for interpretive reliability, particularly as collaborative attribution platforms incorporate expert validation and iterative feedback. As real-time consensus-building and error-correction become more prominent, new demands will arise for effective interoperability between human and artificial agents, shifting the methodological emphasis from static individual judgment to ongoing, multidisciplinary committee-based review.

These recommendations target both institutional and policy design levels within the art authentication ecosystem. Institutionally, museums, galleries, and private collections should revise acquisition protocols and record-keeping practices to ensure compatibility with AI systems, adopt standardised metadata and imaging formats, and provide staff training in data management and AI analysis. At the policy level, regulatory bodies should develop guidelines promoting responsible, unbiased data use and equitable access to AI resources. Sector-wide initiatives, such as establishing open, multilingual metadata repositories and interoperable curation standards, will enhance participation, especially for underrepresented collections [11]. Additionally, harmonising data-sharing and funding practices across national borders will reduce disparities in access and representation. Collectively, these measures address practical stakeholder needs and core technical and ethical priorities essential for implementing a hybrid epistemological model for AI-powered art authentication [1].

Blockchain-supported certification systems, while not AI themselves, complement computational authentication by providing tamper-resistant provenance records and digital certificates anchored to distributed ledgers, boosting disclosure and long-term traceability.

3. Conclusion

To summarise, this paper makes an original contribution by articulating and substantiating a hybrid epistemological model for art authentication that uniquely unites computational forensics, expert interpretation, and institutional validation. The research advances the field by not only synthesising the roles of AI and human expertise, but also by proposing a structured partnership that clarifies how each component contributes functionally to authentication processes. Via systematic analysis of the limitations and synergies inherent in uniting these elements, the paper argues that authenticity is most reliably established through this combined framework rather than through automation alone. This approach offers a novel foundation for both scholarly discussion and institutional policy by enhancing technical accuracy, protecting the integrity of cultural heritage, and strengthening organisational accountability and societal trust. Thus, the proposed synthesis establishes a strong and flexible blueprint for addressing emerging challenges in contemporary art authentication.

The proposed hybrid epistemological model, which integrates computational forensics, expert interpretation, and institutional validation, delivers a sustainable framework for cultural heritage institutions navigating technological transformation while continuing scholarly rigour and cultural trust. Success depends not on automation but on the careful orchestration of human-computer collaboration that preserves the interpretive richness of art-historical scholarship while leveraging AI's analytical strengths. To accelerate the adoption and effectiveness of this hybrid model, policy initiatives and funding opportunities should be prioritised. Policymakers and funding agencies can support institutions by providing resources to upgrade digital infrastructure, develop standardised protocols for AI integration, and facilitate cross-institutional collaboration. Dedicated funding streams for research, staff training, and data curation are essential to ensure equitable access and to address disparities between well-resourced and underrepresented collections. Sector-wide policy frameworks that encourage ethical AI practices and ongoing evaluation will further strengthen the transition to a transparent and accountable authentication ecosystem.

The future of art authentication will be defined by institutions that master this integration, creating transparent, accountable systems in which artificial intelligence functions as a strong evidentiary tool inside broader ecosystems of cultural knowledge and trust.

4. References

1. A Adorni, G. & Bellini, E. (2025). Towards a Manifesto for Cyber Humanities: Paradigms, Ethics, and Prospects. arXiv preprint. <https://doi.org/10.48550/arXiv.2508.02760>, Accessed: 10.02.2026
2. Lawson-Tancred, J., (2025). AI and the Art Market (Hot Topics in the Art World). <https://oceanofpdf.com/authors/jo-lawson-tancred/pdf-epub-ai-and-the-art-market-hot-topics-in-the-art-world-download/>, Accessed: 12.02.2026
3. Polak, A., Kelman, T., Murray, P., Marshall, S., (2017). Hyperspectral imaging combined with data classification techniques as an aid for artwork authentication. https://www.researchgate.net/publication/314670114_Hyperspectral_imaging_combined_with_data_classification_techniques_as_an_aid_for_artwork_authentication, Accessed: 15.02.2026
4. Dobbs, T., Nayeem, A., Cho, I. & Ras, Z. (2023). Contemporary Art Authentication with Large-Scale Classification. Big Data Cogn. Comput. <https://doi.org/10.3390/bdcc7040162>, Accessed: 01.02.2026
5. Elgammal, A., Mazzone, M., Liu, B., Kim, D., & Elhoseiny, M. (2018). The Shape of Art History in the Eyes of the Machine. <https://www.researchgate.net/publication/322694894>, Accessed: 01.02.2026
6. Expert Systems with Applications. (2022). On art authentication and the Rijksmuseum challenge: A residual neural network approach, 202. <https://doi.org/10.1016/j.eswa.2022.116933>, Accessed: 05.02.2026
7. Getty Research Institute. (May 1, 2025). Getty Transforms Art Provenance Data to Support 21st Century Research. Getty News. <https://www.getty.edu/news/getty-transforms-art-provenance-data-to-support-21st-century-research/>, Accessed: 07.02.2026
8. Johnson, C. R., Hendriks, E., Berezhnoy, I., et al. (2008). Image processing for artist identification. <https://ericpostma.nl/publications/Johnsonetal2008.pdf>, Accessed: 11.02.2026
9. Münster, S., et al. (2019). Digital cultural heritage meets digital humanities. https://www.researchgate.net/publication/335417192_DIGITAL_CULTURAL_HERITAGE_MEETS_DIGITAL_HUMANITIES, Accessed: 07.02.2026
10. Silva, R. S., Lotfi, A., Ihianle, I. K., Shahtahmassebi, G., & Bird, J. J. (2024). ArtBrain: An Explainable End-to-End Toolkit for Classification and Attribution of AI-Generated Art and Style. arXiv preprint. <https://doi.org/10.48550/arXiv.2412.01512>, Accessed: 10.02.2026
11. Thiel, S. & Bernhardt, J. C. (2024). AI in Museums: Reflections, Perspectives and Applications. transcript publishing. https://library.oapen.org/bitstream/handle/20.500.12657/87430/1/9783839467107.pdf?trk=public_post_comment-text, Accessed: 02.02.2026
12. Tom's Hardware. (June 14, 2025). ASML's Impasto Project creates 3D digital twins of Vincent van Gogh's art with 100 gigabytes of data per scan. <https://www.tomshardware.com/3d-printing/3d-scanning/asmls-impasto-project-creates-3d-digital-twins-of-vincent-van-goghs-art-with-100-gigabytes-of-data-per-scan-nanometer-capable-chipmaking-tech-used-to-create-google-maps-for-paintings>
13. Wang, Z. & Song, H. (2025). A Fusion Model for Artwork Identification Based on Convolutional Neural Networks and Transformers.. <https://doi.org/10.48550/arXiv.2502.18083>, Accessed: 08.02.2026
14. Centre for Art Law. (2025). Framework for Responsible Use of AI in Art Authentication. <https://itsartlaw.org/ai-and-art-authentication-guidelines/>, Accessed: 04.02.2026

Artificial intelligence as a tool for systemic modelling and simulation of architectural strategies for a sustainable future

Stoyanka Ivanova*

University of Architecture, Civil Engineering and Geodesy (UACEG) – Sofia, Bulgaria

siva_fce@uacg.bg

Abstract: In this work, AI is examined as a tool for the simulation of climate scenarios and the systemic modelling of the future of the built environment under conditions of climate change. Rather than relying on numerical physical simulation, the approach is based on qualitative analysis using a set of possible future scenarios, in which the same building is considered within different climatic and economic contexts up to the end of the twenty-first century. The study examines how deteriorating climatic conditions redirect resources from development toward the compensation of losses and how this transforms the architectural environment over time. The paper discusses the potential for buildings to act as active participants in the carbon balance through mechanisms for CO₂ removal and long-term fixation. AI supports the analysis of causal relationships between climate scenarios, economic incentives, and architectural decisions in order to avoid systemically problematic strategies. The approach is applicable both in professional practice and as an educational exercise for architecture students.

Keywords: CLIMATE CHANGE, ARTIFICIAL INTELLIGENCE, SYSTEMIC MODELLING, SCENARIO SIMULATION, BUILT ENVIRONMENT, CAPITAL REALLOCATION, SUSTAINABILITY

1. Introduction

The buildings we design and construct today will exist in the future – for another 80–100 years. This means that a significant part of the contemporary building stock will function under climatic conditions different from those of today. Scientific research in the field of climate change outlines several possible development trajectories, summarised in four main scenarios: RCP 2.6, 4.5, 6.0 and 8.5 [1-3] – from the most favourable to the most unfavourable in terms of global warming and the associated risks. Each of them assumes different emission dynamics, different final temperatures, frequency of extreme events and sea level changes by 2100, with estimates varying according to the models used.

In most studies, climate scenarios are used to analyse the physical behaviour of buildings [4-6] – temperature loads, humidity, extreme weather events, the impact of rising sea levels. However, these analyses rarely include the economic dimension of the problem. Each climate scenario implies not only a different physical environment, but also a different state of society, a different structure of public and private budgets, and a different distribution of public resources.

The link between climate change and the building sector can be examined in two main aspects. The first is related to the impact on the design, operation and long-term sustainability of buildings. The second is economic – the costs of adaptation, repair, maintenance and damage limitation [7,8]. It is precisely the second aspect that is often underestimated in strategic planning.

The main economic hypothesis put forward by this study is that, as climate conditions deteriorate, capital gradually changes its function: from capital for development to capital for compensating losses. This represents a fundamental change affecting the building stock, cities, public budgets, private investors and long-term economic growth.

Investment costs create future value – new buildings, improved technologies, higher quality of life and increased productivity. Compensation costs, on the other hand, do not create new value, but restore damage, limit losses and maintain minimum functionality. In more unfavourable climate scenarios, the share of the second type of expenditure increases significantly [7,9].

Under relatively favourable conditions (conditionally close to RCP 2.6), the majority of funds are directed towards new construction, value-added renovations and innovation. The economy operates proactively and works towards future growth. In intermediate scenarios (RCP 4.5–6.0), part of the investment resource is gradually "eaten up" by repairs after extreme events, increased insurance costs and adaptation of existing systems. The same funds begin to generate fewer improvements for the future, and the economy becomes increasingly reactive.

In severe scenarios (RCP 8.5), a qualitative change occurs: a significant portion of resources is directed not towards creating new value, but towards maintaining buildings in a habitable condition. Funds are invested in flood recovery, emergency reconstruction and temporary solutions. The economy shifts into maintenance mode, and growth is replaced by survival. This leads to a crowding-out effect – funds that could be directed towards education, scientific research, culture and high-quality architecture are redirected towards compensation and restoration.

The buildings designed today will fall precisely within this transition period (2030–2100), but they are created with today's standards, economic expectations and return models. As a result, we may see earlier economic obsolescence, a shortening of the real life cycle and the need for unplanned capital injections. In severe scenarios, there is also a risk of "locked-in capital" – significant funds invested in protective facilities and adaptation measures that do not generate income or increase productivity, but only delay losses.

It is in this context that there is a need for a tool that allows for a systematic examination of the interrelationships between climate scenarios, economic consequences and architectural solutions. This study proposes an integrated simulation framework, supported by artificial intelligence, which combines climate scenarios and economic dynamics into a common architectural analysis. The new element is the introduction of the economic transformation of capital as a central analytical parameter in assessing the future of buildings.

The study asks the question: how do different climate scenarios transform the economic function of buildings, and can artificial intelligence support the systematic modelling of this transformation?

This study therefore explores the potential of AI-supported scenario simulation as a conceptual tool for linking climate trajectories, economic dynamics and architectural strategies.

2. Methodology

This study uses a qualitative, integrated simulation based on climate scenarios with a horizon until the end of the 21st century. The method is not a numerical physical simulation of building processes, nor is it a purely economic predictive model, but rather a conceptual analysis of the interrelationships between climate conditions, economic consequences and architectural solutions. The approach can be defined as exploratory – aimed at investigating structural dependencies and possible trajectories, rather than calculating precise quantitative values.

The simulation is carried out by sequentially applying the four climate scenarios (RCP 2.6, 4.5, 6.0 and 8.5) to the same

architectural object. A conditionally selected contemporary building, designed in accordance with current standards and typical economic expectations for its life cycle, is taken as the starting point. To ensure comparability, all initial design parameters, structural characteristics and functional assumptions are kept constant, with only the climate and economic environment varying in accordance with the respective scenario.

For each scenario, the following sequence is followed:

change in the physical environment → impact on operation → economic consequences → change in capital allocation → transformation of the role of the building in the urban system.

The input assumptions include published climate forecasts, generally accepted economic relationships between damage from extreme events and budget expenditures, as well as assumptions about the investment cycle in the building sector. The comparison between scenarios is made by analysing the direction, intensity and quality of structural changes, rather than by absolute numerical indicators.

Artificial intelligence is used as a tool for structuring and analysing the cause-and-effect chain and for generating visual interpretations of the results [10]. Two independent large language models (ChatGPT 5.4 [11] by OpenAI and Gemini 3 [12] by Google) were used under identical input conditions, allowing verification of the robustness of the interpretation and reducing potential model-specific biases. The role of AI is analytical and interpretative – supporting the formulation and comparison of alternative scenarios, the tracing of systemic dependencies, and the identification of possible logical contradictions or undesirable effects.

The approach does not aim to predict specific quantitative values, but to clarify structural trends and the risk of shifting investment capital from development to loss compensation. The method is research and educational in nature and aims to develop a systematic understanding of the long-term consequences of today's architectural decisions.

The methodological framework is reproducible because it is based on clearly defined input parameters, including the selected architectural object, the set of climate scenarios and a predefined cause-and-effect logic. By maintaining the same initial conditions and analytical sequence, the simulation can be repeated or applied to other building types or regional contexts. Although the results are qualitative, the logical structure of the study enables comparative verification and adaptation across different disciplinary and geographical contexts.

3. Results

As a result of the integrated simulation based on the four climate scenarios (RCP 2.6, 4.5, 6.0 and 8.5), it was investigated how the same contemporary building would function and evolve in different climatic and economic contexts until the end of the 21st century. The visualisations are not quantitative forecasts, but rather a conceptual representation of the expected structural trends resulting from the respective scenario.

Each scenario is considered as a combination of physical climatic conditions and the corresponding economic environment. Thus, the simulation allows us to track not only the impact on the construction and operation of the building, but also the change in its role in the broader urban and economic context. The differences between the scenarios are not only expressed in the degree of climate pressure, but also in the way the distribution of resources changes – from investments in development to costs for adaptation and compensation for damage.

The visualisations and analyses presented below illustrate this transformation and outline the gradual shift in architectural function – from a source of added value to an object of maintenance and protection in more adverse scenarios.



Fig. 1. Conceptual visualisation of the building under study in climate scenario RCP 2.6, generated as part of the integrated simulation.

In the RCP 2.6 scenario (see Fig. 1), the visualisation presents the building in a relatively stable climate environment, where temperature increases are limited and extreme events are less frequent and less severe. The architectural object retains its original functionality and aesthetics, with a high level of maintenance and integration of sustainable technologies in the surrounding environment – photovoltaic systems, green areas, effective awnings and resource management systems. The building functions not only as a consumer but also as an active element of sustainable urban infrastructure, without the need for large-scale adaptation or emergency interventions. The economic context allows funds to be directed primarily towards development and improvement rather than damage compensation, supporting a proactive model of growth and environmental renewal.



Fig. 2. Conceptual visualisation of the building under study in the RCP 4.5 climate scenario, generated as part of the integrated simulation.

In the RCP 4.5 scenario (see Fig. 2), the visualisation shows the building under conditions of moderately increased temperatures and more frequent extreme weather events, which already require targeted adaptation. The architectural object retains its basic structure and function, but additional sustainability measures are noticeable in the environment – enhanced sun protection screening, more intensive landscaping for cooling, rainwater collection and management systems, and reinforcement of structural elements. Some of the resources are already being directed towards increased maintenance and preventive actions, which gradually reduces the share of investments that create new value. The building remains functional and integrated into the urban environment, but the economic model is beginning to shift from entirely proactive to balanced between development and adaptation.



Fig. 3. Conceptual visualisation of the building under study in the RCP 6.0 climate scenario, generated as part of the integrated simulation.

In the RCP 6.0 scenario (see Fig. 3), the visualisation shows the building under significantly more severe climatic pressure – higher average temperatures, more frequent heat waves and more intense extreme events. The architectural object retains its structural integrity, but the environment now bears the marks of adaptation under stress: additional protective elements, reinforced structures, technical cooling and water management systems, and more limited

green spaces due to resource constraints. Part of the original architectural concept gives way to the need for functional sustainability. A significant portion of the funds is directed towards maintenance, repairs and damage prevention, which limits opportunities for development and innovation. The economy is operating in a transitional mode, in which the balance between investment and compensation is becoming increasingly fragile.



Fig. 4. Conceptual visualisation of the building under study in the RCP 8.5 climate scenario, generated as part of the integrated simulation.

In the RCP 8.5 scenario (see Fig. 4), the visualisation presents the building in an environment with high climate instability and frequent extreme events. The architecture is primarily subject to the requirements for survival and maintenance of minimum functionality. There are protective structures, temporary or emergency solutions, modified facade systems and adaptations that do not improve quality but limit risk. Green and public spaces are reduced or transformed into protective zones. A significant portion of resources is spent on damage restoration, insurance payments and replacement of existing elements, without creating new value. The economic context is dominated by compensatory costs, and growth gives way to a regime of maintenance and loss mitigation.

The analysis of the four climate scenarios shows a clear trend towards a transformation in the structure of building operating costs. As climate risk increases, the share of funds allocated to damage restoration gradually increases at the expense of investment in development and innovation. Table 1 presents a comparison of conceptual estimates of the cost structure obtained using the two independent AI models.

Table 1: Comparison of estimates of the economic cost structure between the two AI models.

RCP	GPT innovation %	GPT damage %	Gemini innovation %	Gemini damage %
2.6	75–80	20–25	80	20
4.5	55–60	40–45	60	40
6.0	35–40	60–65	40	60
8.5	10–20	80–90	15	85

The visualisations generated by the two AI models show a similar overall logic of development across the different climate scenarios, but differ in how they interpret the degree and character of architectural adaptation. In the more favourable scenarios (RCP 2.6 and 4.5), both models depict the building as well maintained and integrated into a high-quality urban environment. The main difference lies in the sustainability strategies: ChatGPT emphasises technological solutions such as energy systems, while Gemini relies more on bioclimatic approaches through vegetation and shading. As climate pressure increases in the RCP 6.0 scenario, the differences become more pronounced. ChatGPT largely preserves the original architectural form with moderate technological improvements, whereas Gemini suggests a stronger architectural transformation with a more engineering-oriented appearance. In the most severe scenario (RCP 8.5), the divergence becomes most visible: ChatGPT presents the building as still functioning under significant climatic stress, while Gemini depicts a more degraded environment and reduced maintenance. These differences indicate that although both

models follow a similar systemic logic, they interpret the scale of climate risk and the resulting architectural transformation in different ways.

4. Discussion

The simulation presented shows that the differences between climate scenarios are not limited to the physical parameters of the environment, but lead to a qualitative transformation in the economic function of capital. The main systemic effect that emerges is the gradual shift of resources from investments that create new value to costs for adaptation, restoration and damage mitigation. This transformation has long-term consequences not only for the building stock, but also for the overall economic potential of society.

In more favourable scenarios, buildings can function as platforms for technological innovation, energy transformation and improving quality of life. In more severe scenarios, they gradually become objects of protection, where a significant portion of the invested capital does not generate new productivity, but only maintains minimal functionality. This raises the question of sustainability not only as an environmental but also as an economic category.

Of particular significance is the risk of "locked-in capital" – investments in protective measures and temporary adaptations that do not increase the long-term value of assets. At the macroeconomic level, this can lead to a crowding-out effect, whereby funds are diverted from education, research and quality urban development to compensatory mechanisms. This creates a cycle in which future growth is constrained by the need to cover accumulated losses.

In this context, the role of artificial intelligence is not to predict a specific future, but to support the systematic understanding of the interrelationships between climatic, economic and architectural processes. Through scenario structuring and comparative analysis, AI can identify logical contradictions, hidden dependencies and potentially undesirable consequences of otherwise locally optimal solutions. This makes it a tool for strategic thinking, not just technical optimisation.

This study is conceptual in nature and does not claim to provide a quantitative forecast. The limitations stem from the use of qualitative simulation and the dependence on assumptions about economic dynamics in different scenarios. Nevertheless, the approach outlines a framework in which future studies can integrate quantitative models, financial analyses and real data on building operation.

In this way, the simulation serves a dual function – analytical and educational – by helping to develop a systematic understanding of the long-term consequences of current design decisions.

This approach is particularly important in the education of architecture and construction students, including at the UACEG, where future designers form their professional thinking. Understanding climate scenarios should not remain an abstract topic in the field of climatology, but should be integrated into the design process as part of the architect's long-term responsibility. The realisation that the buildings designed today will operate in a different climatic and economic environment draws attention to sustainability as a strategic choice rather than a regulatory requirement. Including such simulations in the learning process helps students think beyond current regulations and investment frameworks, assess the risk of premature obsolescence of buildings, and seek solutions that reduce future adaptation costs. In this way, education becomes a tool for reducing future systemic risks through more conscious design today.

The development of more complex artificial intelligence models opens up the possibility for such integrated simulations to gradually move from qualitative to hybrid and subsequently to quantitative modelling. With reliable climate forecasts, real operational data

from the building stock, insurance statistics and macroeconomic indicators, systems can be built that combine physical simulations of building behaviour, econometric dependencies and multi-criteria optimisation in a unified analytical framework.

Integration with the concept of digital twins of buildings and the urban environment is particularly promising. By linking real sensor data, operational indicators and climate forecasts, the digital twin could function as a dynamic platform for strategic planning, not just operational control. In this context, artificial intelligence can support not only the analysis but also the simulation of long-term investment strategies, including the assessment of trade-offs between initial costs and future adaptation needs.

It is also possible to use agent-based models that simulate the behaviour of different actors – investors, public institutions, insurers, residents – under different climatic and economic conditions. This would allow for a better understanding of systemic dynamics and potential non-linear effects, where locally rational decisions lead to globally unfavourable outcomes.

At the same time, the paradox that the development of artificial intelligence itself is associated with increasing energy consumption and carbon footprint should also be taken into account. Therefore, future integrated modelling systems should be developed with attention to their energy efficiency and real contribution to reducing systemic risk, rather than as an additional source of stress on the climate system.

In this sense, the present study can be seen as a conceptual basis for more complex interdisciplinary models that integrate architecture, climatology, economics and artificial intelligence into a common framework for strategic planning. Expanding such simulations would allow not only for more accurate assessments, but also for earlier identification of critical points at which the system transitions from a development mode to a loss compensation mode.

The time aspect of decision-making is also of additional importance. Architectural and investment choices are made in the short term – within a project cycle or investment framework of 5–10 years, while climate processes and economic transformations unfold over decades. It is precisely this time gap that creates systemic risk: decisions that are economically rational today may prove structurally vulnerable within the life cycle of a building. In this context, artificial intelligence can function as a tool for extending the time horizon of analysis, allowing for a comparison between short-term returns and long-term sustainability. In this way, AI becomes a means of reducing strategic myopia and integrating future risks into current design.

5. Conclusion

This study examines artificial intelligence as a tool for integrated simulation of the links between climate scenarios, economic dynamics and architectural solutions. The analysis shows that different climate trajectories lead not only to changes in the physical environment, but also to a fundamental transformation in the distribution of capital – from investments that create new value to costs for adaptation and compensation for damage in more adverse scenarios. In this way, climate processes become not only an environmental but also a structural economic driver that changes the logic of development in the building sector.

This change directly affects buildings designed today, whose life cycle will cover the period of the most intense climatic and economic transformations. The risk of premature obsolescence and "locked-in capital" calls for longer-term and more systemic thinking in the design process.

The proposed approach demonstrates that artificial intelligence can be used not only for technical optimisation, but also as a tool for strategic thinking about the future of the built environment. By integrating climate and economic scenarios into architectural analysis, it is possible to make more informed decisions today,

based on an extended time horizon and an assessment of systemic risk. Such an approach is important both for professional practice and for the training of future designers, as it encourages the development of thinking that links climate, economics and architecture within a single logical framework.

More broadly, artificial intelligence makes it possible to identify earlier the critical points at which the system shifts from development to loss compensation, helping to support more sustainable development trajectories.

6. References

1. Meinshausen M, Smith SJ, Calvin K, Daniel JS, Kainuma M, Lamarque JF, et al. The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*. 2011;109(1–2):213–241. <https://doi.org/10.1007/s10584-011-0156-z>
2. van Vuuren DP, Edmonds J, Kainuma M, Riahi K, Thomson A, Hibbard K, et al. The representative concentration pathways: an overview. *Climatic Change*. 2011;109(1–2):5–31. <https://doi.org/10.1007/s10584-011-0148-z>
3. IPCC. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press; 2021. <https://doi.org/10.1017/9781009157896>
4. IPCC. Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press; 2022. <https://doi.org/10.1017/9781009325844>
5. Choi Y, Ray A, Smith AD, Hain A. Assessment of building energy demand and photovoltaics for urban decarbonization. *Applied Energy*. 2021; 292: 116880. <https://doi.org/10.1016/j.apenergy.2021.116880>
6. Dougherty T, Jain RK. Microclimate effects on building energy consumption: A data-driven analysis using machine learning. *Energy and Buildings*. 2022;255:111676. <https://doi.org/10.1016/j.enbuild.2021.111676>
7. Watkiss P. The Costs of Adaptation and the Economic Costs and Benefits of Adaptation. London: Climate Change Committee; 2023. Available from: <https://www.theccc.org.uk>
8. Felicioni L, Khosravi M, Jenkins D. Economic assessment of building adaptation to climate change: A systematic review of cost evaluation methods. *Sustainable Cities and Society*. 2025;109:105485. <https://doi.org/10.1016/j.scs.2024.105485>
9. Bilal A, Känzig D. The Macroeconomic Impact of Climate Change: Global vs. Local Temperature Effects. *American Economic Journal: Macroeconomics*. 2024;16(3):1–36. <https://doi.org/10.1257/mac.20220173>
10. Buster G, Mendes A, Silva C. Machine learning approaches for urban climate modelling and heat mitigation strategies. *Sustainable Cities and Society*. 2024;100:104989. <https://doi.org/10.1016/j.scs.2023.104989>
11. Conversation with ChatGPT 5.4: <https://chatgpt.com/share/69afd74a-8864-8012-af61-fa0aa474b565>
12. Conversation with Gemini 3: <https://gemini.google.com/share/1f42da8f138e>