# Evaluating the potential of LLMs as analysis-support tools in the prioritization of critical infrastructures threats

**Ana-Maria CHISEGA-NEGRILĂ, Petrișor PĂTRAȘCU***

The Command and Staff Faculties "Carol I" National Defence University, Bucharest, Romania

anachisega@yahoo.com, patrascupetrisor@yahoo.com (*Corresponding author)

**Abstract:** Critical infrastructure protection represents one of the greatest challenges of the past decade due to new cyber and physical threats, as well as those generated by climatic, geopolitical, or other forms of instability. This complex of threats can disrupt interconnected and complex systems, leading to cascading failures that can no longer be easily remedied. Therefore, critical infrastructures, which include essential systems aiming at economic activities, public health, and national security, should be analyzed not only from the perspective of reliability, but also from that of their capacity to adapt in case of a major crisis. The study aims to analyze the capability of large language models (LLMs) to prioritize threats to critical infrastructures and to compare these assessments with those made by certified experts in the field. The analysis is based on a set of ten threats and three criteria: the effect on society, economic consequences, and potential casualties. These are ranked through three different LLM models (ChatGPT, Gemini, and Claude), using repeated iterations, while in parallel, the same set of threats is evaluated by a sample of human specialists. Through this approach, it is intended to determine the points where these hierarchies converge, as well as their points of divergence, in order to understand to what extent artificial intelligence can be integrated into decision-support mechanisms designed for resilience and crisis management.

**Keywords:** Artificial Intelligence, Critical infrastructure, Decision-making process, LLMs, AI evaluation.

# Evaluarea potențialului LLM-urilor ca instrumente de analiză în prioritizarea amenințărilor la adresa infrastructurilor critice

**Rezumat:** Protecția infrastructurilor critice reprezintă una dintre cele mai mari provocări ale ultimului deceniu, datorită noilor amenințări cibernetice și fizice, dar și a celor generate de instabilitatea climatică, geopolitică sau de altă natură. Acest complex de amenințări poate perturba sistemele interconectate și complexe, conducând la defecțiuni în lanț care nu mai pot fi remediate cu ușurință. De aceea, infrastructurile critice, care includ sistemele esențiale ce vizează activitățile economice, de sănătate publică și de siguranță națională, ar trebui analizate nu numai din punctul de vedere al fiabilității, ci și din cel al capacității de adaptare în caz de criză majoră. Studiul are ca scop analiza capacității modelelor de limbaj de mari dimensiuni (LLM-uri) de a prioritiza amenințările la adresa infrastructurilor critice și compararea acestor evaluări cu cele realizate de experți certificați în domeniu. Analiza se bazează pe un set de zece amenințări și pe trei criterii: efectul asupra societății, consecințele economice și posibilele victime. Acestea sunt ierarhizate prin intermediul a trei modele LLM diferite (ChatGPT, Gemini și Claude), folosind iterații repetate, iar în paralel, același set de amenințări este evaluat de un eșantion de specialiști umani. Prin această abordare, se dorește determinarea punctelor în care aceste ierarhii converg, cât și a punctele lor de divergență, pentru a înțelege în ce măsură inteligența artificială poate fi integrată în mecanismele de suport decizional destinate rezilienței și managementului crizelor.

**Cuvinte cheie:** Inteligență Artificială, infrastructură critică, proces decizional, LLM-uri, evaluarea inteligenței artificiale.

## 1. Introduction

In the third decade of the twenty-first century, global security is no longer shaped only by isolated actors or single events; instead, the international community faces a complex crisis due to the combined effects of cyber and physical attacks and drastic climate changes. As Davis et al. (2024) note, the combined effects of a global pandemic, a potential large-scale military invasion in Europe,

https://doi.org/10.33436/v36i2y202602

and the resulting international energy crisis could provide a concrete example of a volatile environment in which multiple threats reinforce one another, increasing the vulnerability of critical infrastructures. Critical infrastructure (CI) is defined as a set of systems and networks necessary to support vital societal functions, including public health, safety, security, and economic activity (Alcaraz & Zeadally, 2015).

Due to these risks, legislation has adapted in order to become less focused on local issues and to adopt a broader perspective with cross-border applicability. In the European Union, the "Critical Entities Resilience (CER) Directive" (European Parliament and Council of the European Union, 2022a) and the "NIS2 Directive" (European Parliament and Council of the European Union, 2022b) form a framework that emphasizes physical and digital security requirements across eleven essential sectors. These instruments require organizations to conduct comprehensive risk assessments that should take into account the interdependencies between physical infrastructure and the digital systems that operate them.

The unprecedented technological progress of the past decade has created the need to align the protection of CI with AI systems, as personnel working in key areas have increasingly begun to interact with artificial intelligence systems. This interaction can be beneficial, since AI can analyze large volumes of data and provide valuable information for the decision-making process; however, the same automation, in the absence of adequate human oversight, can also lead to events that are difficult to anticipate and manage (Sarker et al., 2024). Within this context, the purpose of this article is to evaluate how LLMs can prioritize CI threats based on criteria defined by the authors, and whether their outputs resemble those produced by a group of CI experts, in order to assess whether these models can be reliably integrated as decision-support tools for real-world risk management.

## 2. Literature review

CIs play a vital role in modern society, so their reliability, performance, continuous operation, safety, maintenance, and protection have become national priorities for many states (Alcaraz & Zeadally, 2015). By providing essential goods and services, CIs are interdependent, and the failure of one may affect the functionality of others that depend on it. The resulting domino effect can result in disruptions or, more importantly, in an escalation that can lead to major crises (Iturriza et al., 2018). In order to avoid such situations, states need to identify all types of threats and, among them, those with a higher degree of risk (De Felice, Baffo, & Petrillo, 2022). Over the past decade, the list of risks and threats to CIs has expanded, complementing traditional risks such as natural disasters and human errors with emerging and unpredictable threats, such as cyberattacks (Markopoulou & Papakonstantinou, 2021). At the same time, threats to CIs are considered threats to national security (Żaboklicka, 2020; Pătrașcu, 2022). Although complete protection of CI cannot be guaranteed and the costs are not always efficient in relation to actual threats (Pursiainen, 2017), critical infrastructures require protection through the implementation of preventive and reactive actions by state authorities and by the owners or operators of critical infrastructures. In fact, the development of resilient systems provides adaptability to change and solutions in unforeseen situations (Osei-Kyei et al., 2021). Moreover, the concept of resilience is used to improve the performance of CI and can, in turn, be further enhanced itself (Mottahedi et al., 2021).

In recent years, artificial intelligence and large language models (LLMs) have become widely used in society, taking over areas such as education, healthcare, global freight of food and goods, and telecommunication (Chaudhry et al., 2018). As the interaction between AI and humans is relatively recent, including in the field of CI, both members of this partnership can have difficulties understanding each other. Currently, there is an enthusiasm, whether justified or not, about AI's ability to perform tasks better than its human partners in almost any situation. However, this belief is coupled with a deep distrust of these LLMs, which remain opaque systems to the general public. Although they have been trained on data, articles, and scenarios created and provided by humans, it is still debated whether the way in which the AI understands the world aligns with the way we perceive it, and to what extent. Humans interact with the surrounding environment based on a set of information and beliefs that help them make correct decisions to the extent that these correspond to

the reality of the world. This implies the existence of models built on probabilistic data, which, however, must be updated as new information appears (Qiu et al., 2026). Bayesian theory can be used for the aggregation and refinement of outputs produced by several large language models (LLMs) involved in a process of ranking threats to CI, if the risk level of each threat is not treated as a variable. Since each LLM can be considered an imperfect source of information that provides partial estimates on the severity of a threat in relation to a set of predefined criteria, if disagreements between LLMs are not treated as errors that must be eliminated, but as sources of uncertainty that contribute to the shape of the final distribution, these estimates can be interpreted as observations that contribute to establishing the real level of risk of the threat.

The use of LLMs in the decision-making process must also take into account the existence of systemic problems that can have a negative impact on the result, such as the one related to aleatoric uncertainty, which comes from the random character of the data or the environment (Felicioni et al., 2024). Also, LLMs tend to establish a fixed hypothesis or explanation at the beginning of a dialogue, which they will not later reanalyze. There are also other problems, such as the presence of biases or alignment with the objectives and beliefs of the user (Ferrara, 2023). Therefore, for an LLM to provide the best possible answer to a request, frameworks such as DeLLMa are used, which is designed to improve the performance of LLMs when they are used for decision-making governed by uncertainty and where classical direct prompting methods often lead to poor results, especially as the complexity of the problems increases. DeLLMa introduces a structured reasoning process, inspired by decision theory and utility theory, which decomposes the decision into several steps, thus transforming LLMs from reactive systems of response generation into decision systems capable of integrating uncertainty and the user's objectives in a coherent and auditable way. Empirical results show that such a framework can significantly increase decision accuracy (up to approximately 40%), demonstrating that the integration of principles from decision theory improves both performance and the transparency of reasoning in different applications such as finance or agriculture (Liu et al., 2024).

The monitoring and control of multiple CIs, such as energy, water distribution, sewage systems, and production control, are carried out through SCADA (Supervisory Control and Data Acquisition) and DCS (Distributed Control Systems) systems, often referred to as Operational Technologies (OT). The evolution of technologies and the development of data-driven and remote operations have produced a convergence between OT systems and IT systems, followed by interconnection from the perspective of the Internet and cloud facilities. Consequently, these characteristics entail vulnerabilities to cyberattacks, as well as cybersecurity that is more oriented toward operational technologies (Murray et al., 2018). At the same time, the IT and OT domains are conceptually framed as parts of an Industrial Control System (ICS). The IT component provides general operational services through workstations, servers, and databases, all interconnected via IP-based networks. In contrast, the OT component provides machinery-specific operational services through optimized systems, such as Programmable Logic Controllers (PLCs) and Variable-Frequency Drives (VFDs), manufactured to withstand the test of time (Makrakis et al., 2021). Across multiple critical infrastructure sectors, ICS systems are encountered that facilitate the functionality of critical infrastructures. Taking into account the essential services delivered, the convergence and interconnectivity between OT and IT systems, and the value of trade secrets and technological patents held, critical infrastructure owners and operators focus on ensuring the cybersecurity of OT and IT systems against cyberattacks and sabotage.

In a holistic study on cyber incidents, vulnerabilities, and threats targeting ICS and critical infrastructures, the authors produce a categorization of vulnerabilities based on the analysis of attack methodologies, comprising: zero-day vulnerabilities (Type 0); known vulnerabilities (Type 1); vulnerabilities stemming from inherently insecure services and protocols (Type 2); vulnerabilities relevant to insecure configuration of networks and equipment (Type 3); and vulnerabilities relevant to insecure configuration (Type 4). In the same context, the authors present the main ICS-specific protocols, such as DNP3, IEC-104, IEC-61850, Modbus, PROFINET, ZigBee, WirelessHART, EtherNet/IP, and OPC, along with the identified vulnerabilities of these protocols (Makrakis et al., 2021). Moving toward Large Language Models (LLMs) as support for the cybersecurity of industrial control systems, existing evaluation frameworks are directed more toward IT systems than OT systems. These shortcomings are addressed through evaluation frameworks, a good example being

CritBench, an innovative framework designed to assess the cybersecurity capabilities of LLM agents in IEC 61850 Digital Substation environments (Keppler et al., 2026).

# 3. Method

## 3.1. Problem selection and purpose

The purpose of the current selection was to test whether Large Language Models (LLMs) and human experts produce similar results in risk assessment across a diverse range of threats, including natural disasters, technological failures, intentional attacks, and economic disruptions. In order to analyze how AI systems perceive and rank risk, a series of threats targeting critical infrastructure was selected and presented for hierarchization to multiple AI tools first, and subsequently to a group of critical infrastructure specialists. The ten threats selected and their abbreviations are presented in Table 1.

**Table 1.** Threats and abbreviations

| Threats | Abbreviation |
|---|---|
| Cyber threats (Infrastructure) | T 1 |
| Hybrid attacks and disinformation | T 2 |
| Extreme weather events | T 3 |
| Energy outages or grid instability | T 4 |
| Supply chain disruption | T 5 |
| Pandemics or biological emergencies | T 6 |
| Terrorist threats | T 7 |
| Organized crime and AI-enabled fraud | T 8 |
| Aging physical infrastructure | T 9 |
| Space weather events (solar storms) | T 10 |

To guide this inquiry, this article seeks to answer two central research questions:

RQ1: What is the statistical correlation between the aggregated threat assessments generated by LLMs and those provided by certified liaison officers for national critical infrastructure?

RQ2: To what extent do LLM non-determinism and model generational evolution affect the reliability and stability of threat classifications across repeated studies?

## 3.2. Model selection

To obtain a comparative perspective on the analytical capacity of LLMs in the field of critical infrastructure, this study utilizes three of the most advanced language models currently available: Gemini 3.1 Pro, ChatGPT GPT-5.5, and Claude Sonnet 4.6, in their May–June 2026 versions. Interaction with all three models was conducted exclusively in English. This design choice was made to better leverage the primary corpus of these systems' training data, which is predominantly in English.

## 3.3. Prompt design and evaluation criteria

For each of the selected threats, the AI systems were tasked, through specific prompts, with providing three distinct impact scores on a scale from 1 to 10. These scores were associated with the three primary evaluation criteria: the potential number of victims, as well as economic and material losses. Although many states have established their own sets of criteria for defining critical infrastructures (da Silva et al., 2025), the motivation for choosing these three specific indicators is rooted in European policy and in the criteria established by the European Programme for Critical Infrastructure Protection for the purpose of identifying European critical infrastructures (Council of the European Union, 2008), which formed the basis for the identification of national critical

infrastructures, serving as tangible and quantifiable indicators for EU Member States (Trinchillo, 2025).

The experiment is based on a prompt that has been designed to assign the models the role of a CIP (Critical Infrastructure Protection) Risk Analyst, but still limiting their free responses. The LLMs were asked to evaluate 10 predefined threats (T1–T10) based on three criteria, each detailed on a scale of 1 to 10:

- Public effects: Measures social stability, trust, and service continuity. The scale ranges from minor localized discomfort (Score 1) up to total loss of public trust and structural societal instability at a global level (Score 10).
- Economic effects: Measures the impact of market volatility and supply chain costs. It ranges from financial losses (Score 1) up to global market collapse and the disintegration of international trade networks (Score 10).
- Human casualties: Measures mortality and serious physical injury. It ranges from zero deaths and minor injuries (Score 1) up to widespread mortality exceeding 100,000 victims (Score 10).

The models were instructed to calculate a final score based on the mathematical average of the three criteria, thereby enforcing a logical and mathematical rigor in the generated response. Also, to measure the models' capacity for internal reasoning rather than their ability to synthesize news, the following limitations were introduced.

The models were instructed not to use internet search tools, external databases, or the retrieval of real-time event information. The purpose of the research is to test the internal knowledge base of the models and their ability to apply expert structural reasoning; if internet access had been permitted, the classification would have transformed into a "search-and-summarize" exercise. By imposing a closed-book regime and prohibiting speculations, the experiment tests the alignment of LLMs with human professional standards in the security field, verifying whether the model can distinguish a structural vulnerability from a theoretical threat lacking historical precedent.

- Data Sufficiency: The models were required to use the phrase "Data Insufficient for Assessment" if they lacked verifiable historical data, instead of inventing plausible scenarios.
- Consistency: The models were prohibited from adjusting their relative scores (to create an artificial hierarchy) or modifying the interpretation of the grid during the generation of the response.

Furthermore, to ensure a uniform evaluation, the models were provided with a specific global context, that of a modern industrialized society (reference 2020-present) characterized by infrastructure interdependence, aging physical networks, normal climate variability, and a baseline level of cyber activity. The scores provided should strictly reflect the severity of the structural impact on society, focusing on the consequences of the threat manifesting rather than its probability. Additionally, the models were required to perform an internal absolute assessment, to assign scores, and then to display the mathematical calculations before generating the final table.

To ensure statistical validity and reproducibility of the results, the experiment is not based on a single generation but on 10 distinct iterations for each of the three models, carried out on different days and in different chat sessions. This was done to prevent, on the one hand, the tendency of a model to be self-consistent within the same session, where results after the first iteration would be biased by the previous ones. On the other hand, running tests on different days prevents the API provider from delivering a cached response for identical prompts sent successively, and also, due to the probabilistic nature of LLMs, testing at different times allows for the identification of the model's natural variation influenced by internal temperature and server decoding settings. The LLMs were queried through the web interface with standard temperature settings; however, the clear definitions for the 1 to 10 scale and the constraints aimed at reducing the AI's tendency to improvise to a certain extent.

## 3.4. Constraint violations

Even if the prompt gave the LLMs explicit instructions, the following violations were recorded, the most common being the fact that they did not flag "Data Insufficient for Assessment" for the required period of time, and expanded it so that they would provide examples and accomplish the task (Table 2):

**Table 2.** Constraint violations

| Common Violation | Gemini | ChatGPT | Claude |
|---|---|---|---|
| Deficient Rubric Mapping (Failure to explicitly align the reasoning text to the exact numeric criteria thresholds) | Used generalized phrasing and implied mapping instead of explicitly mirroring the rubric's wording for the assigned scores. | Described impacts generally but failed to explicitly cite the text of the rubric anchors, relying entirely on implied alignment. | Blended general historical context into the bullets rather than strictly disciplining each bullet to match a specific numeric criterion. |
| Inference Over Strict Evidence (Substituting infrastructure risk modelling or speculation where direct historical data was missing) | Assigned arbitrary low scores (like 2) to undocumented categories (e.g., T8 casualties) instead of executing the mandatory "Data Insufficient" fallback. | Scored Space Weather and Cyber using predictive engineering models and plausible cascading potential rather than realized historical outcomes. | Admitted to inflating the data profile to produce a "richer" output, blending generalized context where explicit historical anchors were missing. |
| Output Format Contamination (Injecting unrequested structural layers, headers, or commentary outside the explicit schema) | Injected unauthorized section headers (PHASE 1 & 2, PHASE 3) and truncated the mandatory text string for the AIS formula. | Formatted confidence levels qualitatively and conceptually without an explicit deterministic rule, while omitting strict anchor constraints. | Appended an unrequested "Key structural observations" commentary section and introduced an unauthorized interactive widget. |

## 3.5. Method application

First, the threats were selected by querying several LLM systems (ChatGPT, Claude, Gemini), and the results (T1–T10) were incorporated into a questionnaire, broken down according to the three criteria: public effects, economic effects, and casualties. The questionnaire was administered between December 2025 and January 2026 to a group of 19 specialists from Romania, chosen by the authors based on their certification as Liaison Officers for the Security of National Critical Infrastructures (Occupational Standard 242227). Respondents were asked to assign a value from 1 to 10 to each threat, where 1 represented the lowest value and 10 the highest, with the possibility of assigning the same value to multiple threats. The order of the threats from T1 to T10 was maintained for all specialists and across all LLM systems. After the artificial intelligence models and the human specialists had completed the task independently of one another, the results were compared in order to determine whether any recognizable patterns could be identified.

The authors first analyzed the results provided by the specialists and then compared them with the results provided by the AI models to see whether they shared a common logic. This step was essential in highlighting whether AI systems tend to over-prioritize popular threats frequently encountered in news reports, such as pandemics or AI-assisted fraud, while at the same time underestimating other risks that human experts consider more dangerous for long-term operational stability.

# 4. Results

## 4.1. Analysis of specialist results

The data obtained from the specialists were compiled in Table 3, aggregated from the 19 questionnaires, and organized according to the three criteria specific to the impact (C1 - Public effects, C2 - Economic effects, C3 – Casualties) and the 10 types of threats. Table 3 was filled in with the average of the values assigned by the 19 respondents for each individual criterion, for each of the 10 threats, from T1 to T10.

After that, to highlight the degree of consensus of the human experts, the standard deviation (SD) was applied for each threat. The next step consisted of determining the general average of the impact, from the values corresponding to the three criteria, then calculating the standard deviation for all three criteria, and finally determining the normalized average by applying the Z-Score method for the 10 threats. Looking at the standard deviation values, a relatively narrow dispersion of the 19 respondents' answers around the mean is highlighted, a fact that denotes that most specialists assigned close scores. The deviation values on a scale from 1 to 10 are relatively small, which shows favourable and grouped opinions, and with a moderate level of consensus among the group of specialists, a consensus justified to some extent by the diversity of infrastructures and areas of expertise represented by these specialists. Their opinions are not strongly polarized, which offers a stable framework for a small-scale study without major differences of opinion. For the Extreme Weather Events (T3) threat, the answers tend to be positive and located in the upper part of the scale, with an average of 7.63, but the standard deviation of 1.76 shows a moderate dispersion of opinions, which means that not all respondents had the same opinion. In contrast, for the Pandemic (T6) threat, the statistical data show a very homogeneous group, with similar and favourable opinions. In the last column of Table 3, the standardized scores are calculated using the Z-Score method. Thus, threats T1, T2, T3, T4, T6 (this one having the highest value), and T9 are above average, while threats T5, T7, T8, and T10 are below average, the latter having the lowest value.

**Table 3.** Results according to specialists

| Criteria / Threats | C1 | | C2 | | C3 | | C1,2,3 | C1,2,3 | Z-Score |
|---|---|---|---|---|---|---|---|---|---|
| | **Mean** | **SD** | **Mean** | **SD** | **Mean** | **SD** | **Mean** | **SD** | |
| **T 1** | 6.00 | 2.27 | 9.00 | 1.08 | 8.53 | 1.04 | 7.84 | 1.12 | 0.62 |
| **T 2** | 6.16 | 2.25 | 8.42 | 0.99 | 8.37 | 1.46 | 7.65 | 1.31 | 0.38 |
| **T 3** | 7.32 | 1.81 | 7.63 | 1.84 | 7.95 | 2.09 | 7.63 | 1.76 | 0.28 |
| **T 4** | 5.58 | 2.30 | 8.37 | 1.42 | 8.53 | 1.63 | 7.49 | 1.43 | 0.24 |
| **T 5** | 3.89 | 1.94 | 7.74 | 1.74 | 6.42 | 1.93 | 6.02 | 1.30 | -0.87 |
| **T 6** | 8.53 | 1.98 | 7.95 | 1.67 | 9.21 | 1.15 | 8.56 | 1.36 | 1.04 |
| **T 7** | 6.16 | 2.68 | 6.11 | 2.10 | 6.21 | 1.44 | 6.16 | 1.65 | -0.60 |
| **T 8** | 4.89 | 2.40 | 8.11 | 1.37 | 7.37 | 1.84 | 6.79 | 1.48 | -0.24 |
| **T 9** | 6.84 | 2.08 | 8.00 | 1.41 | 7.89 | 1.68 | 7.58 | 1.42 | 0.31 |
| **T 10** | 3.74 | 2.20 | 6.74 | 1.65 | 6.74 | 1.65 | 5.74 | 1.34 | -1.05 |

In the same context, Figure 1 highlights the standard deviation of threats according to the three criteria. In this graph, the standard deviation was applied for the common criterion (C1,2,3) of the impact resulting from the three analyzed criteria, C1, C2, and C3, and as a benchmark for them. Thus, by analyzing the error bars, it is observed that for the Public effects (C1) criterion, threats T1, T4, T5, T8, and T10 fall outside the deviation interval. On the other hand, three compact groupings for all three criteria are highlighted for the threats T3, T6, and T9.

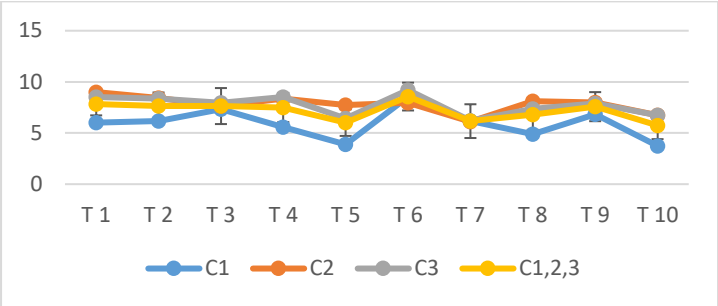


**Figure 1.** Standard deviation and error bars for the specialists' data

## 4.2. Analysis of LLMs results

The following section presents the results obtained through LLMs. The results were generated by three artificial intelligence systems (ChatGPT, Claude, Gemini), following the same rule of assigning values to threats on a scale from 1 to 10 and on the same criteria: C1 – Public effects, C2 – Economic effects, C3 – Casualties. In this regard, the three platforms were queried for a number of iterations (N=10). The resulting data were entered into tables and analyzed through the lens of standard deviations and average scores.

In Table 4, the means and standard deviations calculated for the data generated by the ChatGPT artificial intelligence system over 10 iterations are presented. An analysis of the standard deviation values reveals that all 10 threats show a low degree of variability, indicating that the data is concentrated around the mean (6.42). Slight tendencies toward moderation can be seen in threats T2 and T10, reflecting minor variations between iterations. The highest rating is identified for T6, where a clear clustering of data across the iterations is observed.

**Table 4.** Results according to ChatGPT

| Criteria / Threats | C1 | | C2 | | C3 | | C1,2,3 | C1,2,3 |
|---|---|---|---|---|---|---|---|---|
| | **Mean** | **SD** | **Mean** | **SD** | **Mean** | **SD** | **Mean** | **SD** |
| **T 1** | 8 | 0.00 | 7.5 | 0.92 | 4 | 0.45 | 6.50 | 1.88 |
| **T 2** | 7.7 | 0.78 | 5 | 1.10 | 2.2 | 0.60 | 4.97 | 2.40 |
| **T 3** | 8 | 0.00 | 7.8 | 0.40 | 7.8 | 0.40 | 7.87 | 0.34 |
| **T 4** | 7.9 | 0.30 | 6.9 | 0.70 | 5.3 | 0.46 | 6.70 | 1.19 |
| **T 5** | 6.9 | 0.30 | 7.7 | 1.27 | 3.8 | 0.40 | 6.13 | 1.86 |
| **T 6** | 9.1 | 0.30 | 9 | 0.45 | 10 | 0.00 | 9.37 | 0.55 |
| **T 7** | 6.2 | 0.40 | 5.1 | 0.70 | 6.5 | 0.81 | 5.93 | 0.89 |
| **T 8** | 5 | 0.00 | 5.5 | 1.50 | 1.7 | 0.46 | 4.07 | 1.91 |
| **T 9** | 6.7 | 0.46 | 6.6 | 0.66 | 4.5 | 0.50 | 5.93 | 1.15 |
| **T 10** | 8.1 | 0.30 | 7.8 | 1.33 | 4.2 | 1.33 | 6.70 | 2.08 |

The means resulting from the repeated querying of the artificial intelligence system ChatGPT are graphically represented in Figure 2, across the three criteria. The standard deviation for all three criteria (C1, C2, and C3) was taken as a common benchmark, and the error bars were applied. Therefore, a strong homogeneity is observed for threats T6 and T3, framed within the error bars for T4 and T7. On the other hand, a slight deviation can be observed for more than half of the specific threats of criterion C3.

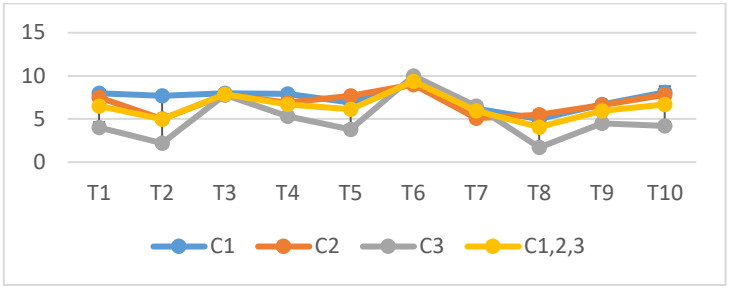


**Figure 2.** Standard deviation and error bars according to ChatGPT

In Table 5, the means and standard deviations calculated for the data generated by the Claude artificial intelligence system over 10 iterations are presented. An analysis of the standard deviation values reveals a low degree of variability across all 10 threats. The lowest standard deviations (SD) are found in threats T3, T6, and T7. Furthermore, threat T7 is associated with a mean that is very close to the overall mean of the 10 threats (6.38), in contrast to T6, where the scores are very close to the maximum value of the scale, indicating a high level of satisfaction. The highest standard deviations are observed in T1, T5, and T8. The first two have scores close to the average score, whereas the score for T8 tends toward the lower end of the scale.

**Table 5.** Results according to Claude

| Criteria / Threats | C1 | | C2 | | C3 | | C1,2,3 | C1,2,3 |
|---|---|---|---|---|---|---|---|---|
| | **Mean** | **SD** | **Mean** | **SD** | **Mean** | **SD** | **Mean** | **SD** |
| **T 1** | 7.2 | 0.40 | 7.7 | 0.46 | 4 | 0.63 | 6.30 | 1.72 |
| **T 2** | 6.7 | 0.46 | 4.9 | 0.30 | 3.3 | 0.64 | 4.97 | 1.47 |
| **T 3** | 7.2 | 0.60 | 7.2 | 0.60 | 7.7 | 0.64 | 7.37 | 0.66 |
| **T 4** | 7.2 | 0.60 | 6.9 | 0.30 | 5.4 | 0.80 | 6.50 | 0.99 |
| **T 5** | 6.7 | 0.78 | 8.3 | 0.46 | 4 | 0.77 | 6.33 | 1.90 |
| **T 6** | 9.1 | 0.30 | 9.1 | 0.30 | 10 | 0.00 | 9.40 | 0.49 |
| **T 7** | 6.3 | 0.46 | 6.1 | 0.54 | 6.2 | 0.75 | 6.20 | 0.60 |
| **T 8** | 4.7 | 0.46 | 6.7 | 0.46 | 2.2 | 0.40 | 4.53 | 1.89 |
| **T 9** | 5.6 | 0.49 | 6 | 0.63 | 4.2 | 0.40 | 5.27 | 0.93 |
| **T 10** | 7.8 | 0.60 | 7.6 | 0.49 | 5.3 | 1.00 | 6.90 | 1.35 |

The means resulting from the repeated querying of the Claude artificial intelligence system are graphically represented in Figure 3 across the three criteria. The standard deviation for all three criteria (C1, C2, C3) was taken as a common baseline to apply the error bars. Consequently, a strong homogeneity is observed for threats T3 and T7, whereas a random distribution within the limits of the error bars is noted for T4, T6, and T9. In contrast, a slight deviation can be observed for four of the threats specific to criterion C3, as well as one for criterion C1 at threat T2, and one for criterion C2 at threat T8.

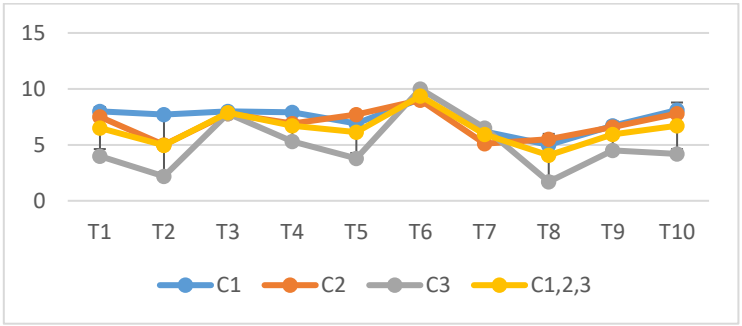


**Figure 3.** Standard deviation and error bars according to Claude

In Table 6, the means and standard deviations calculated for the data generated by the Gemini artificial intelligence system, also run over 10 iterations, are presented. An analysis of the standard deviation values reveals a low degree of variability for most of the threats, with minor tendencies toward moderation observed in threats T1 and T5. The lowest standard deviations (SD) are found in threats T3, T4, T6, and T9. Furthermore, threat T4 is associated with a score that is very close to the overall average score of the 10 threats (5.09), in contrast to T6, where the score is very close to the upper limit of the scale, indicating a high level of satisfaction. The highest standard deviations are observed in T1 and T5, though their scores remain quite close to the average score.

**Table 6.** Results according to Gemini

| Criteria / Threats | C1 | | C2 | | C3 | | Mean | SD |
|---|---|---|---|---|---|---|---|---|
| | **Mean** | **SD** | **Mean** | **SD** | **Mean** | **SD** | | |
| **T 1** | 6.1 | 1.37 | 6.3 | 0.90 | 1.8 | 0.87 | 4.73 | 2.34 |
| **T 2** | 5.5 | 2.06 | 3.8 | 0.87 | 2.5 | 1.20 | 3.93 | 1.91 |
| **T 3** | 5.7 | 0.46 | 6.2 | 0.40 | 6.3 | 0.64 | 6.07 | 0.57 |
| **T 4** | 5.8 | 0.40 | 5.9 | 0.54 | 4.8 | 0.40 | 5.50 | 0.67 |
| **T 5** | 4.4 | 1.56 | 7.1 | 0.94 | 2.2 | 1.33 | 4.57 | 2.39 |
| **T 6** | 8.8 | 0.60 | 9 | 0.00 | 10 | 0.00 | 9.27 | 0.63 |
| **T 7** | 6.8 | 0.98 | 6.3 | 1.35 | 6.8 | 0.60 | 6.63 | 1.05 |
| **T 8** | 3.2 | 1.72 | 4.2 | 1.40 | 1 | 0.00 | 2.80 | 1.85 |
| **T 9** | 4.3 | 0.64 | 4.2 | 0.60 | 4.1 | 0.30 | 4.20 | 0.54 |
| **T 10** | 4.2 | 1.47 | 4.4 | 1.36 | 1 | 0.00 | 3.20 | 1.94 |

The means resulting from the repeated querying of the Gemini artificial intelligence system are graphically illustrated in Figure 4 across the three criteria. The standard deviation for all three criteria (C1, C2, C3) was taken as a common baseline to apply the error bars. Consequently, a strong homogeneity is observed for threats T3, T7, and T9, whereas a random distribution within the limits of the error bars is noted for T2, T4, T5, T6, and T8. In contrast, a slight deviation can be observed for two of the threats, T1 and T10, which are specific to criterion C3.

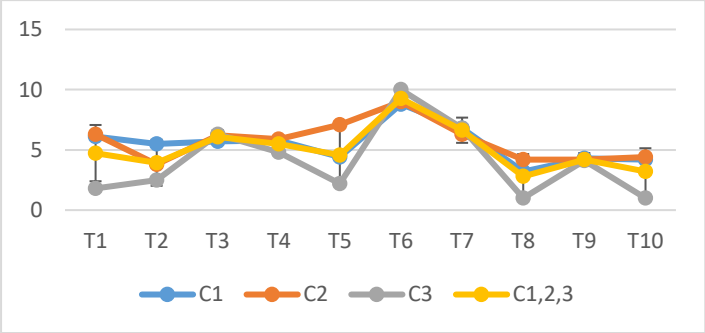


**Figure 4.** Standard deviation and error bars according to Gemini

## 4.3. External validation of results

To demonstrate the decision-making independence of LLM algorithms and the fact that the threat ranking was not based on an existing internet hierarchy reflecting users' search preferences, a statistical correlation test was performed for each of the results provided by the three LLMs, using data extracted from the Google Trends platform. Initially, the use of academic jargon generated incomplete data (values of 0), so in order to correlate the dataset, the academic jargon was simplified where necessary, and commonly used keywords optimized for user search behavior were employed (e.g., Energy Outages replaced with Power outage, or cyber threats replaced with Cyber attack). Furthermore, since Google Trends limits simultaneous comparison to a maximum of 5 terms and provides only relative values, the keyword "supply chain" was used as an anchor across three separate queries. The average interest scores were then mathematically scaled in Excel to a common reporting

baseline. Correlation coefficients of 0.29697 (Gemini) and 0.018181818 (ChatGPT and Claude) were obtained, which suggests that the classification system developed through LLMs is not strongly associated with fluctuations in online search interest, providing a ranking that appears less dependent on the popularity patterns reflected in Google Trends.

Additionally, a dummy ranking text was also created, in which 1,000 dummy rankings were generated to verify whether the hierarchy provided by LLMs was not randomly generated. Thus, it was observed that the variance of standard deviations shows a large mathematical difference between the real and simulated data. More specifically, the statistical analysis showed that the variance of the data provided by the LLMs is 0.029758294 for ChatGPT, 0.140388664 for Gemini, and 0.014547707 for Claude; these values indicate a natural dynamic, while the variance of dummy data (0.000956 for ChatGPT, 0.002184957 for Gemini, and 0.001218886 for Claude) is very close to zero, indicating a more homogeneous distribution. Because the 1,000 hierarchies were generated randomly, they produced stable statistical distributions, generating results that lack the structured differentiation observed in the rankings that were generated by LLMs.

## 4.4. Comparative analysis

The final part of this section consists of the comparative analysis between the specialists' results and those generated by the LLM systems. Before completing the questionnaire, the specialists were informed by the authors not to use LLM systems, as this could have altered the study's results. The questionnaires were first completed by the specialists, and only afterward were the three LLM systems queried by the authors. In Table 7, the Threat Priority Means resulting from the specialists' questionnaire and those generated by the three LLM systems are summarized. The comparison between the experts' results and the results generated by the LLM systems is highlighted using Spearman's Rank Correlation Coefficient (SRCC) method. This statistical method measures the strength and direction of the association between two variables, with values ranging between -1 and +1. The sign of the value (r) indicates either a direct (+) or inverse (–) correlation. A value of 0 indicates no correlation, while the extreme values -1 and +1 indicate a perfect correlation.

The values of the correlation coefficient (r) do not have a universally fixed interpretation; however, in the specialized literature, correlation coefficients (r) < 0.20 are considered "very weak" or "negligible," values between 0.30–0.40 are considered a "low," "fair," or "mild" relationship, values between 0.40–0.70 indicate a "moderate" relationship, values between 0.70–0.90 indicate a "strong" or "high" relationship, and (r) > 0.90 indicates a "very high" relationship (Alsaqr, 2021). The resulting values fall within the acceptance interval of -1 and 1, where the coefficient corresponding to Gemini shows a moderate positive correlation (0.41), ChatGPT shows a weak positive correlation (0.26), and Claude shows a very weak positive correlation (0.13). Furthermore, these results support and address the answer to the first research question (RQ 1), complemented by the results illustrated in the previous figures regarding the alignment of most criteria within the error bars.

**Table 7.** Statistical correlation between expert results and LLMs

| SD / Threats | **Specialists' Results** | **Chat GPT** | **Claude** | **Gemini** |
|---|---|---|---|---|
| **T 1** | 7.84 | 6.50 | 6.30 | 4.73 |
| **T 2** | 7.65 | 4.97 | 4.97 | 3.93 |
| **T 3** | 7.63 | 7.87 | 7.37 | 6.07 |
| **T 4** | 7.49 | 6.70 | 6.50 | 5.50 |
| **T 5** | 6.02 | 6.13 | 6.33 | 4.57 |
| **T 6** | 8.56 | 9.37 | 9.40 | 9.27 |
| **T 7** | 6.16 | 5.93 | 6.20 | 6.63 |
| **T 8** | 6.79 | 4.07 | 4.53 | 2.80 |
| **T 9** | 7.58 | 5.93 | 5.27 | 4.20 |

| | | | | |
|---|---|---|---|---|
| **T 10** | 5.74 | 6.70 | 6.90 | 3.20 |
| **SRCC** | - | **0.26** | **0.13** | **0.41** |

To answer RQ2, the analysis shows that the models do not always provide the same response, as they change the category of a threat from one run to another, a fact confirmed by the values in the table. Furthermore, the results reveal a moderately weak correlation between LLMs and human experts, which may be attributed to factors such as the models' inherent non-determinism, the absence of deterministic anchors for grounding assessment scales, and the lack of explicit utility models. These findings are consistent with the conclusions put forward by Felicioni et al. (2024) and the DeLLMa framework (Liu et al., 2024), which suggest that the direct use of prompts with LLMs is insufficient for conducting reliable automated critical infrastructure risk assessments. Consequently, assessments generated by LLMs may benefit from aggregation and refinement mechanisms, including approaches such as Bayesian updating and agentic evaluation frameworks (e.g., CritBench), in order to improve the results.

## 5. Conclusions

Artificial intelligence (AI) and large language models (LLMs) have advanced rapidly in recent years, demonstrating in many cases remarkable capabilities in processing complex information and are positioned as potential tools to support decision-makers. In the context of critical infrastructures, where rapid and accurate threat assessment is essential, LLMs can identify patterns, anticipate risks, and prioritize threats. Thus, the use of these technologies allows decision-makers to enhance resilience, optimize resource allocation, and develop strategies to protect vital systems by anticipating various threats to critical infrastructures.

This study highlights a hierarchy of threats specific to critical infrastructures from the perspective of two approaches. The first was supported by a group of Romanian specialists consisting of 19 individuals certified as liaison officers for the security of national critical infrastructures, who established a score of threats by completing a questionnaire. The second approach consisted of querying three artificial intelligence systems, applying the same rules used in the questionnaire. The hierarchies resulting from these two approaches were compared, and the specialists' results served as the baseline for comparing the three LLM systems.

The experts' results were based on the standard deviation calculated for each individual criterion and on the mean of the assigned values. Furthermore, the standard deviation represented a statistical element used in calculating the results generated by the LLM systems. In both cases, the resulting scores were compared, taking into account the standard deviation and the error bars. Finally, to determine the Threat Priority Mean between the specialists' and the LLMs' results, the Spearman's Rank Correlation Coefficient method was applied. This method highlighted three types of relationships between the specialists and the LLM systems, validated by the values obtained, which are situated within the acceptance interval of -1 and 1. Among the evaluated models, Gemini exhibited a moderate positive correlation (0.41), ChatGPT demonstrated a weak positive correlation (0.26), while Claude showed a very weak positive correlation (0.13).

At the same time, this article does not demonstrate which platform is more useful; rather, it presents an analytical model for improving the efficiency of risk analysis and decision-making processes within critical infrastructures. These results may change depending on the number of repetitions and the time interval considered. At the same time, the present study may also be viewed as a model for further research, through the expansion of the number of specialists, potentially even at the international level, or through its application in other fields of activity.

Submission received: 19 March 2026; Revised: 25 May 2026; Accepted: 05 June 2026; Published: 30 June 2026.

# REFERENCES

Alcaraz, C. & Zeadally, S. (2015) Critical infrastructure protection: requirements and challenges for the 21st century. *International Journal of Critical Infrastructure Protection*. 8, 53–66. https://doi.org/10.1016/j.ijcip.2014.12.002.

Cyber Incident Reporting for Critical Infrastructure Act of 2022 (2022) Public Law 117-103, Division Y. Enacted 15 March 2022. https://www.congress.gov/public-laws/117th-congress/public-law/117-103 [Accessed: 7 May 2026].

Alsaqr, A.M. (2021) Remarks on the use of Pearson's and Spearman's correlation coefficients in assessing relationships in ophthalmic data. *African Vision and Eye Health*. 80(1), 1–6. https://doi.org/10.4102/aveh.v80i1.612.

Chaudhry, J., Pathan, A.S.K., Rehmani, M.H. et al. (2018) Threats to critical infrastructure from AI and human intelligence. *Journal of Supercomputing*. 74, 4865–4866. https://doi.org/10.1007/s11227-018-2614-0 [Accessed: 10 November 2025].

Council of the European Union. (2008) Council Directive 2008/114/EC of 8 December 2008 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection. *Official Journal of the European Union*. L 345/75. Available at: https://eur-lex.europa.eu/eli/dir/2008/114/oj/eng [Accessed: 21 November 2025].

Davis, R., Bace, B. & Tatar, U. (2024) Space as a critical infrastructure: an in-depth analysis of U.S. and EU approaches. *Acta Astronautica*. 225, 263–272. https://doi.org/10.1016/j.actaastro.2024.08.053.

da Silva, E.G., Georg, M.A.C., Júnior, L.A.R. et al. (2025) International perspectives on critical infrastructure: evaluation criteria and definitions. *International Journal of Critical Infrastructure Protection*. 49, 100761. https://doi.org/10.1016/j.ijcip.2025.100761.

De Felice, F., Baffo, I. & Petrillo, A. (2022) Critical infrastructures overview: past, present and future. *Sustainability*. 14(4), p. 2233. https://doi.org/10.3390/su14042233.

European Parliament and Council of the European Union. (2022a) *Directive (EU) 2022/2555 of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS2 Directive)*. Available at: https://eur-lex.europa.eu/eli/dir/2022/2555/oj [Accessed: 8 May 2026]

European Parliament and Council of the European Union. (2022b) *Directive (EU) 2022/2557 of 14 December 2022 on the resilience of critical entities (CER Directive).* Available at: https://eur-lex.europa.eu/eli/dir/2022/2557/oj [Accessed: 7 November 2025].

Felicioni, N., Maystre, L., Ghiassian, S. & Ciosek, K. (2024) On the importance of uncertainty in decision-making with large language models. *Transactions on Machine Learning Research.* https://doi.org/10.48550/arXiv.2404.02649.

Ferrara, E. (2023) Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*. 28(11). https://doi.org/10.5210/fm.v28i11.13346.

Iturriza, M., Labaka, L., Sarriegi, J.M. & Hernantes, J. (2018) Modelling methodologies for analysing critical infrastructures. *Journal of Simulation*. 12(2), 128–143. https://doi.org/10.1080/17477778.2017.1418640.

Keppler, G., Gstür, M. & Hagenmeyer, V. (2026) CritBench: *A framework for evaluating cybersecurity capabilities of large language models in IEC 61850 digital substation environment*. https://doi.org/10.1007/978-3-032-16089-8_31.

Liu, O., Fu, D., Yogatama, D. & Neiswanger, W. (2024) DeLLMa: Decision making under uncertainty with large language models. *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024).* https://doi.org/10.48550/arXiv.2402.02392.

Maguire, E. (n.d.) Introducing Lumo, the AI where every conversation is confidential. *Proton*. Available at: https://proton.me/blog/lumo-ai [Accessed: 27 January 2026].

Makrakis, G.M., Kolias, C., Kambourakis, G., Rieger, C.G. & Benjamin, J. (2021) Industrial and critical infrastructure security: Technical analysis of real-life security incidents. *IEEE Access*. 9, 165286–165312. https://doi.org/10.1109/ACCESS.2021.3133348.

Markopoulou, D. & Papakonstantinou, V. (2021) The regulatory framework for the protection of critical infrastructures against cyberthreats: identifying shortcomings and addressing future challenges: the case of the health sector in particular. *Computer Law & Security Review*. 41, 105502. https://doi.org/10.1016/j.clsr.2020.105502.

Mottahedi, A., Sereshki, F., Ataei, M. et al. (2021) The resilience of critical infrastructure systems: a systematic literature review. *Energies*. 14(6), p. 1571. https://doi.org/10.3390/en14061571.

Murray, G., Johnstone, M.N. & Valli, C. (2017) The convergence of IT and OT in critical infrastructure. In Valli, C. (ed.), *Proceedings of the 15th Australian Information Security Management Conference. Perth: Edith Cowan University*. pp. 149–155. https://doi.org/10.4225/75/5a84f7b595b4e.

Nacanabo, M.W., Bayala, Y.L.T., Seghda, A.A.T. et al. (2026) Comparative study of the performance of ChatGPT-4, Claude, Gemini, Mistral, and Perplexity on multiple-choice questions in cardiology. *BMC Cardiovascular Disorders*. 26, p. 32. https://doi.org/10.1186/s12872-025-05431-y.

Osei-Kyei, R., Tam, V.W.Y., Ma, M. & Mashiri, F.R. (2021) Critical review of the threats affecting the building of critical infrastructure resilience. *International Journal of Disaster Risk Reduction*. 60, p. 102316. https://doi.org/10.1016/j.ijdrr.2021.102316.

Pătrașcu, P. (2022) National security strategies and critical infrastructure: an analysis of the European Union member states. *Romanian Military Thinking*. (3), 10–29. https://doi.org/10.55535/RMT.2022.3.01.

Pursiainen, C. (2017) Critical infrastructure resilience: a Nordic model in the making? *International Journal of Disaster Risk Reduction* 27, 632–641. https://doi.org/10.1016/j.ijdrr.2017.08.006.

Qiu, L., Sha, F., Allen, K., Kim, Y., Linzen, T. & van Steenkiste, S. (2026) Bayesian teaching enables probabilistic reasoning in large language models. *Nature Communications*. 17, p. 1238. https://doi.org/10.1038/s41467-025-67998-6.

Sarker, I.H., Janicke, H., Ferrag, M.A. et al. (2024) Multi-aspect rule-based AI: methods, taxonomy, challenges and directions towards automation, intelligence and transparent cybersecurity modeling for critical infrastructures. *Internet of Things*. 25, 101110. https://doi.org/10.1016/j.iot.2024.101110

Trinchillo, F. (2025) Critical infrastructures: European context and evolution of the law. *Journal of Defense Resources Management*. 16(1), 309–320. https://doi.org/10.64404/jodrm.2025.1.15

Żaboklicka, E. (2020) Critical infrastructure in the shaping of national security. *Security and Defence Quarterly*. 28(1), 70–81. https://doi.org/10.35467/sdq/118585.

**Ana-Maria CHISEGA-NEGRILĂ** este conf. univ. dr. în cadrul Universității Naționale de Apărare „Carol I” din București, România. Domeniile sale de cercetare includ inteligența artificială, aplicarea tehnologiilor emergente în predare și învățare și, mai larg, impactul tehnologiei asupra proceselor educaționale. A publicat articole științifice despre predarea și învățarea într-o lume condusă de inteligență artificială și despre legătura dintre AI și educație.

**Ana-Maria CHISEGA-NEGRILĂ** is an Associate Professor at the Carol I National Defence University in Bucharest, Romania. Her research areas include artificial intelligence, the application of emerging technologies in teaching and learning and, more broadly, the impact of technology on educational processes. She has published scientific articles on teaching and learning in an AI-driven world and on the relationship between AI and education.

**Petrișor PĂTRAȘCU** is an associate professor at the "Carol I" National Defense University in Bucharest, Romania. His current research interests include critical infrastructure protection, cybersecurity, and strategic studies.

**Petrișor PĂTRAȘCU** este conferențiar universitar la Universitatea Națională de Apărare "Carol I" din București, România. Domeniile sale actuale de cercetare includ protecția infrastructurilor critice, securitatea cibernetică și studiile strategice.