doi: https://doi.org/10.15407/visn2026.04.037

ХАРЧЕНКО
Вячеслав Сергійович —
член-кореспондент НАН України, доктор технічних наук, професор, завідувач кафедри кібербезпеки та інтелектуальних інформаційних технологій факультету радіоелектроніки, комп'ютерних систем та інфокомунікацій Національного аерокосмічного університету «Харківський авіаційний інститут»

НАУКОВІ ЗАСАДИ, МЕТОДИ СТВОРЕННЯ ТА ВПРОВАДЖЕННЯ ГАРАНТОЗДАТНИХ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ

Стенограма доповіді на засіданні Президії НАН України 25 лютого 2026 року

У доповіді наведено актуальні результати робіт, спрямованих на розроблення теорії, методів і технологій аналізу, оцінювання, створення та використання гарантоздатних систем штучного інтелекту для ефективного керування, оброблення даних, забезпечення готовності, безпечності та резильєнтності об'єктів критичної інформаційної та енергетичної інфраструктури, інтелектуальних безпілотних систем та інших систем, комплексів та інфраструктур різного призначення, важливих для оборони та безпеки країни.

Вельмишановний Анатолію Глібовичу!
Вельмишановні члени Президії, присутні!
Мою доповідь присвячено не лише новим можливостям, які відкриваються завдяки розробленню та впровадженню методів і технологій штучного інтелекту (ШІ), а й пов'язаним з цим викликам щодо надійності, прогнозованості й безпеки функціонування критичних систем, де використовуються засоби ШІ.

Системи штучного інтелекту і виклики безпеки. Системи штучного інтелекту (СШІ) — це програмно-апаратні системи, здатні виконувати завдання нарівні з можливостями людського інтелекту або перевершувати їх завдяки використанню сучасних інформаційних технологій. Поряд із незрівнянними перевагами від впровадження відповідних методів і засобів ШІ в енергетику, логістику, оборону та інші галузі, особливо в умовах війни і в період повоєнного відновлення, — від ШІ-моделей для дронів та інтелектуальних гетерогенних мобільних систем (UXV-систем), які поєднують різні типи мобільних апаратів (літальні, наземні, морські безекіпажні та ін.), до предиктивної аналітики, індустріального інтернету речей (IoT) та новітніх методів медичної діагностики — зростають ризики відмови програмно-апаратних платформ, неконтрольованої та непередб-

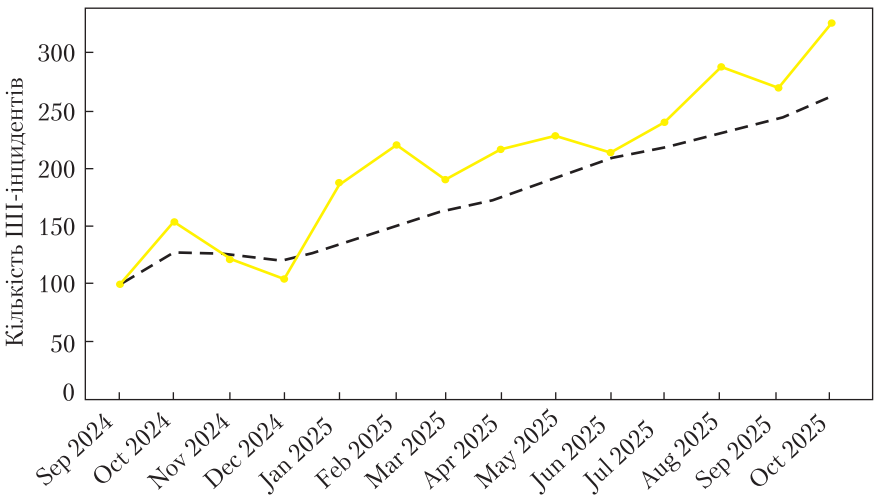

Рис. 1. Динаміка зареєстрованих інцидентів та аварій, пов'язаних зі штучним інтелектом (AIM OECP)

бачуваної поведінки СШІ, успішних кібератак на вразливості та інших негативних наслідків. Так, за даними Департаменту штучного інтелекту США (U.S. AI Governance)*, кількість інцидентів та аварій, пов'язаних із використанням ШІ, лише протягом 2025 р. зросла більш як утричі (рис. 1).

З точки зору вивчення надійності та безпеки системи штучного інтелекту є відносно новим об'єктом дослідження. Наразі у світі спостерігається стрімкий розвиток методів і засобів машинного / глибокого навчання (ML/DL) та великих мовних моделей (LLM), зростання об'ємів і концентрації даних, розроблення синтетичних методів генерації датасетів — наборів даних, що використовують для навчання (однак тут, на мій погляд, переважають екстенсивні шляхи), і водночас накопичуються проблеми, пов'язані з невизначеністю й невпорядкованістю характеристик моделей і систем штучного інтелекту. Особливо гостро ці питання постають у разі використання ШІ в так званих критичних системах, у яких ціна відмови є дуже високою.

Наведу лише один приклад. 5 листопада 2022 р. в китайському місті Чаочжоу (провінція Гуандун) сталася автомобільна аварія. 55-річний водій Tesla Model Y збирався припаркуватися біля власного магазину, але замість гальмування автомобіль раптом почав неконтрольовано прискорюватися. Розігнавшись до 164 км/год і маневруючи між транспортними засобами на дорозі, авто врізалося в стіну будівлі. У цій дорожній пригоді загинуло двоє людей, ще троє отримали поранення. Водій, який залишився живим, стверджував, що відчайдушно намагався загальмувати, але «програмний збій у системі штучного інтелекту, навпаки, розганяв автівку». Інцидент набув широкого розголосу і став предметом тривалих фахових дискусій, навіть попри те, що суд зрештою визнав винним водія. Цей випадок яскраво ілюструє, наскільки небезпечними можуть бути непередбачувані збої в роботі критичних інтелектуальних систем.

Гарантоздатність — поєднання надійності та безпеки програмно-апаратних комплексів. Новизна систем штучного інтелекту як об'єктів забезпечення надійності та безпеки зумовлює необхідність формування концептуальної бази для створення гарантоздатних СШІ. Гарантоздатність (dependability) — це здатність системи надавати послуги (виконувати специфіковані функції), яким можна обґрунтовано довіряти.

Розвиток ШІ додав до традиційних і добре відомих причин відмов у системах (ідеться про фізичні дефекти апаратних засобів та проєктні дефекти програмного забезпечення) ще один

* AI Incidents Are Rising. It's Time for the United States to Build Playbooks for When AI Fails. November 12, 2025. https://thefuturesociety.org/us-ai-incident-response/

ISSN 1027-3239. Visn. Nac. Acad. Nauk Ukr. 2026. (4)

вимір — кібератаки (втручання) на специфічні вразливості систем, пов'язані з використанням штучного інтелекту. Тому концепція гарантоздатності поєднує безвідмовність та надійність апаратних і програмних засобів, функціональну безпечність та кібербезпеку системи.

В Україні проблемами гарантоздатності інтелектуальних систем займаються досить давно. Так, з 2006 р. щороку проводиться започаткована нами міжнародна науково-технічна конференція «Гарантоздатні системи, сервіси та технології» (Dependable Systems, Services and Technologies — DESSERT). У 2010 р. було ухвалено настанову Національного космічного агентства України СОУ-Н НКАУ 0060:2010 «Галузева система управління якістю. Гарантоздатність програмно-технічних комплексів критичного призначення», яка встановлює вимоги щодо надійності, безпечності та безвідмовності критичних систем в аерокосмічній галузі.

Наукові основи гарантоздатних систем штучного інтелекту. Сьогодні дослідження науковців кафедри кібербезпеки та інтелектуальних інформаційних технологій Національного аерокосмічного університету «Харківський авіаційний інститут» спрямовано на розроблення концептуальних основ, принципів та методів аналізу й оцінювання гарантоздатності систем штучного інтелекту.

Зокрема, сформовано й розвинено теорію та розроблено моделі гарантоздатних інтелектуальних обчислень (DIC-моделі — dependable intelligent computing), які побудовано об'єднанням розробленої свого часу для комп'ютерних систем таксономії гарантоздатності (так званих ALR-моделей — Avizienis—Laprie—Randell) і моделей якості ШІ та програмно-апаратної платформи СШІ (AIQM-моделей — artificial intelligence quality model). Такий підхід, запропонований нами чотири роки тому, дозволяє впорядкувати і гармонізувати класичні складові гарантоздатності та специфічні характеристики штучного інтелекту.

Використана і розвинена нами парадигма гарантоздатного комп'ютингу (dependable computing), або гарантоздатних обчислень, являє собою комплекс моделей, методів і засобів, які забезпечують виконання вимог до безвідмовності, готовності, функціональної безпечності, кібербезпеки та її складових (цілісності, доступності, конфіденційності), а за певних умов — живучості та резильєнтності. Термін «резильєнтність» (resilience) означає тут здатність системи еволюціонувати в разі зміни вимог, параметрів середовища чи виникнення неспецифікованих відмов. Ми показали, що системи штучного інтелекту за своєю природою є резильєнтними системами [1].

Розроблено і досліджено множину сценаріїв і моделей поведінки СШІ (AIS-моделей — artificial intelligent systems) на основі комплексного підходу, який ґрунтується на так званому тріад-базованому аналізі ШІ як ресурсу, який: а) захищається; б) може підсилювати захист і відмовобезпечність; в) може підсилювати кіберфізичні впливи (атаки) на СШІ. Крім того, розроблено низку відповідних змагальних сценаріїв, які дають змогу досліджувати поведінку систем штучного інтелекту в різних умовах.

За результатами досліджень багатоверсійних систем і технологій, зокрема в атомній енергетиці та аерокосмічних комплексах, запропоновано й теоретично обґрунтовано принцип комбінованої диверсності (версійної, версійно-структурної або версійно-часової надмірності). Диверсність — це принцип, що використовують у деяких складних системах, у яких традиційного структурного резервування недостатньо і потрібно застосовувати різні варіанти реалізації резервних каналів із використанням різних (диверсних) програмно-апаратних платформ, різних процесів розроблення, тестування тощо. Розроблено також методи оперативного контролю, реконфігурації та донавчання резервних підсистем для забезпечення гарантоздатності СШІ.

Розвинено парадигму фон Неймана (VNP), сформульовану близько 70 років тому для простих цифрових засобів. У сучасному її застосуванні (VNP*-парадигма) йдеться про створення гарантоздатних систем штучного інтелекту з недостатньо надійних і недостатньо безпечних інтелектуальних підсистем (компонент) в умовах дії агресивного фізичного й інформа-

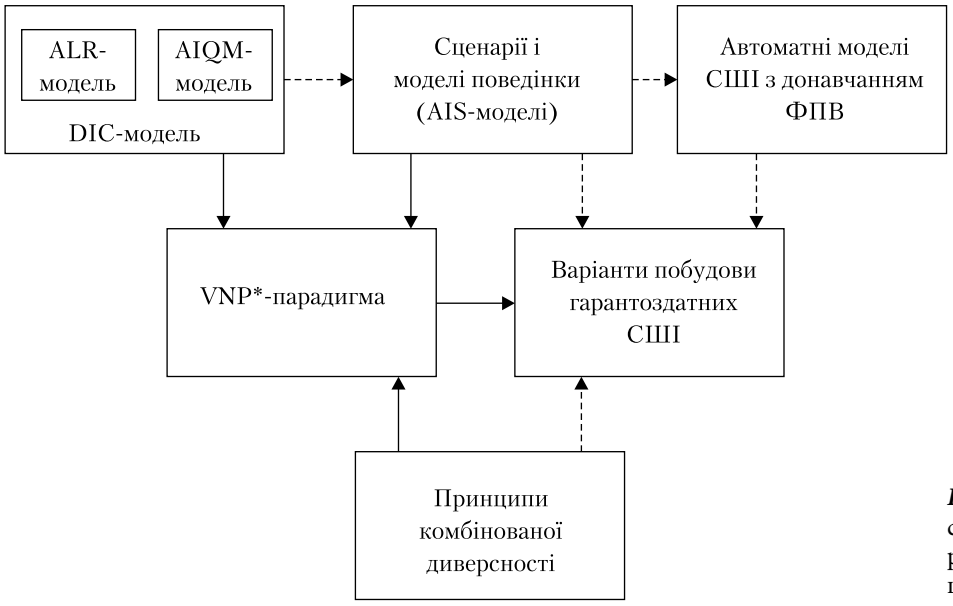

Рис. 2. Концептуальна схема забезпечення гарантоздатності систем штучного інтелекту [3]

ційного середовища зі змінними параметрами [2]. Можна сказати, що зараз ми перебуваємо на сьомому етапі еволюції VNP*-парадигми, на якому формується теорія систем синергетичного інтелекту, що поєднує штучний і людський інтелект у багатоверсійну систему з мінімізацією ризиків відмови із загальної причини. Саме такі ризики є критичними факторами для систем, у яких визначальною вимогою є функційна безпечність і надійність, наприклад для систем аварійного захисту енергоблоків АЕС чи бортових комп'ютерів безпілотних систем.

На рис. 2 наведено концептуальну схему забезпечення гарантоздатності систем штучного інтелекту, яка поєднує комплекс моделей власне гарантоздатного комп'ютингу, моделей і сценаріїв поведінки (AIS-моделі) та принципи використання комбінованої диверсності [3]. На відміну від добре відомих моделей якості програмного забезпечення, концепція моделі якості для штучного інтелекту зараз тільки вибудовується. Вона має багаторівневу ієрархічну структуру, яка охоплює 46 характеристик моделей та ШІ-платформ, базовими з яких є довірчоздатність (trustworthiness), поясненність (explainability), відповідальність (responsibility), етичність (ethics), законність (lawfulness) [4].

Приклади впровадження. Сферами застосування отриманих нами результатів є системи аварійного захисту реакторів АЕС, аерокосмічні й безпілотні інтелектуальні системи, мобільні системи екологічного моніторингу та розмінування тощо. Наведу лише окремі приклади.

На особливу увагу заслуговує запропонована для гуманітарного розмінування система пошуку та ідентифікації вибухонебезпечних предметів із застосуванням мультисенсорних інтелектуальних платформ. У таких комплексах використовують різні типи сенсорів, для кожного з яких створюють спеціалізовані нейромоделі. Потім інформацію з них акумулюють та аналізують увесь масив даних, що дозволяє значно підвищити достовірність виявлення вибухонебезпечних предметів. На основі моделей якості ШІ формуються вимоги до систем, що дозволяє знизити ризики, пов'язані із застосуванням саме компонентів штучного інтелекту [5]. У 2024—2025 рр. ця розробка (патент України на корисну модель № 1542266) успішно пройшла полігонні випробування. Крім того, ми поєднали цю систему з технологіями доповненої реальності. При цьому результати дронового обстеження території та її тривимірного картографування передають саперам

у вигляді 3D-моделей у шарі доповненої реальності, що значно підвищує безпеку їхньої роботи та прискорює процес розмінування.

Слід зазначити, що модельна частина цієї розробки має свою специфіку. Ми використовуємо класичний апарат марковських випадкових процесів і його модифікацію — багатофрагментні марковські моделі, які враховують різні чинники можливих відмов і кібератак, а також здатність систем штучного інтелекту донавчатися, крок за кроком поліпшуючи свої характеристики [6].

Інший приклад практичного застосування наших розробок — це створення та верифікація систем безпеки для об'єктів атомної енергетики, які знижують ризики критичних відмов із загальної причини. Впродовж багатьох років ми плідно співпрацюємо з Науково-виробничим підприємством «Радій» щодо розроблення і впровадження систем аварійного захисту енергоблоків АЕС та інших систем, важливих для їхньої безпеки. Такі системи багатоверсійні, тобто їх побудовано за принципом диверсності, коли основна та диверсна системи виконують однакові функції, але реалізовано їх на різних програмно-апаратних платформах. Зокрема, йдеться про двоверсійну систему аварійного захисту на базі платформи Radiy та платформи наступної генерації — RadICS, яку було сертифіковано на відповідність найжорсткішим у світі вимогам американського ядерного регулятора — Комісії з ядерного регулювання США (United States Nuclear Regulatory Commission — US NRC).

Керуючі системи безпеки систем аварійного захисту реакторів впроваджено в ядерній галузі України, Болгарії та інших країн, а платформи НВП «Радій» сертифіковано в США і Канаді. Український досвід використання цих систем у складних умовах війни та кіберфізичних збурень засвідчив їхню високу безпечність та резильєнтність.

Стратегії диверсності реалізовано не лише в системах безпеки ядерної енергетики, а й у аерокосмічній галузі та для різних видів моніторингу територій критичних об'єктів. Для таких систем ми розробили і постійно оновлюємо класифікатор видів диверсності — різних варіантів надмірності, які доповнюють традиційне резервування і дозволяють мінімізувати ризик відмов із загальної причини.

Загалом впровадженням принципу диверсності для безпеки критичних систем займаються вже понад 20 років, але з появою штучного інтелекту цей напрям набув нового розвитку [7—9]. Зокрема, розроблено і запатентовано (патент України на корисну модель № 156654) спосіб резервування систем штучного інтелекту, що ґрунтується на використанні різних джерел і наборів даних для навчання, створення альтернативних моделей та алгоритмів, а також незалежних каналів контролю їхньої працездатності. Такий підхід дозволяє створювати кілька незалежних версій системи, результати роботи яких можна порівнювати і перевіряти.

У співпраці з нашими партнерами ми розробили 12 національних і галузевих стандартів, які містять вимоги до функційної безпечності та кібербезпеки інформаційних систем управління і програмно-технічних комплексів атомних станцій, космічних систем, методів регулювання, проєктування, верифікації, забезпечення ІТ-безпеки та гарантоздатності критичних систем. Ці стандарти заклали методологічну базу для подальшої адаптації підходів до забезпечення надійності й безпеки складних технічних систем з використанням штучного інтелекту. Опубліковано кілька монографій, зокрема в іноземних видавництвах [10—12]. Цього року в США має вийти ще одна монографія, присвячена проблемам безпеки цифрової інфраструктури малих модульних реакторів. Отримано 5 патентів на наші розробки. Видано дві білі книги (white paper) з впровадження стандартів IEC 62443 (кібербезпека) та IEC 61508 (функційна безпечність) в індустріальних системах. Разом з установами НАН України ми постійно беремо участь у виконанні міжнародних проєктів за програмою «Горизонт Європа», програмою НАТО «Наука заради миру і безпеки» та ін. Наприклад, проєкт ECHO був спрямований на створення європейської міжсекторальної мережі компетенцій у сфері кібербезпеки, і наша команда розробила європейську дорож-

ню карту з використання засобів ШІ для забезпечення кібербезпеки автономних транспортних систем (безпілотних літальних, наземних і морських апаратів, супутникових систем) та брала участь у розробленні інших програм досліджень та інновацій у сфері кібербезпеки.

Висновки. Отже, в Національному аерокосмічному університеті «Харківський авіаційний інститут» у співпраці з іншими закладами вищої освіти, установами НАН України та підприємствами активно розвивається новий напрям з розроблення наукових засад, методів і технологій оцінювання та забезпечення гарантоздатності й резильєнтності систем штучного інтелекту, які сьогодні стають важливою складовою об'єктів критичної інформаційної та енергетичної інфраструктури, засобів автоматизації промислових підприємств, інтелектуальних безпілотних систем та інших комплексів різного призначення.

Надалі ми плануємо зосередити наші зусилля на таких напрямах:

• розроблення систем для гетерогенних інтелектуальних мобільних комплексів, що об'єднують безпілотні літальні апарати, наземні роботизовані системи, морські безекіпажні комплекси;

• створення гарантоздатних систем і сервісів штучного інтелекту, методів оцінювання їхньої довірчоздатності, поясненності, відповідальності;

• формування чітких регуляторних вимог до програмного забезпечення систем штучного інтелекту;

• розвиток принципів синергетичного інтелекту як складової концепції ноокомп'ютингу;

• поєднання методів штучного інтелекту, доповненої реальності, цифрових двійників та інтернету речей у різних сферах застосування, зокрема в інтерактивному мистецтві, реабілітаційних комплексах тощо.

Дякую за увагу!

За матеріалами засідання підготувала О.О. Мележик

REFERENCES

1. Moskalenko V., Kharchenko V., Moskalenko A., Kuzikov B. Resilience and resilient systems of artificial intelligence: taxonomy, models and methods. *Algorithms*. 2023. **16**(3): 165. https://doi.org/10.3390/a16030165
2. Kharchenko V., Odarushchenko O. Trustworthy AI systems from untrustworthy components: development von Neumann's paradigm using principle of diversity. In: Proc. 4th Int. Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024). Cambridge, MA, USA. P. 392—404. https://ceur-ws.org/Vol-3777/
3. Kharchenko V.S. Conceptual fundamentals of dependable artificial intelligence systems. *Reports of the National Academy of Sciences of Ukraine*. 2025. (2): 11—23. https://doi.org/10.15407/dopovidi2025.02.011
4. Kharchenko V., Fesenko H., Illiashenko O. Quality models for artificial intelligence systems: characteristic-based approach, development, application. *Sensors*. 2022. **22**(13): 4865. https://doi.org/10.3390/s22134865
5. Fedorenko G., Fesenko H., Kharchenko V., Kliushnikov I., Tolkunov I. Robotic-biological systems for detection and identification of explosive ordnance: concept, general structure, and models. *Radioelectronic and Computer Systems*. 2023. (2): 143—159. https://doi.org/10.32620/reks.2023.2.12
6. Kharchenko V., Ponochovnyi Yu., Zemlianko H. Markov's models of AI systems availability considering re-learning processes. In: Proc. 5th Int. Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2025). Liverpool, UK, 2025. P. 240—249. https://ceur-ws.org/Vol-4164/
7. Kharchenko V., Shchehlov V., Ivasiuk O., Morozova O. Digital twin-based lifecycle methodology for ensuring safety of NPP/SMR I&C systems. *Technologies*. 2026. **14**(1): 46. https://doi.org/10.3390/technologies14010046
8. Kharchenko V., Ponochovnyi Y., Ruchkov E., Babeshko E. Safety Assessment of the Two-Cascade Redundant Information and Control Systems Considering Faults of Versions and Supervision Means. In: Zamojski W., Mazurkiewicz J., Sugier J., Walkowiak T., Kacprzyk J. (eds) *New Advances in Dependability of Networks and Systems*. Springer Cham, 2022. https://doi.org/10.1007/978-3-031-06746-4_9
9. Panarin A., Kharchenko V., Zemlianko H. Advanced classifier for expert assessment of diversity and cost of NPP I&C systems. *ISTCMTM*. 2025. **86**(4): 52—57. https://doi.org/10.23939/istcmtm2025.04.052

10. Yastrebenetsky M., Kharchenko V. *Cyber Security and Safety of Nuclear Power Plant Instrumentation and Control Systems*. IGI Global, Hershey, United States, 2014.
11. Yastrebenetsky M., Kharchenko V. *Nuclear Power Plant Instrumentation and Control Systems for Safety and Security*. IGI Global, Hershey, United States, 2014.
12. Kharchenko V., Paturej A., Potii O. (eds). *Manual on Cybersecurity, Reliability and Resilience Assurance in the Critical Industries*. International Centre for Chemical Safety and Security, Warsaw, 2024.

Vyacheslav S. Kharchenko
National Aerospace University "Kharkiv Aviation Institute," Kharkiv, Ukraine
ORCID: https://orcid.org/0000-0001-5352-077X

SCIENTIFIC PRINCIPLES, METHODS OF CREATING AND IMPLEMENTING DEPENDABLE AI SYSTEMS

Transcript of scientific report at the meeting of the Presidium of NAS of Ukraine, February 25, 2026

The report presents current results of work aimed at developing the theory, methods and technologies of analysis, evaluation, creation and use of dependable AI systems for effective management, data processing, ensuring the readiness, safety and resilience of critical information and energy infrastructure facilities, intelligent unmanned systems and other systems, complexes and infrastructures of various purposes important for the defense and security of the country.

Cite this article: Kharchenko V.S. Scientific principles, methods of creating and implementing dependable AI systems (transcript of scientific report at the meeting of the Presidium of NAS of Ukraine, February 25, 2026). *Visn. Nac. Akad. Nauk Ukr*. 2026. (4): 37—43. https://doi.org/10.15407/visn2026.04.037