ДОПОВІДІ
НАЦІОНАЛЬНОЇ
АКАДЕМІЇ НАУК
УКРАЇНИ

ІНФОРМАТИКА І КІБЕРНЕТИКА
INFORMATICS AND CYBERNETICS

https://doi.org/10.15407/dopovidi2025.02.011
УДК 004.8.05

В.С. Харченко, https://orcid.org/0000-0001-5352-077X

Національний аерокосмічний університет «Харківський авіаційний інститут», Харків
E-mail: v.kharchenko@csn.khai.edu

Концептуальні основи гарантоздатних систем штучного інтелекту

Представлена членом-кореспондентом НАН України С.В. Яковлевим

Запропоновано концепцію гарантоздатних систем штучного інтелекту (ШІ) на базі розвитку парадигми фон-Неймана (von Neumann paradigm, VNP), яку представлено теоретико-множинним описом з урахуванням різних складових — характеристик якості ШІ та систем ШІ (СШІ). Модель якості СШІ описується як упорядкована ієрархія характеристик довірчоздатності, поясненності, етичності, законності, відповідальності та їх підхарактеристик, що дозволяє визначити можливості застосування VNP для забезпечення виконання вимог до окремих характеристик. Розроблено модель відповідності перетворення вхідних та вихідних даних СШІ з врахуванням декомпозиції універсальної множини наборів даних (датасетів) на підмножини тих, що використовувалися для навчання та можливих некоректностей за певними характеристиками якості та різних типів відмов та кібератак на СШІ. Запропоновано використання принципу диверсності для впровадження VNP для забезпечення довірчоздатності та інших характеристик ШІ та створення гарантоздатних СШІ.

Ключові слова: *гарантоздатні обчислення, довірчоздатний штучний інтелект, модель якості систем штучного інтелекту, парадигма фон-Неймана, принцип диверсності.*

Вступ. Розроблення та впровадження методів і технологій штучного інтелекту (ШІ) надає багато нових можливостей, але також формує доленосні виклики. Одними з найбільш чутливих є загрози порушення надійного та безпечного функціонування систем ШІ (СШІ), обумовлені невизначеністю їх поведінки в ситуаціях поза межами випадків, перевірених на обмежених послідовностях вхідних даних і наборах даних, які використовувалися для навчання (датасетів), а також відмовами внаслідок недосконалої програмно-апаратної реалізації, фізичними дефектами та кібератаками [1].

Крім того, важливими є задачі формування і забезпечення виконання зрозумілих вимог до специфічних характеристик ШІ, а саме етичності (ethics), поясненності (explainability), довірчоздатності (trustworthiness) тощо, які визначено і упорядковано в [2, 3].

Цитування: Харченко В.С. Концептуальні основи гарантоздатних систем штучного інтелекту. *Допов. Нац. акад. наук Укр.* 2025. № 2. С. 11—23. https://doi.org/10.15407/dopovidi2025.02.011

Слід зазначити, що більшість публікацій з розвитку ШІ, машинного і глибокого навчання, інших методів спрямовано на їх удосконалення шляхом розширення та більш інтенсивного використання множини даних для навчання, але це ще більше ускладнює перевірку СШІ і залишає невизначеність внаслідок збільшення складності моделей [4, 5]. Такі проблеми мають місце і з Large Language Models (LLM), які навчаються на великих об'ємах стиснутих даних, що може призвести не тільки до невизначеності в поведінці систем, але й до так званих галюцинацій [6].

Новизна СШІ як об'єктів забезпечення надійності та безпеки обумовлює необхідність формування концептуальної бази для створення гарантоздатних систем штучного інтелекту. Гарантоздатність відповідно до класичної роботи авторів A. Avizienis, J.-C. Laprie, B. Randell [7], де було сформовано принципи об'єднання надійності і безпеки (ALR-модель), є властивість системи надавати послуги (виконувати специфіковані функції), яким можна обґрунтовано довіряти. Принципи гарантоздатного комп'ютингу або обчислень (dependable computing) і таксономія гарантоздатності (dependability, DPD) як комплексної властивості, яка узагальнює, перш за все, безвідмовність (Reliability, RLB), готовність (Availability, AVL), функційну безпечність (Safety, SFT), кібербезпеку (Cybersecurity, CSC) та її складові — цілісність (Integrity, INT), конфіденційність (Confidentiality, CFD), а за певних умов — живучість (Survivability, SRV) та резильєнтність (resilience , RSL), були запропоновані і розвинуті в [7—11], але надалі недостатньо поширені на СШІ та інтелектуальні обчислення (сукупність моделей, методів та технологій обчислень з використанням ШІ) з урахуванням їх особливостей. Зазначимо, що переклад терміну resilience як стійкість не є достатньо коректним, оскільки описує тільки одну з її складових і не враховує додаткові властивості, пов'язані зі здатністю еволюціонувати за умов зміни вимог, параметрів середовища та виникнення неспецифікованих відмов [8].

Ці обставини пояснюють також обмеженість теоретичних і системних рішень щодо розроблення гарантоздатних СШІ, забезпечення надійності, безпечності та споріднених характеристик штучного інтелекту впродовж використання таких систем внаслідок недосконалості методологічної бази. Немає прийнятної відповіді на запитання, як мають розроблятися нові і удосконалюватися існуючі принципи і методи резервування компонент ШІ, зокрема, на підставі модернізації парадигми J. von Neumann (VNP) «синтез надійних цифрових систем із ненадійних компонентів» [12], враховуючи етапи її еволюції впродовж семи десятиліть від використання простого резервування (дублювання, мажоритування) логічних елементів до застосування потужних комбінованих методів резервування і реконфігурації складних систем та інфраструктур [13].

Мета статті — розроблення концепції гарантоздатних систем штучного інтелекту і принципів їх створення на підставі розвитку та узагальнення результатів автора, викладених в [2, 3, 11, 14—19].

Концептуальні основи гарантоздатних СШІ базуються на:

- формуванні моделі гарантоздатних інтелектуальних обчислень (Dependable Intelligent Computing, DIC-моделі) шляхом об'єднання таксономічної моделі гарантоздатності (ALR-моделі) [7] і моделі якості ШІ та програмно-апаратної платформи СШІ (відомої як AIQM-модель) [3], яка упорядковує і гармонізує класичні складові гарантоздатності та специфічні характеристики штучного інтелекту;
- описі сценаріїв і моделей поведінки СШІ (AIS-моделей) на підставі декомпозиції простору вхідних даних з можливими порушеннями за різними характеристиками з врахуванням визначеної множини причин на модельному та платформному рівнях;

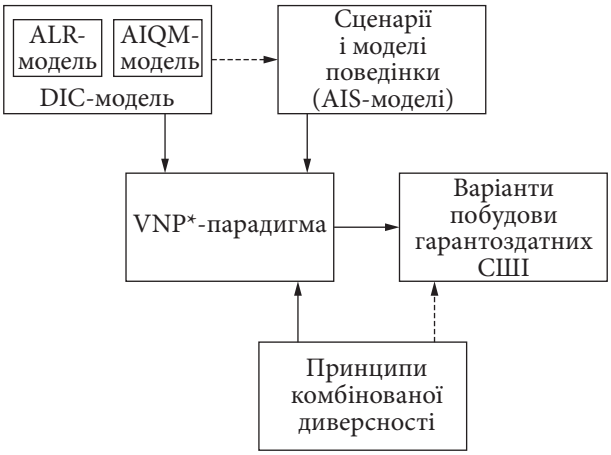

Рис. 1. Схема взаємозв'язків концептуальних моделей гарантоздатних СШІ

▪ розвинутій VNP-парадигмі (VNP*-парадигмі) — її представленню як «синтез гарантоздатних СШІ з недостатньо гарантоздатних інтелектуальних підсистем (компонент)» з можливою вибірковою декомпозицією за різними характеристиками;

▪ використанні принципу комбінованої диверсності (версійної, версійно-структурної або версійно-часової надмірності), а також різних варіантів оперативного контролю та донавчання резервних підсистем для забезпечення гарантоздатності СШІ [14].

Проаналізуємо ключові поняття та концептуальні моделі гарантоздатних інтелектуальних обчислень та СШІ та їх взаємозв'язки за схемою, представленою на рис. 1.

Модель гарантоздатності СШІ. Щоб відповісти на питання про те, як VNP можна розвинути та використовувати для покращення специфічних характеристик, пов'язаних із надійністю інтелектуальних систем, необхідно визначити особливості характеристик ШІ. Для цього доцільно використовувати модель якості, запропоновану в [2, 3]. Модель якості СШІ об'єднує 46 характеристик і складається з двох частин: власне моделі якості штучного інтелекту і моделі якості програмно-апаратної платформи, що реалізує функціональні алгоритми штучного інтелекту відповідно до вимог. Табл. 1 містить спрощену модель якості ШІ — дворівневу структуру характеристик і підхарактеристик та їх відповідності гарантоздатності системи.

Ця модель є частиною так званої базової моделі якості ШІ [2], яка має трирівневу структуру та включає 32 характеристики. Спрощена дворівнева модель має п'ять характеристик першого рівня та 16 підхарактеристик, які утворюють другий рівень і деталізують характеристики ШІ. В табл. 1 виконано зіставлення характеристик гарантоздатності, які стосуються системи в цілому, і характеристик (підхарактеристик) ШІ.

Такі характеристики як етичність (Ethics, ETH) і законність (Lawfulness, LFL) мають зв'язки тільки з тими характеристиками СШІ, які стосуються функційної та інформаційної безпеки. Інші характеристики — поясненність (Explainability, EXP), відповідальність (Responsibility, RSB) і довірчоздатність (Trustworthiness, TST) та їх підхарактеристики мають зв'язки з усіма характеристиками ALR-моделі. Отже, дані табл. 1 дозволяють формувати DIS-модель, поєднуючи ALR- і AIQM-моделі.

Сценарії поведінки СШІ з порушеннями. Коректне функціонування СШІ визначається кількома обставинами. Крім можливого порушення її працездатності як звичайної комп'ютерної системи внаслідок фізичних дефектів апаратних засобів, проєктних дефектів

Таблиця 1

Характеристики	Визначення (здатність ШІ)	Підхрактеристики ШІ	Гарантоздатність (Dependability, DPD)						
			RLB	AVL	MNT	INT	CFD	SFT	SRV
Етичність (Ethics, ETH)	Відповідати сучасним стандартам моралі за результатами функціонування.	Справедливість (Fairness, FRN)				+	+		
		Сприйнятливість (Graspability, GRS)				+			
		Людська автономність (Human agency, HMA)					+		
		Відшкодовуваність (Redress, RDR)				+	+	+	
Законність (Lawfulness, LFL)	Дотримуватися законів і правил	—				+	+	+	
Поясненність (Explainability, EXP)	Бути зрозумілим і передбачуваним з точки зору мети та поведінки	Повнота (Completeness, CMT)		+	+			+	
		Зрозумілість (Comprehensibility, CMH)		+	+	+			+
		Інтерпретованість (Interpretability, INP)		+	+	+	+	+	+
		Інтерактивність (Interactivity, INR)			+			+	+
		Прозорість (Transparency, TRP)	+	+	+	+	+	+	+
		Верифікованість (Verifiability, VFB)	+	+	+	+	+	+	+
Відповідальність (Responsibility, RSP)	Працювати з урахуванням очікувань користувача відповідно до етичних і правових норм та інформувати його у разі порушень	—	+	+		+		+	+
Довірчоздатність (Trustworthiness, TST)	Забезпечувати необхідну ступінь впевненості зацікавлених сторін, розробників, аудиторів…, що ШІ відповідає та виконує свої функції передбачуваним чином	Точність (Accuracy, ACR)	+	+	+	+		+	+
		Диверсність (Diversity, DVS)		+		+		+	+
		Резильєнтність (Resilience, RSL)		+		+			+
		Робастність (Robustness, RBS)						+	+
		Функційна безпечність (Safety, SFT)		+				+	+
		Інформаційна безпека (Security, SCR)		+		+	+	+	+

програмних засобів та кібератак на вразливості платформної складової СШІ, додатковим і ключовим чинниками з точки зору особливостей власне модельної складової, некоректне, невизначене або потенційно небезпечне функціонування обумовлюється тим, на яких множинах даних здійснювали навчання, донавчання або перенавчання моделі ШІ.

Загальна множина вхідних даних ID(U) може бути декомпозована на підмножини даних, на яких ШІ навчався $ID(U_L)$ і був навчений коректно і коректно реагує на вхідні дані та їх послідовності $ID(U_{LC})$:

$$ID(U_{LC}) \subset ID(U_L) \subset ID(U).$$

На підмножинах вхідних даних

$$\Delta ID_N = ID(U_L) \backslash ID(U_{LC}),\ \Delta ID_U = ID(U) \backslash ID(U_L)$$

система працюватиме некоректно і невизначено відповідно. Отже множина

$$\Delta ID = \Delta ID_N \cup \Delta ID_U$$

утворює простір потенційно некоректного і небезпечного функціонування. Він може бути зменшений завдяки більш ретельному і повному навчанню або відповідні ризики мають толеруватися відповідними проєктними рішеннями залежно від різних сценаріїв функціонування.

Простір даних ID(U) може бути також декомпозовано за принципом впливу на характеристики ШІ (TST, RSP, EXP, ETH, LFL), для кожної з яких X є пімножини коректних, некоректних і невизначених даних:

$$ID(U_X) = ID(U_{X,C}) \cup ID(U_{X,N}) \cup ID(U_{X,U}).$$

Множину $ID(U_{EXP})$ на рис. 2 показано пунктиром, оскільки характеристика поясненості не пов'язана безпосередньо з процесами навчання. Додатковим чинником, який мають враховувати при побудові гарантоздатних СШІ, є визначення ролі ШІ з огляду на власну структуру і взаємодію із зовнішнім середовищем. Йдеться про те, що ШІ може бути [15—17]:

- кіберактивом C(AI), як є об'єктом захисту (в більш широкому контексті забезпечення гарантоздатності), що виконує відповідні функції системи;
- засобом підсилення захисту P(AI) кіберактивів (AI powered protection);
- засобом підсилення атакуючої сторони A(AI) (AI powered attacks).

Тоді маємо класичний змагальний ланцюжок (рис.3) і сім базових сценаріїв, в яких ШІ виконує хоча б одну з функцій: C(AI), P(AI), A(AI).

Найпростішим є сценарій «штучний інтелект захищається звичайними засобами від звичайних атак». Повним є сценарій «штучний інтелект захищається засобами, підсиленими штучним інтелектом проти атак, підсилених засобами штучного інтелекту». Деталізацію множини сценаріїв для різних типів P(AI) та A(AI) надано в [17], а розроблення засобів пошуку і колекціонування вразливостей C(AI) з використанням аналітики великих даних — в [19]. Крім того, ця множина розширюється залежно від аналізу та реалізації потенційних можливостей різних акторів (розробників стандартів і нормативних вимог, аудиторів функційної і кібербезпеки, розробників систем, операційного персоналу та замовників) [16].

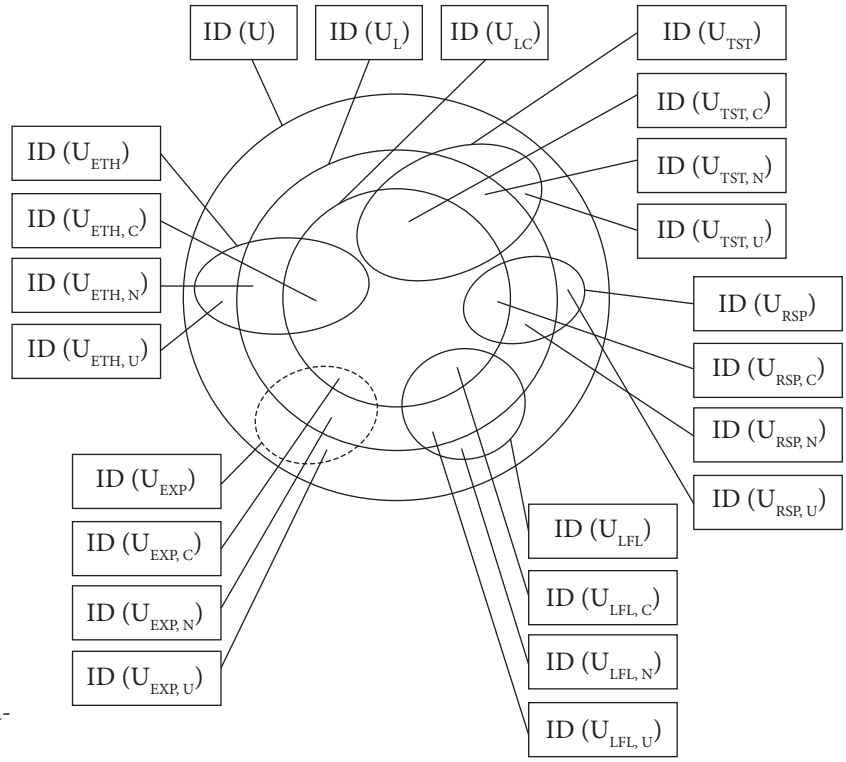

Рис. 2. Модель вхідних даних гарантоздатних СШІ

Для аналізу ризиків порушення безпеки і гарантоздатності для різних сценаріїв та їх зменшення розроблено різні модифікації IMECA-методології (Intrusion Modes and Effects Criticality Analysis) [20], зокрема, з використанням модельно-інформованого підходу (Security Informed Safety) [16].

Розвиток VPN для гарантоздатних СШІ. На підставі моделей гарантоздатності ШІ, можливих сценаріїв порушення функціонування важливо сформулювати концепцію побудови гарантоздатних СШІ**.** Природним є використання ідей фон Неймана [12] з огляду на еволюцію систем [13, 14] з врахуванням специфіки СШІ та рівня розвитку технологій. Формально парадигма VNP може бути описана кортежем [14]:

VNP = < {Proc}, {CharS}, {Syst}, from, {CharC}, {Comp}>,

де {Proc} — множина процесів (синтезу, розроблення, створення,…); {CharS} — множина характеристик системи (властивостей гарантоздатності — надійності, безпечності тощо); {Syst} — множина систем (пристроїв, комп'ютерних систем, IT-інфраструктур тощо), які синтезуються, розробляються, створюються,…; from — прийменник, який поєднує систему Syst та її компоненти Comp (Syst X from Comp Y); {CharC} — множина характеристик компонентів системи (зазвичай вони є антиподами множини CharS, на кшталт: ненадійний (небезпечний) або недостатньо надійний (безпечний); {Comp} — множина компонентів (логічні елементи, чипи, апаратні та програмні засоби, підсистеми…), які використовуються для побудови Syst.

Рис. 3. Схема ланцюжка ШІ проти ШІ

Для систем ШІ (сьомий етап еволюції, 2020 роки, відповідно до [14]), можливо кілька формулювань

VNP* залежно від того, які характеристики чи підхарактеристики моделі гарантоздатності розглядаються. Найзагальнішою і доволі зрозумілою, враховуючи набір підхарактеристик, є «Розроблення довірчоздатних СШІ (або ШІ) з недовірчоздатних або недостатньо довірчоздатних компонентів ШІ».

Потенційна можливість і методи реалізації VNP* для покращення характеристик СШІ надано в табл. 2.

Принцип диверсності для забезпечення гарантоздатності СШІ. З огляду результатів аналізу можливих варіантів впровадження VNP* для забезпечення гарантоздатності СШІ є використання принципу диверсності [14]. Пропонуються три стратегії впровадження цього принципу:

- диверсність датасетів і методів побудови версій (каналів) СШІ (наприклад, різних типів нейромереж, методик їх розроблення та реалізації);
- диверсність донавчання при виявленні проблем або отриманні додаткових датасетів;
- диверсність засобів контролю працездатності окремих версій (каналів) СШІ.

Це потребує розвитку теорії багатоверсійних гарантоздатних систем стосовно СШІ [21]. Опис способу резервування систем штучного інтелекту, який частково реалізує сформульовані принципи, надано в [22]. Він базується на багатоверсійній системі, яка скла-

Таблиця 2

Характеристика	Підхарактеристика	Можливість використання VNP*	Метод використання	Примітка
TST	TST	+	Методи, що використовуються для підхарактеристик	Сумісність методів, які використовуються для різних підхарактеристик, має бути врахована
	DVS	+	Версійна надмірність	Існує проблема розроблення і вибору версій з необхідною диверсністю
	RSL	+	Проактивність, версійна і структурна надмірність, динамічна реконфігурація	Так само. Крім того, існує проблема розроблення методів достовірного проактивного аналізу
	RBS	+	Версійна надмірність	Існує проблема розроблення і вибору версій з точки зору вхідних даних
	SFT	+	Версійна і структурна надмірність	Основний критерій — мінімізація ризиків відмов за загальною причиною
	SCR	+/–	Версійна надмірність для цілісності і доступності	Необхідно враховувати вплив версійної надмірності на інші підхарактеристики
	ACR	+	Версійна і структурна надмірність	Основний критерій — зменшення похибок обчислень
EXP	EXP	—	—	Надмірність може зменшити рівень поясненності
RSP	RSP	+	Версійна і структурна надмірність, динамічна реконфігурація	RSP залежить від довірчоздатності, поясненності та ін., що необхідно враховувати при виборі методів
ETH	ETH	—	—	Версійна надмірність може бути застосована, якщо версії зможуть забезпечити різну реакцію в ситуаціях з етично неприйнятними альтернативами
LFL	LFL	—	—	Так само з точки зору законності

дається з паралельно функціонуючих інтелектуальних автоматів — версій AIV_1,…, AIV_n; аналізатора даних —підмножин ID(U); перетворювача результатів аналізу в код керування вибором версій з врахуванням стану працездатності AIV_i; автомата реконфігурації, який визначає і керує конфігуруванням системи; засобів до навчання при федеративній організації. В [14] надано імовірнісні моделі для оцінювання гарантоздатності з врахуванням модельної диверсності та диверсності програмно-апаратної платформи.

Практичне використання. Запропоновану концепцію і принципи побудови і забезпечення гарантоздатності СШІ впроваджено при виконанні проєктів:

- створення мультисенсорних інтелектуальних платформ для систем пошуку і виявлення вибухонебезпечних предметів на базі роїв безпілотних літальних апаратів та робототобіологічних систем [17, 23, 24];
- багатоканальних інтелектуальних систем моніторингу об'єктів критичної енергоінфраструктури (АЕС, ММР) з використанням мобільної і стаціонарної підсистем, а також приватної хмарної структури [25—27];
- розроблення дорожньої карти використання засобів ШІ для забезпечення безпеки і гарантоздатності автономних транспортних систем в рамках європейського проєкту ECHO за програмою Horizon 2020 [16].

Крім того, продовжується розроблення та дослідження комплексів систем БПЛА і наземних роботів з розподіленими засобами ШІ для моніторингу та запобігання лісових пожеж в рамках шведсько-естонсько-українського проєкту WILDCAT (2024—2026), а також проєкту за замовленням МОН України 0124U000945 «Методи, засоби і технологія забезпечення гарантоздатності і резільєнтності інтелектуальних комплексів безпілотних літальних апаратів та безекіпажних апаратів із комбінованими стратегіями технічного обслуговування» (2024—2026) [28, 29].

Висновки. Концептуальні основи гарантоздатних систем ШІ представлено в даній роботі комплексом високорівневих моделей і принципів, що розвивають і гармонійно поєднують:

по-перше, ALR-концепцію гарантоздатних обчислень та AIQM-модель подання характеристик якості СШІ. Визначено їх зв'язки та спільні складові;

по-друге, можливі сценарії їх поведінки залежно від сегментації вхідних даних і датасетів, а також відмов внаслідок неповної визначеності, фізичних, проєктних дефектів та кібератак на вразливості моделей і платформ СШІ. Крім того, множина сценаріїв доповнюється варіантами використання ШІ як захищувального кіберактиву, засобу підсилення атак та засобу захисту з різними акторами процесів розроблення, аудиту та застосування СШІ;

по-третє, модифіковану парадигму VNP* побудови гарантоздатних СШІ та принцип диверсності, який використовується на модельному та платформному рівнях, що надає змогу підвищити довірчоздатність та інші характеристики ШІ з використанням множини багатоверсійних самоконтрольованих автоматів з донавчанням.

Представлені концептуальні моделі та принципи формують новий науковий напрям критичного інтелектуального комп'ютингу для систем та інфраструктур, важливих для безпеки. Термін комп'ютинг узагальнює методи та засоби комп'ютерних технологій, а також процеси їх використання. Практичне значення результатів підтверджується їх впровадженням в низці національних і міжнародних проєктів створення гарантоздатних СШІ.

Наступні кроки розвитку концепції гарантоздатних СШІ пов'язані з деталізацією принципів і розробленням методів створення природно резильєнтних інтелектуальних

систем та СШІ з обмеженими ресурсами на базі ідей, сформованих в роботах [11, 18, 28, 29]. Цікавими і важливими напрямами, на наш погляд, є:

▪ поєднання технологій ШІ, доповненої реальності та інтернету речей в інтерактивному мистецтві [30];

▪ розроблення інтелектуальних пацієнтоцентричних систем та дружніх інтерфейсів для СШІ в медицині [31, 32], де вимоги до надійності і безпеки є особливо жорсткими;

▪ розроблення гарантоздатних систем на базі ШІ як сервісу (AIaaS) [33], аналітичних та експериментальних методів їх оцінювання та розгортання для використання в мобільних системах;

▪ поєднання концепцій гарантоздатного інтелектуального комп'ютингу та ШІ на базі онтолінгвістичних моделей, нейромереж з глибоким навчанням і експертних систем [34], а також високопродуктивних обчислень [35];

▪ формування регулюючих вимог до програмного забезпечення СШІ на підставі загальних стандартів та їх профілювання [36], а також методів оцінювання виконання вимог.

ЦИТОВАНА ЛІТЕРАТУРА

1. Ding W., Abdel-Basset M., Hawash H., Ali A.M. Explainability of AI methods, applications and challenges: A comprehensive survey. *Information Science*. 2022. **615** (C). C. 238—292. https://doi.org/10.1016/j.ins.2022.10.013
2. Kharchenko V., Fesenko H., Illiashenko O. Basic model of non-functional characteristics for assessment of artificial intelligence quality. *Radioelectron. Comput. Syst*. 2022. No. 2. P. 1—14. https://doi.org/10.32620/reks.2022.2.11
3. Kharchenko V., Fesenko H., Illiashenko O. Quality Models for Artificial Intelligence Systems: Characteristic-Based Approach, Development and Application. *Sensors*. 2022. **22** (13). P. 1—32. https://doi.org/10.3390/s22134865
4. Williams R., Yampolskiy R. Understanding and Avoiding AI Failures: A Practical Guide. *Philosophies*. 2021. **53**, № *6*. P. 1—25. https://doi.org/10.3390/philosophies6030053
5. Steimers A., Schneider M. Sources of Risk of AI Systems. *Int. J. Environ. Res. Public Health*. 2022. **19**, № 6. P. 1—26. https://doi.org/10.3390/ijerph19063641
6. Ziwei Xu, Sanjay Jain, Mohan Kankanhalli. Hallucination is Inevitable: An Innate Limitation of Large Language Models. 2022. **22**, № 1. P. 1—26. https://doi.org/10.48550/arXiv.2401.11817
7. Avizienis A., Laprie J.-C., Randell B., Landwehr C. Basic concepts and taxonomy of dependable and secure computing. *IEEE Trans. Dependable Secur. Comput*. 2004. **1**. P. 11—33. https://doi.org/10.1109/TDSC.2004.2
8. Харченко В.С. Гарантоздатні системи та багатоверсійні обчислення: аспекти еволюції. *Радіоелектронні і комп'ютерні системи*. 2009. № 7. С. 46—59.
9. Харченко В.С. Гарантоздатність комп'ютерних систем: межа універсальності у контексті інформаційно-технічних станів. *Радіоелектронні і комп'ютерні системи*. 2007. № 8. С. 7—14.
10. Nikolas Guelfi. A Formal Framework for Dependability and Resilience from a Software Engineering Perspective. *Cent. Eur. J. Comp. Sci*. 2011. **1**(3). P. 294—328. https://doi.org/10.2478/s13537-011-0025-x
11. Moskalenko V., Kharchenko V., Moskalenko A., Kuzikov B. Resilience and Resilient Systems of Artificial Intelligence; Taxonomy, Models and Methods. *Algorithms*. 2023. **165,** № *16,* P. 1—34. https://doi.org/10.3390/a16030165
12. Neumann J. von. Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components. Princeton Univ. Press, 1956. P. 43—98. https://doi.org/10.1515/9781400882618-003
13. Kharchenko V., Gorbenko A. Evolution of von Neumann's paradigm: Dependable and green computing, *East-West Design & Test Symposium*. 2013. P. 1—6. https://doi.org/10.1109/EWDTS.2013.6673090
14. Kharchenko V., Odarushchenko O. Trustworthy AI Systems from Untrustworthy Components: Development von Neumann's Paradigm using Principle of Diversity. Proceedings 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024), September 25–27, 2024, Cambridge, MA, USA. P. 392—404.
15. Veprytska O., Kharchenko V. AI powered attacks against AI powered protection: classification, scenarios and risk analysis, 12th International Conference on Dependable Systems, Services and Technologies (DESSERT), IEEE. 2022. P. 1—7. https://doi.org/10.1109/dessert58054.2022.10018770

16. Illiashenko O., Kharchenko V., Babeshko I., Fesenko H., Giandomenico Di F., Security-Informed Safety Analysis of Autonomous Transport Systems Considering AI-Powered Cyberattacks and Protection. *Entropy.* 2023. **25**, № 8. P. 1—35. https://doi.org/10.3390/e25081123
17. Veprytska O., Kharchenko V. Analysis of AI powered attacks and protection of UAV assets: quality model-based assessing cybersecurity of mobile system for demining Proceeding of 5th International Workshop on Intelligent Information Technologies and Systems of Information Security, March 28, 2024, Khmelnytskyi, Ukraine. P. 356—374.
18. Moskalenko V., Kharchenko V. Resilience-aware MLOps for AI-based medical diagnostic system. *Front. Public Health*. 2024. **12**. 1342937. https://doi.org/10.3389/fpubh.2024.1342937
19. Neretin O., Kharchenko V., Fesenko H. Multi-source Analysis of AI Vulnerabilities: Methodology and Algorithms of Data Collection, *2023. IEEE 12th Int. Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Dortmund, Germany. 2023. P. 972—977. https://doi.org/10.1109/IDAACS58523.2023.10348671
20. Babeshko I., Illiashenko O., Kharchenko V., Leontiev K. Towards Trustworthy Safety Assessment by Providing Expert and Tool-Based XMECA Techniques. *Mathematics.* 2022. **10** (13). P. 1—29. https://doi.org/10.3390/math10132297
21. Сиора А.А., Краснобаев В.А., Харченко В.С. Отказоустойчивые системы с версионно-информационной избыточностью. Ред. В.С. Харченко. Х.: Нац. аерокосм. ун-т ім. Н.Е. Жуковского «ХАИ», 2009. 321 с.
22. Харченко В.С., Одарущенко О.М., Фесенко Г.В., Одарущенко О.Б. Спосіб резервування системи штучного інтелекту: патент на корисну модель № 156654, заявка № u202304653 ; заявл. 03.10.2023; опубл. 24.07.2024, бюл. № 30. https://sis.nipo.gov.ua/uk/search/detail/1809682
23. 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. *Radioelectron. Comput. Syst*. 2023. № 2. P. 143—159.
24. Федоренко Г.Л., Клюшніков І.М., Харченко В. С. та ін. Спосіб пошуку та розпізнавання вибухонебезпечних предметів: патент на корисну модель № 1542266, заявка № u202300129; заявл. 03.01.2023; опубл. 25.10.2023, бюл. № 43. https://sis.nipo.gov.ua/en/search/detail/1768299/
25. Fesenko H., Illiashenko O., Kharchenko V., Kliushnikov I., Morozova O., Sachenko A., Skorobohatko S. Flying Sensor and Edge Network-Based Advanced Air Mobility Systems: Reliability Analysis and Applications for Urban Monitoring. *Drones.* 2023. № *7*. P. 409. https://doi.org/10.3390/drones7070409
26. Ivanchenko O, Kharchenko V, Smyrynska N., Veprytska O. Cybersecurity of unmanned surface vessels: IMECA based assessment and protection against ai powered attacks. *Annual of Navigation*. 2024. https://doi.org/10.36163/aon-2024-0004
27. Mishchuk V., Fesenko H., Kharchenko V. Deep learning models for detection of explosive ordnance using autonomous robotic systems: trade-off between accuracy and real-time processing speed. *Radioelectronic and Computer Systems*. 2024. № 4. P. 99—111. https://doi.org/10.32620/reks.2024.4.09
28. Ponochovnyi Y., Kharchenko V. Dependability Assurance Methodology of Information and Control Systems using multipurpose service strategies. Radioelectron. *Comput. Systems*. 2020. № 3. P. 43—58. https://doi.org/10.32620/reks.2020.3.05
29. Moskalenko V., Kharchenko V., Semenov S. Model and Method for Providing Resilience to Resource-Constrained AI-System. *Sensors*. 2024. **24**. P. 5951. https://doi.org/10.3390/s24185951
30. Golembovska Olena, Kharchenko Vyacheslav. Technologies of Interactive Art. Encyclopedia of Libraries, Librarianship, and Information Science. Elsevier, 2025. **4**. P. 506—527 https://doi.org/10.1016/b978-0-323-95689-5.00152-8
31. Barmak O., Krak I., Yakovlev S., Radiuk P., Kuznetsov V. Toward explainable deep learning in healthcare through transition matrix and user-friendly features. *Frontiers in Artificial Intelligence*. 2024. **7**. 1482141. https://doi.org/10.3389/frai.2024.1482141
32. Vakulenko D.V., Palagin O.V., Sergienko I.V. et al. Algorithmization and Optimization Models of Patient-Centric Rehabilitation Programs*. *Cybern. Syst. Anal*. 2024. **60**. P. 736—752. https://doi.org/10.1007/s10559-024-00711-5
33. Veprytska O., Kharchenko V. Analysis of requirements and quality model-oriented assessment of explainable AI as a service. *Electronic Modeling*. 2022. **44**, № 5. P. 36—50. https://doi.org/10.15407/emodel.44.05.036

34. Palagin O., Kaverinskiy V., Malakhov K., Petrenko M. Fundamentals of the Integrated Use of Neural Network and Ontolinguistic Paradigms: A Comprehensive Approach. *Cybern. Syst. Anal.* 2024. **60**, № 1. P. 111—123. https://doi.org/10.1007/s10559-024-00652-z
35. Sergienko I.V., Molchanov I.N., Khimich A.N. Intelligent technologies of high-performance computing. *Cybern. Syst. Anal.* 2010. **46,** №5. P. 833—844. https://doi.org/10.1007/s10559-010-9265-3
36. Gordieiev O., Rainer A., Kharchenko V., Pishchukhina O., Gordieieva D. A Unified Approach to the Development of Technology-Based Software Quality Models on the Example of Blockchain Systems. *IEEE Access*. 2024. **12**. P. 11887—118889. https://doi.org/10.1109/ACCESS.2024.3448271

Надійшла до редакції 27.01.2025

REFERENCES

1. Ding, W., Abdel-Basset, M., Hawash, H. & Ali, A. M. (2022). Explainability of AI methods, applications and challenges: A comprehensive survey. Information Science, 615 (C), pp. 238-292. https://doi.org/10.1016/j.ins.2022.10.013
2. Kharchenko, V., Fesenko, H. & Illiashenko, O. (2022). Basic model of non-functional characteristics for assessment of artificial intelligence quality. Radioelectron. Comput. Syst., No. 2, pp. 1-14. https://doi.org/10.32620/reks.2022.2.11
3. Kharchenko, V., Fesenko, H. & Illiashenko O. (2022). Quality Models for Artificial Intelligence Systems: Characteristic-Based Approach, Development and Application. Sensors, 22(13), pp. 1-32. https://doi.org/10.3390/s22134865
4. Williams, R. & Yampolskiy, R. (2021). Understanding and Avoiding AI Failures: A Practical Guide. Philosophies, 53*, 6*, pp. 1-25. https://doi.org/10.3390/philosophies6030053
5. Steimers, A. & Schneider, M. (2022). Sources of Risk of AI Systems. Int. J. Environ. Res. Public Health, 19, 3641, pp. 1-26. https://doi.org/10.3390/ijerph19063641
6. Ziwei, Xu, Sanjay, Jain & Mohan Kankanhalli. (2022). Hallucination is Inevitable: An Innate Limitation of Large Language, 22, No. 1, pp.1-26. https://doi.org/10.48550/arXiv.2401.11817
7. Avizienis, A., Laprie, J.-C., Randell, B. & Landwehr, C. (2004). Basic concepts and taxonomy of dependable and secure computing. IEEE Trans. Dependable Secur. Comput., 1, pp. 11-33. https://doi.org/10.1109/TDSC.2004.2
8. Kharchenko, V. S. (2009). Dependable systems and multi-version computing: evolution issues. Radioelecntronni ta compjuterni systemy, No. 7, pp. 46-59 (in Ukrainian).
9. Kharchenko, V. S. (2007). Dependability of computer systems: boarder of universality in context of information-technical states. Radioelecntronni ta compjuterni systemy, No. 8, pp. 7-14 (in Ukrainian).
10. Nikolas, Guelfi. (2011). A Formal Framework for Dependability and Resilience from a Software Engineering Perspective. Cent. Eur. J. Comp. Sci., 1(3), pp. 294-328. https://doi.org/:10.2478/s13537-011-0025-x
11. Moskalenko, V., Kharchenko, V., Moskalenko, A. & Kuzikov, B. (2023). Resilience and Resilient Systems of Artificial Intelligence: Taxonomy, Models and Methods. Algorithms, *16*, 165, pp. 1-34. https://doi.org/10.3390/a16030165
12. Neumann, J. von. (1956). Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components. Ed. by C. E. Shannon, J. McCarthy, Princeton University Press, pp. 43-98. https://doi.org/10.1515/9781400882618-003
13. Kharchenko, V. & Gorbenko, A. (2013). Evolution of von Neumann's paradigm: Dependable and green computing, East-West Design & Test Symposium, pp. 1-6. https://doi.org/10.1109/EWDTS.2013.6673090
14. Kharchenko, V. & Odarushchenko, O. (2024). Trustworthy AI Systems from Untrustworthy Components: Development von Neumann's Paradigm using Principle of Diversity. Proceedings 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024), September 25–27, Cambridge, MA, USA, pp. 392-404.
15. Veprytska, O. & Kharchenko, V. (2022). AI powered attacks against AI powered protection: classification, scenarios and risk analysis, 12th International Conference on Dependable Systems, Services and Technologies (DESSERT), IEEE, pp. 1-7. https://doi.org/10.1109/dessert58054.2022.10018770
16. Illiashenko, O., Kharchenko, V., Babeshko, I., Fesenko, H. & Di, F. (2023). Giandomenico, Security-Informed Safety Analysis of Autonomous Transport Systems Considering AI-Powered Cyberattacks and Protection. Entropy, 25, No. 8, pp.1-35. https://doi.org/:10.3390/e25081123

17. Veprytska, O. & Kharchenko, V. (2024). Analysis of AI powered attacks and protection of UAV assets: quality model-based assessing cybersecurity of mobile system for demining Proceeding of 5th International Workshop on Intelligent Information Technologies and Systems of Information Security, March 28, Khmelnytskyi, Ukraine. P. 356-374.
18. Moskalenko, V. & Kharchenko, V. (2024). Resilience-aware MLOps for AI-based medical diagnostic system. Front. Public Health, 12, 1342937. https://doi.org/10.3389/fpubh.2024.1342937
19. Neretin, O., Kharchenko, V. & Fesenko, H. *(2023).* Multi-source Analysis of AI Vulnerabilities: Methodology and Algorithms of Data Collection, IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Dortmund, Germany, pp. 972-977. https://doi.org/10.1109/IDAACS58523.2023.10348671
20. Babeshko, I., Illiashenko, O., Kharchenko, V. & Leontiev, K. (2022). Towards Trustworthy Safety Assessment by Providing Expert and Tool-Based XMECA Techniques. Mathematics. 10, pp. 1-29. https://doi.org/10.3390/math10132297
21. Siora, A. A., Krasnobaev, V. A. & Kharchenko, V. S. (2009). Fault-tolerant systems with version-information redundancy. Ed. V. S. Kharchenko. Kharkiv: National aerospace university «KhAI», 321 p. (in Russian).
22. Kharchenko, V. S., Odarushchenko, O. M., Fesenko, H. V. & Odarushchenko, O. B. Method of intelligent system reservation: patent on know-how model № 156654, request № u202304653; submit. 03.10.2023; publish. 24.07.2024, bul. № 30 (in Ukrainian).
23. Fedorenko, G., Fesenko, H., Kharchenko, V., Kliushnikov, I. & Tolkunov, I. (2023). Robotic-biological systems for detection and identification of explosive ordnance: Concept, general structure, and models. Radioelectron. Comput. Syst., No. 2, pp. 143-159.
24. Fedorenko, G.., Kliushnikov, I., Kharchenko, V. et al. Method of search and recognition of explosive ordnance: patent on know-how model № 1542266, request № u202300129; submit. 03.01.2023; publish. 25.10.2023, bul. № 43 (in Ukrainian).
25. Fesenko, H., Illiashenko, O., Kharchenko, V., Kliushnikov, I., Morozova, O., Sachenko, A. & Skorobohatko, S. (2023). Flying Sensor and Edge Network-Based Advanced Air Mobility Systems: Reliability Analysis and Applications for Urban Monitoring. Drones, No. 7, pp. 409. https://doi.org/10.3390/drones7070409
26. Ivanchenko, O., Kharchenko, V., Smyrynska, N. & Veprytska, O. (2024). Cybersecurity of unmanned surface vessels: IMECA based assessment and protection against ai powered attacks. Annual of Navigation. https://doi.org/10.36163/aon-2024-0004
27. Mishchuk, V., Fesenko, H. & Kharchenko, V. (2024). Deep learning models for detection of explosive ordnance using autonomous robotic systems: trade-off between accuracy and real-time processing speed. Radioelectronic and Computer Systems, No. 4, pp. 99-111. https://doi.org/10.32620/reks.2024.4.09
28. Ponochovnyi, Y. & Kharchenko, V. (2020). Dependability Assurance Methodology of Information and Control Systems using multipurpose service strategies. Radioelectron. Comput. Syst., 3 ,pp. 43-58. https://doi.org/10.32620/reks.2020.3.05
29. Moskalenko, V., Kharchenko, V. & Semenov, S. (2024). Model and Method for Providing Resilience to Resource-Constrained AI-System. *Sensors*, *24*, 5951. https://doi.org/10.3390/s24185951
30. Golembovska Olena & Kharchenko Vyacheslav. (2025). Technologies of Interactive Art Encyclopedia of Libraries, Librarianship, and Information Science. Elsevier, pp. 506-527. https://doi.org/10.1016/b978-0-323-95689-5.00152-8
31. Barmak, O., Krak, I., Yakovlev, S., Radiuk, P. & Kuznetsov, V. (2024). Toward explainable deep learning in healthcare through transition matrix and user-friendly features. Frontiers in Artificial Intelligence, 7, 1482141. https://doi.org/10.3389/frai.2024.1482141
32. Vakulenko, D. V., Palagin, O. V., Sergienko, I. V. et al. (2024). Algorithmization and Optimization Models of Patient-Centric Rehabilitation Programs. Cybernetics and Systems Analysis, 60, pp. 736-752. https://doi.org/10.1007/s10559-024-00711-5
33. Veprytska, O. & Kharchenko, V. (2022). Analysis of requirements and quality model-oriented assessment of explainable AI as a service. Electronic Modeling, 44, No. 5, pp. 36-50. https://doi.org/10.15407/emodel.44.05.036
34. Palagin, O., Kaverinskiy, V., Malakhov, K. & Petrenko, M. (2024). Fundamentals of the Integrated Use of Neural Network and Ontolinguistic Paradigms: A Comprehensive Approach. Cybern. Syst. Anal., 60, No. 1, pp. 111-123. https://doi.org/10.1007/s10559-024-00652-z

35. Sergienko, I. V., Molchanov, I. N. & Khimich, A. N. (2010). Intelligent technologies of high-performance computing. Cyber. Syst. Anal., 46, No. 5, pp. 833-844. https://doi.org/10.1007/s10559-010-9265-3
36. Gordieiev, O., Rainer, A., Kharchenko, V., Pishchukhina, O. & Gordieieva, D. (2024). A Unified Approach to the Development of Technology-Based Software Quality Models on the Example of Blockchain Systems. *IEEE Access,* 12, pp. 118875-118889. https://doi.org/10.1109/ACCESS.2024.3448271

Received 27.01.2025

V.S. Kharchenko, https://orcid.org/0000-0001-5352-077X

National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine
E-mail: v.kharchenko@csn.khai.edu

CONCEPTUAL FUNDAMENTALS OF DEPENDABLE ARTIFICIAL INTELLIGENCE SYSTEMS

The concept of dependent artificial intelligence (AI) systems based on the development of the von Neumann paradigm (VNP) is proposed. It is presented by a set-theoretic description that takes into account different qualitative characteristics of AI and AI systems (AIS). The stages and formulations of the evolution of VNPs from simple relay units to complex digital infrastructures and AIS are analyzed. One of the stages of VNP development is related to the fundamental work on the concepts and taxonomy of reliable and secure computing (A. Avizienis et. al., 2004). The AIS Quality Model (QM) is described as an ordered hierarchy of attributes (characteristics) of trustworthiness, explainability, ethicality, legality, responsibility and their specific sub-characteristics, which allows to determine the possibilities of applying VNP to ensure the required values of the characteristics. VNP is formulated for AIS in various representations such as " trustworthy AIS form untrustworthy components". AIS QM consists of the AI quality model and the QM of the system's hardware-software platform. Application examples of AIS QM are analyzed. A model for matching and transforming input data into AIS output data is proposed, taking into account the decomposition of a universal data set into subsets used for training and possible anomalies in certain quality characteristics, as well as different types of failures and cyberattacks on AIS. It is proposed to use the principle of diversity in the implementation of VNP to ensure reliability and other characteristics of AI and to create dependent AIS. Models of reliable multiversion AIS are described and methods of reliability improvement are considered. These methods are based on various online testing schemes and application of versioning and structural redundancy.

Keywords*: dependable computing, trustworthy artificial intelligence, AI system quality model, von Neumann's paradigm, principle of diversity.*