УДК 004.8.05 **doi: 10.32620/reks.2022.2.11**

В. С. ХАРЧЕНКО, Г. В. ФЕСЕНКО, О. О. ІЛЛЯШЕНКО

Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна

БАЗОВА МОДЕЛЬ НЕФУНКЦІЙНИХ ХАРАКТЕРИСТИК ДЛЯ ОЦІНКИ ЯКОСТІ ШТУЧНОГО ІНТЕЛЕКТУ

Предметом дослідження *є моделі якості штучного інтелекту (ШІ).* ***Метою статті*** *є розроблення моделі якості ШІ на основі визначення та упорядкування його характеристик.* ***Задачі:*** *сформулювати принципи та обґрунтувати послідовність аналізу і розроблення моделей якості ШІ як впорядкованих множин характеристик; запропонувати моделі якості ШІ задля подальшого використання, перш, за все, оцінювання окремих характеристик і якості в цілому; продемонструвати профілювання моделей якості ШІ для систем, де використовується штучний інтелект. Було отримано наступні* ***результати****. Запропоновано послідовність побудови моделей якості ШІ. На підставі аналізу посилань сформовано список характеристик ШІ і здійснено гармонізацію їх визначень. Представлено загальну модель якості ШІ з наданням опису покрокової процедури реалізації її ієрархічної побудови. Запропонована базова модель ШІ зі скороченими множинами характеристик з огляду на їх важливість. Надано приклади профілювання моделей якості для двох систем – моніторингу інженерних комунікацій і розпізнавання дорожніх знаків.* ***Висновки.*** *Основним результатом проведеного дослідження є розробка моделі якості для штучного інтелекту, яка базується на аналізі та гармонізації визначень та залежностей характеристик якості, які є специфічними для ШІ. Вибір характеристик та побудова моделі якості здійснювалось таким чином, щоб виключити повторення, забезпечити повноту представлення, а також визначити специфічні ознаки кожної з характеристик. Зрозуміло, що зробити модель, яка б повністю відповідала таким вимогам вкрай важко, тому представлені варіанти мають доповнюватися та удосконалюватися з урахуванням швидкого розвитку технологій і застосувань ШІ. Запропоновані моделі якості є відкритими і можуть доповнюватися і деталізуватися відповідно до специфіки призначення та сфери використання ШІ.*

Ключові слова: *штучний інтелект; характеристики штучного інтелекту; модель якості штучного інтелекту; профілювання моделей якості штучного інтелекту*.

Вступ

Мотивація. Якість життя, безпека окремих людей і, навіть, країн залежать від інформаційних технологій, серед яких найбільш складними і дещо суперечливими є технології штучного інтелекту (ШІ). Динаміка впровадження систем ШІ (СШІ) в різних сферах, інтенсивні розробки і дослідження супроводжуються стрімким збільшенням кількості публікацій за останні три роки [1], численних технічних звітів і стандартів європейських інституцій [2, 3], ISO/IEC [4-6], IEEE [7], NIST [8-10], OECD [11], UNESCO [12].

В індустріальних системах, медицині, транспорті, системах озброєння, юриспруденції тощо вплив ШІ стає, з одного боку, все більш відчутним і сталим, а з іншого, - вельми суперечливим, що обумовлено кількома чинниками:

- складністю рішень, які приймаються при розробленні та застосуванні систем, в які вбудовано засоби ШІ;
- змінним та не завжди визначеним фізичним та інформаційним середовищем, в якому вони функціонують СШІ. Зростає інтенсивність і розширюється номенклатура зовнішніх впливів, кібератак, які, з одного боку, спрямовані на штучний інтелект, з іншого, – базуються на методах ШІ;
- накопиченням експертної інформації та розширенням баз знань, які можуть бути використані для підвищення ефективності СШІ. Принцип людиноцентричності при їх створенні та застосуванні має бути збалансованим задля мінімізації ризиків прийняття помилкових рішень внаслідок суб'єктивних причин;
- зростанням ваги етичних і безпекових аспектів впродовж використання. Цей чинник є особливо важливим і специфічним для СШІ. Відповідно до [12] та інших документів, які зосереджуються на гуманітарних аспектах, людська гідність, персональна і колективна безпека та благополуччя є ціннісними орієнтирами при розробленні та впровадженні систем ШІ.

Означені обставини ускладнюють, на відміну від «традиційних» систем, формулювання специфікацій та перевірку виконання вимог при створенні та модернізації СШІ. Крім того, зростає кількість і

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

різноманітність характеристик ШІ та СШІ, які мають бути враховано, зокрема, таких як етичність, пояснюваність, довірчоздатність тощо [13-15]. В свою чергу, урізноманітнюються методи оцінювання, що мають базуватися на чіткому уявленні про сутність і взаємозалежність характеристик ШІ.

Слід підкреслити, що зростання кількості публікацій і стандартів супроводжується суттєвою невпорядкованістю характеристик ШІ, яка, з одного богу, обумовлює, з іншого, - обумовлюється неузгодженістю їх визначень. Отже, вкрай важливими є дослідження задля гармонізації та ієрархізації характеристик, що надасть об'єктивності і спростить розроблення інструментарію нормування, оцінювання та забезпечення вимог при створенні та впровадженні СШІ.

Мета і структура. Метою дослідження є розроблення моделі якості штучного інтелекту на основі визначення та упорядкування характеристик. Ці характеристики називаємо нефункційними за аналогією з характеристиками програмного забезпечення та ІТ-систем, оскільки вони є загальними для різних застосунків. Задачі дослідження полягають у наступному:

- сформулювати принципи та обґрунтувати послідовність аналізу і розроблення моделей якості ШІ як впорядкованих множин характеристик;
- запропонувати моделі якості ШІ задля подальшого використання, перш, за все, оцінювання окремих характеристик і якості в цілому;
- продемонструвати варіанти профілювання моделей якості ШІ для систем моніторингу інженерних комунікацій і розпізнавання дорожніх знаків.

Стаття структурується у такий спосіб. В наступному розділі обґрунтовуються принципи та послідовність розроблення моделей якості. В *другому* розділі пропонуються підходи щодо формулювання визначень характеристик ШІ на підставі аналізу існуючих та їх гармонізації з урахуванням різних груп джерел, а також надається таблиця з визначеннями і класифікацією характеристик якості ШІ. Наступний *третій* розділ описує загальну модель ШІ з наданням опису покрокової процедури реалізації її ієрархічної побудови; представлено так звану базову модель зі скороченими множинами характеристик з огляду на їх важливість. У *четвертому* розділі описуються приклади моделей якості для двох систем штучного інтелекту. Останній розділ надає висновки і описує напрямки подальших досліджень.

1. Принципи та послідовність досліджень

Сукупність характеристик СШІ, які аналізуються в статті, об'єднуються поняттям «якість» за аналогією з тим, як це зазвичай робиться для програмного забезпечення, де існують сталі моделі якості, що розвивалися і удосконалювалися впродовж майже 55 років еволюції [16-18]. Поняття «якість» ШІ, на наш погляд, є прийнятною узагальнюючою характеристикою, не зважаючи на те, що деколи її використовують як часткову характеристику СШІ або розглядають якість штучного інтелекту суто в контексті якості програмного забезпечення [19].

Контекст якості програмного забезпечення є дійсно дуже важливим, але він має використовуватися як підхід для формування більш загальної моделі якості ШІ. В [20] формується думка, аналогічна позиції авторів даного дослідження щодо важливості саме якості АІ. Однак робота [20] дещо звужує зміст якості, оскільки набір характеристик ШІ, які аналізуються, є обмеженим. Отже надалі ми використовуємо поняття якості ШІ як системоутворюючої, верхньорівневої сутності в ієрархії всіх характеристик згідно із загальним тлумаченням якості за стандартом ISO 9001:2015, тобто ступеня, в який набір властивих об'єкту (в даному випадку ШІ) характеристик відповідає вимогам.

Якість СШІ складається з якості власне штучного інтелекту як узагальненого об'єкту і якості програмно-апаратної платформи, за допомогою якої ШІ реалізується. Це дослідження розглядає складові якості (характеристики) тільки ШІ.

Ключовим поняттям, яке використовується в дослідженні, є «характеристика» - складова якості, що описує різні властивості ШІ. Характеристика є базою для формулювання вимог до системи штучного інтелекту та її компонентів шляхом:

- урахування відповідної характеристики при розробленні специфікації – переліку вимог до СШІ;
- визначення метрик, за допомогою яких оцінюється значення характеристики, а саме, шкали і методики вимірювань (оцінювання);
- обґрунтування необхідних «меж» цієї характеристики, тобто вимог до її якісного або кількісного рівня, що визначаються відповідними метриками.

Якщо характеристика штучного інтелекту ChAI-1 є залежною від характеристики ChAI-2, то ChAI-2 називатимемо підхарактеристикою характеристики ChAI-1. Відповідно ChAI-1 і ChAI-2 мають розташовуватися на верхньому та наступному нижньому рівнях ієрархії моделі якості. Для кожної з характеристик (підхарактеристик) мають бути визначені метрики для їх оцінювання, а також сформульовано вимоги до значень характеристик, розроблено профіль вимог та оцінено його якість [22].

Послідовність побудови моделей якості ШІ є такою: на підставі аналізу посилань формується список характеристик ШІ і здійснюється гармонізація їх визначень. Результатом є таблиця 1 (розділ 2);

далі пропонуються моделі якості ШІ у графовій формі і надаються приклади профілювання моделей якості для двох систем – моніторингу інженерних комунікацій і розпізнавання дорожніх знаків (підрозділи 4.1 і 4.2).

2. Відбір і гармонізація визначень характеристик

Відбір і гармонізацію визначень характеристик ШІ було виконано наступним чином.

1. При аналізі визначень враховувалося, що окремі характеристики можуть бути тотожними, тобто такими, що мають різну назву, але однакову сутність. З підмножин таких характеристик залишалася одна для подальшого використання. Наприклад, з поміж характеристик «explicability» та «explainability», які означають «пояснюваність» (ability to be explained), для подальшого розгляду обрана характеристика «explainability».

2. Деякі характеристики з несуттєвою відмінністю були об'єднані, а у відповідних визначеннях ці відмінності враховувалися. Наприклад, характеристика «людський нагляд і рішучість» («human oversight and determination») була поглинута характеристикою «людський нагляд» («human oversight») з урахуванням особливостей проведення нагляду за ШІ, запропонованих в поглинутій характеристиці, у кінцевому визначенні.

3. Кілька характеристик були виключені, оскільки вони не мали специфічних ознак для ШІ, а є загальними для технічних систем або їх програмно-апаратного забезпечення. До таких характеристик, зокрема, належать «впевненість» («confidence») та «відповідність» («compliance»).

4. Для характеристик, які є суттєвими для ШІ, визначення надано шляхом:

- повторення (цитування) або несуттєвого коригування визначення з одного з документів, яке є найбільш адекватним і точним, на думку авторів (позначено літерою R – referred). Наприклад, визначення характеристики «цілісність» («integrity») була подано у відповідності з [21];
- гармонізації визначення на підставі визначень, які надаються в різних публікаціях (позначено літерою H – harmonized). Сутність гармонізації полягала у виявленні ключових термінів і поєднанні суттєвих складових різних визначень характеристики, що аналізувалася. Наприклад, визначення характеристики «цілісність» («integrity») було отримано шляхом поєднання суттєвих складових визначень цієї характеристики, запропонованих у [5, 21, 23];
- визначення, яке надано авторами у разі відсутності або незадовільного на їх думку формулювання для характеристики в доступних джерелах (позначено літерою A – authored). Наприклад, у такий спосіб було отримано визначення характеристики «resiliency» («резильєнтність»).

Результати аналізу джерел [1-3, 5, 6, 8-15, 20, 21, 23-36] і гармонізації визначень характеристик штучного інтелекту надано в таблиці 1. Таким чином, було відібрано 32 характеристики.

Визначення чотирьох характеристик вибрано з відповідних джерел без змін; визначення 25 характеристик було гармонізовано, визначення трьох характеристик є авторським.

Таблиця 1

Результати аналізу і гармонізації визначень характеристик штучного інтелекту

№ з/п	Назва характеристики	Визначення	Спосіб отримання	Джерело
1	Верифікованість (verifiability, VFB)	здатність ШІ, яка характеризується ступенем пристосованості до проведення верифікації різними методами.	H	[36]
2	Відповідальність (responsibility, RSP)	здатність ШІ функціонувати з урахуванням очікувань замовника (користувача) у відповідності до етичних норм, законодавчих нормативно-правових актів, а також інформувати його у разі можливого їх порушення.	H	[12, 26, 31]
3	Відстежуваність (accountability, ACN)	здатність ШІ надавати звіти за визначеною формою про результати функціонування у прозорий спосіб.	R	[21, 27]
4	Відшкодовуваність (redress, RDR)	здатність ШІ надавати доступні механізми забезпечення адекватного відшкодування наслідків негативного впливу на людей.	H	[2, 3]

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

№ з/п	Назва характеристики	Визначення	Спосіб отримання	Джерело
5	Диверсність (diversity, DVS)	здатність ШІ мінімізувати ризик невиконання специфікованих (визначених за необхідністю) функцій або завдань внаслідок відмов, обумовлених фізичними та інформаційними чинниками, з використанням різних моделей, алгоритмів та інших засобів.	A	[6]
6	Довірчоздатність (trustworthiness, TST)	здатність ШІ, яка характеризується ступенем впевненості користувача або іншої зацікавленої особи (розробника, аудитора тощо) в тому, що ШІ відповідає вимогам і виконує функції у передбачуваний спосіб.	H	[5, 10, 15 23]
7	Етичність (ethics, ETH)	здатність ШІ відповідати діючим нормам моралі за результатами функціонування.	H	[2, 3]
8	Завершеність (completeness, CMT)	здатність ШІ бути цілісним з точки зору ступеня відповідності всім вимогам замовника.	H	[21]
9	Законність (lawfulness, LFL)	здатність ШІ відповідати законодавчим і нормативно-правовим актам.	R	[2, 3]
10	Захищеність (security, SCR)	здатність ШІ захищати інформаційні та фізичні активи таким чином, щоб інші невизначені (неавторизовані) особи чи системи, включаючи ШІ, не мали б доступу до них або мали б такий доступ відповідно до визначеного типу і рівня авторизації.	H	[12, 21, 27]
11	Зміщеність (bias, BIS)	характеристика ШІ, яка визначає ризики появи результатів, які упереджені через хибні припущення та помилки в процесі налаштування моделей (наприклад, машинного навчання).	H	[2, 3, 6, 9]
12	Зрозумілість (comprehensibility, CMH)	здатність ШІ забезпечувати для користувача (або полегшувати користувачеві) розуміння пояснень, достатніх для того, щоб надати змогу застосувати ШІ або інформацію, отриману за його допомогою, для виконання інших завдань.	A	[29]
13	Інтерактивність (interactivity, INR)	здатність ШІ забезпечувати ефективну і проактивну взаємодію з користувачем.	H	[25]
14	Інтерпретабельність (interpretability, INP)	здатність ШІ надавати та інтерпретувати інформацію у зрозумілий для користувача спосіб.	H	[32]
15	Людська автономність (human agency, HMA)	здатність ШІ надавати користувачу можливість приймати автономні обґрунтовані рішення щодо застосування ШІ.	H	[2, 3]
16	Людський нагляд (human oversight, HMO)	здатність ШІ надавати можливості користувачу контролювати і при необхідності втручатися визначеним чином в функціонування ШІ.	H	[2, 3]
17	Недискримінативність (non-discrimination, NDS)	здатність ШІ забезпечувати виконання етичних норм щодо відсутності дискримінації за будь-якими ознаками.	H	[2, 3, 12]
18	Об'єктивність (objectivity, OBC)	здатність ШІ запобігати використанню скомпроментованих або сфальсифікованих даних.	R	[33]
19	Пояснюваність (explainability, EXP)	здатність ШІ бути зрозумілим і передбачуваним з точки зору призначення та поведінки.	H	[1-3, 8, 31]
20	Приватність (privacy, PRV)	здатність ШІ забезпечувати право розпоряджатися особистою інформацією у відповідності до вимог користувача.	H	[2, 3, 20, 33]
21	Прийнятність (acceptability, ASP)	здатність ШІ забезпечувати хоча б часткову його відповідність вимогам замовника або очікуванням споживача	H	[24]
22	Причинність (causability, CSL)	здатність ШІ визначати причино-наслідкові зв'язки між подіями, що виникають під час його застосування.	H	[28]
23	Простежуваність (traceability, TRC)	здатність ШІ простежувати виконання вимог у зручний для користувача спосіб, здійснювати пошук та документування помилок, невідповідностей на кожному етапі життєвого циклу.	H	[2, 3, 34]

№ з/п	Назва характеристики	Визначення	Спосіб отримання	Джерело
24	Резильєнтність (resiliency, RSL)	здатність ШІ продовжувати функціонування в умовах зміни вимог, параметрів фізичного та інформаційного середовища, а також виникнення неспецифікованих порушень і відмов.	A	[27]
25	Робастність (robustness, RBS)	здатність ШІ коректно працювати в широкому діапазоні вхідних даних та умов експлуатації і переходити у стан призупинення системи у разі виходу цих даних і умов за специфіковані межі.	H	[2, 3, 11, 20]
26	Соціальне благополуччя (societal well-being, SWB)	здатність ШІ враховувати соціальні процеси і не шкодити фізичному та психічному почуттю людей та благополуччю суспільства в цілому.	H	[2, 3]
27	Справедливість (fairness, FRN)	здатність ШІ мінімізувати ризики аномалій, обумовлених упередженістю при прийнятті рішень, які пов'язані з виконанням етичних норм (включаючи відсутність фаворитизму, дискримінацію за релігійними, расовими та іншими ознаками, тощо), а також хибних припущень і помилок в процесі налаштування моделей.	H	[2, 3, 11]
28	Сприйнятливість (graspability, GRS)	здатність ШІ забезпечувати можливості користувачу критичного сприйняття ШІ в рамках відкритого і демократичного середовища.	H	[30]
29	Точність (accuracy, ACR)	здатність ШІ забезпечувати близькість результатів виконання вимог та/або функцій, які представляються певними даними, до їх справжніх значень.	H	[2, 3, 23, 35]
30	Транспарентність (transparency, TRP)	здатність ШІ описувати, перевіряти та відтворювати моделі, окремі компоненти та алгоритми, за якими приймаються рішення.	H	[2, 3, 26, 27]
31	Функційна безпечність (safety, SFT)	здатність ШІ не припускати ризики неприйнятних пошкоджень і втрат внаслідок відмов, обумовлених внутрішніми і зовнішніми причинами, та мінімізувати їх наслідки з використанням засобів, вбудованих в ШІ.	H	[2, 3, 12]
32	Цілісність (integrity, ING)	здатність ШІ, яка характеризується ступенем запобігання несанкціонованому доступу задля модифікації алгоритмів або даних, використовуваних системою.	R	[21]

3. Базова модель якості ШІ

3.1. Послідовність побудови моделі якості ШІ

При побудові ієрархії на цьому і подальших етапах використовуємо наступну процедуру:

Крок 1. Кожну з характеристик S_{ChAI} співставляємо з усіма іншими і вибираємо такі, які залежать від інших, і є такими, від яких не залежать всі інші (відношення залежності визначається експертним шляхом). Такі характеристики мають бути віднесено до першого рівня ієрархії S_{ChAI-1} (з потужністю m_1);

Крок 2. Характеристики, які не ввійшли до S_{ChAI-1}, тобто сформували множину

$$S_{ChAI-2} = S_{ChAI} \setminus S_{ChAI-1},$$

розділяються на m_1 підмножин $S_{ChAI-2i}$, які не перетинаються і впливають на відповідні характеристики з множини S_{ChAI-1}:

$$S_{ChAI-2i} = \cup S_{ChAI-2i};\ i = \{1,2,\ldots, m_1\},$$
$$\forall\ i, j = \{1,2, \ldots, m_1\},\ i \neq j:\ S_{ChAI-2i} \cap S_{ChAI-2j} = \varnothing.$$

Крок 3. Операції 1,2 повторюються для кожної з підмножин $S_{ChAI-2i}$ потужністю m_{2i}, що надає змогу сформувати другий і третій рівень ієрархії.

Ця процедура продовжується далі, у разі більшої кількості півнів ієрархії. Моделі якості представлено у *графовій* формі, найбільш наочній та зручній для подальшого використання з метою оцінювання якості ШІ. В графі вершини відповідають характеристикам і підхарактеристикам, а ребра – відношенням залежності між ними.

Відповідно до покрокової процедури формуємо множину характеристик першого півня. До таких віднесено характеристики S_{ChAI-1} = {ETH, EXP, LFL, RSP, TST}, оскільки вони є найбільш вживаними і впливають безпосередньо на якість ШІ.

Характеристиками другого рівня (підхарактеристиками) є такі:

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

- для ETH: $S_{ChAI-21}$ = {FRN, GRS, HMA, HMO, RDR};
- для EXP: $S_{ChAI-22}$ = {ACN, CSL, CMT, CMH, TRP, INP, INR, VFB};
- для LFL, RSP: $S_{ChAI-23} = S_{ChAI-24} = ø$;
- для TST: $S_{ChAI-25}$ = {DVS, RSL, RBS, SFT, SCR, ASP, ACR}.

Далі характеристиками третього рівня є S_{ChAI-1} = {BIS, NDS, TRC, SWB, PRV, ING, OBC}.

Графова форма моделі надана на рис. 1.

3.2. Особливості базової моделі якості ШІ

Базова модель якості ШІ розробляється для того, щоб зробити її більш компактною і зручною для інженерної практики для оцінювання реальних систем ШІ. Базова модель може бути отримана шляхом її оптимізації по «вертикалі» і «горизонталі» і відрізняється від загальної тим, що:

- оптимізація по вертикалі здійснюється шляхом представлення моделі двома рівнями. Підхарактеристики третього рівня ураховуються на рівні метрик відповідних характеристик другого рівня;
- відповідні складові характеристик, які видаляються або об'єднуються, можуть бути враховані на рівні метрик, що використовуються для оцінювання, та їх зважування відповідним чином при оцінюванні характеристики верхнього рівня;
- видалено характеристику RSP: вона перетинається з іншими характеристиками цього рівня:

а) довірчоздатністю TST – з точки зору відповідальності за виконання вимог користувача в цілому. Крім того, вимога щодо інформування у разі можливого їх порушення, яка є складовою відповідальності, може розглядатися як обов'язкова і враховуватися при оцінюванні довірчоздатності;

в) пояснюваністю EXP – з точки зору пристосованості до перевірки та надання інформації у разі порушення відповідних норм і вимог, що є складовими підхарактеристик TRP, VFB;

- характеристики HMA і HMO об'єднуються, оскільки вони, зазвичай, розглядаються разом і можуть доповнюватися на рівні метрик. Нове визначення HMA: здатність ШІ на підставі контролю надавати користувачу можливість приймати автономні обґрунтовані рішення щодо застосування і втручатися визначеним чином в функціонування ШІ;
- відстежуваність ACN і причинність CSL об'єднуються з TRP, оскільки можуть розглядатися як додаткові метрики транспарентності. Тоді транспарентність може визначатися як здатність ШІ описувати, перевіряти та відтворювати моделі, окремі компоненти та алгоритми, за якими приймаються рішення, визначати причино-наслідкові зв'язки між подіями і надавати звіти за визначеною формою про результати функціонування;
- характеристика ACP виключена як окрема, оскільки вона фактично є «м'якою» складовою власне TST, визначення якої не потребує коригування.

Базова модель описується графом (рис. 2), який є підграфом загальної моделі і містить 19 характеристик.

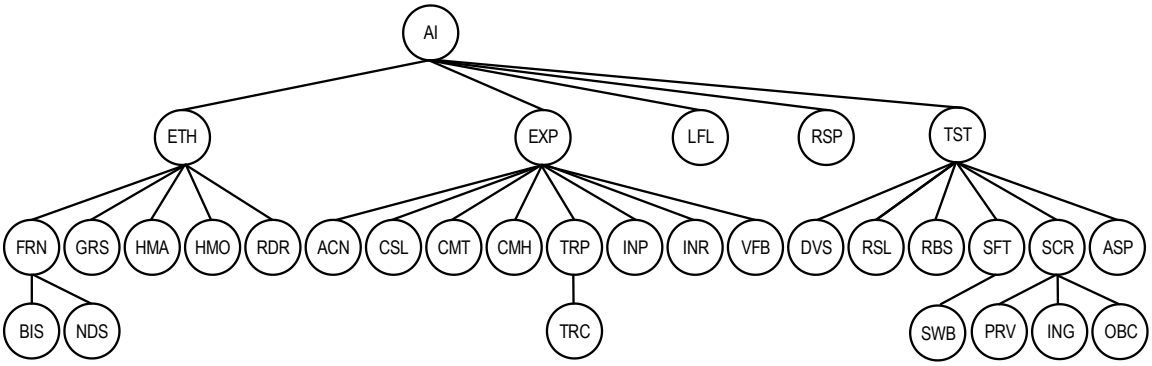

Рис. 1. Графова форма вихідної моделі якості штучного інтелекту

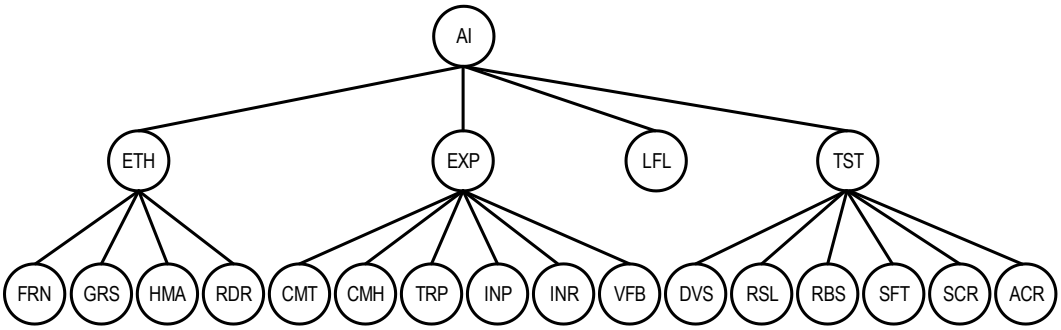

Рис. 2. Графова форма базової моделі якості штучного інтелекту

4. Приклад побудови моделей якості для систем штучного інтелекту

Розглянемо приклади побудови моделі якості для реальних систем штучного інтелекту на підставі запропонованої базової моделі.

Такі моделі можуть бути використано для обґрунтування вимог до розроблюваних систем або перевірки їх виконання і коригування проєктних рішень. Процес побудови моделей для реальних СШІ може називатися розробленням профілю або профілюванням вимог. Профілювання реалізується шляхом визначення характеристик якості на кожному рівні ієрархії моделі, які є важливими для системи, що аналізується.

Ця задача розв'язується далі у експертний спосіб для двох систем з використанням базової моделі якості ШІ (див. рис. 2). Мета цієї частини дослідження – продемонструвати як можуть застосовуватися моделі на підставі запропонованих в розділі 3.

4.1. Система моніторингу інженерних комунікацій

Перший приклад стосується системи моніторингу інженерних комунікацій (СМІК), завдання якої полягає у розпізнаванні дефектів на стінках стічних труб [37]. У системі реалізується багатоетапний метод машинного навчання, перший етап якого полягає в контрастному самонавчанні на нерозмічених даних, а наступні етапи пов'язані з визначенням двійкового коду кожного класу, що використовується як мітка під час точного налаштування моделі. Модель якості СМІК як системи штучного інтелекту представлено на рис. 3.

Її особливості є наступними:

- на першому рівні включено дві характеристики якості ШІ: пояснюваність EXP і довірчоздатність TST; етичність ETH і законність LFL виключені з розгляду, оскільки необхідність відповідати нормам моралі і права для системи не є природною;
- серед підхарактеристик для пояснюваності EXP включено всі підхарактеристики за виключенням зрозумілості CMH, враховуючі автономний режим роботи СМІК; для довірчоздатності TST також включено всі підхарактеристики за виключення диверсності DVS, оскільки застосування принципу багатоверсійності в СМІК обмежується необхідністю мінімізувати габаритно-масові та енергетичні показники.

4.2. Система розпізнавання дорожніх знаків

Другий кейс ілюструє побудову профіля якості для системи розпізнавання дорожніх знаків (СРДЗ), яка базується на технології розпізнавання зображень з використанням згорткових нейронних мереж [38]. У запропонованій системі вдосконалено етапи нормалізації та сегментації. На етапі нормалізації перед еквалізацією проводиться афінне перетворення зображення. Для сегментації та розпізнавання номерного знаку використовується нейронна мережа Mask R-CNN.

Модель якості СРДЗ як системи штучного інтелекту представлено на рис. 4.

Її особливості є наступними:

- на першому рівні модель включає три характеристики (етичність ETH, пояснюваність EXP і довірчоздатність TST), законність LFL виключена з розгляду з причини відсутності нормативно-правових актів, які б регулювали правові аспекти застосування подібних систем;
- для етичності включена одна характеристика – людська автономність HMA, оскільки користувачу в ряді випадків може надаватися можливість остаточного прийняття рішень щодо результатів роботи СРДЗ;
- серед підхарактеристик для пояснюваності EXP включено всі підхарактеристики за виключенням зрозумілості CMH та інтерактивності INR, враховуючі автономний режим роботи СМІК;

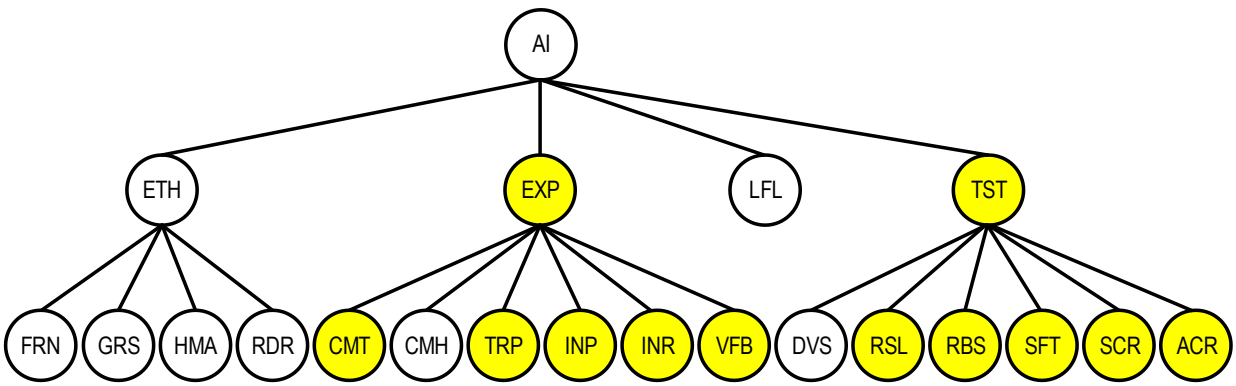

Рис. 3. Графова форма базової моделі якості системи моніторингу інженерних комунікацій

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

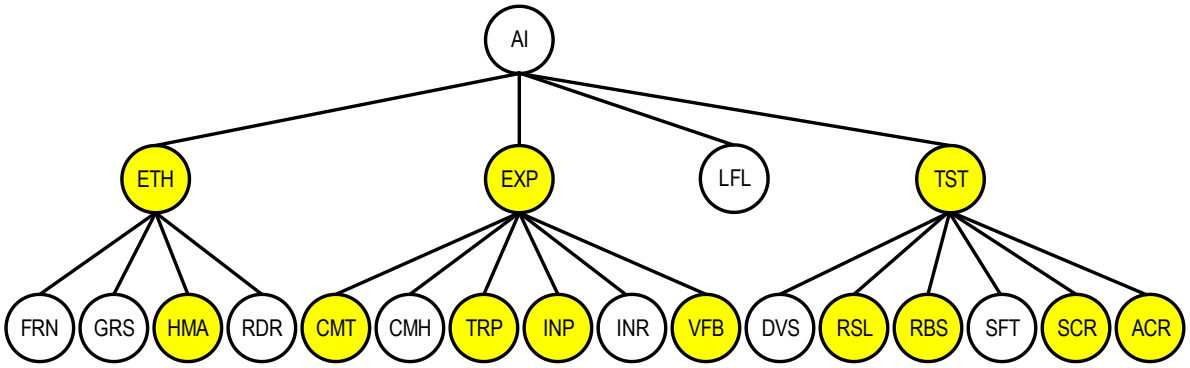

Рис. 4. Графова форма базової моделі якості системи розпізнавання дорожніх знаків

- довірчоздатність представлена чотирма підхарактеристиками з шести за виключенням диверсності DVS і безпечності SFT, враховуючи відсутність функцій формування керуючих впливів на людину.

Висновки

Не зважаючи на велику кількість публікацій та документів, виданих поважними національними та міжнародними інституціями, на даний час відсутня повна, упорядкована і несуперечлива сукупність характеристик ШІ, яку можна було б називати моделлю якості за аналогією з існуючими та загальноприйнятими моделями якості, розробленими для програмного забезпечення. Тому основним результатом дослідження вважаємо запропоновану модель якості штучного інтелекту, яка базується на аналізі та гармонізації визначень та залежностей характеристик якості, специфічних для ШІ.

Вибір характеристик та побудова моделі якості здійснювалось таким чином, щоб виключити повторення, забезпечити повноту представлення, а також визначити специфічні ознаки кожної з характеристик. Зрозуміло, що зробити модель, яка б повністю відповідала таким вимогам вкрай важко, тому представлені варіанти мають доповнюватися та удосконалюватися з урахуванням швидкого розвитку технологій і застосувань ШІ.

Основна та базова моделі якості надано в даному дослідженні у графовій формі, найбільш наочній та зручній для подальшого використання з метою оцінювання якості ШІ. Вони забезпечують можливість отримання часткових профілів якості з урахуванням специфіки відповідних систем, що було продемонстровано для двох СШІ. Вони далі можуть використовуватися як основа для метрично-базованого оцінювання якості таких систем.

Запропоновані моделі якості є відкритими і можуть доповнюватися і деталізуватися відповідно до специфіки призначення та сфери використання ШІ. На нашу думку, на базі запропонованих моделей можливе розроблення міжгалузевого стандарту якості та вимог до ШІ.

Подальші дослідження доцільно проводити за такими напрямами:

- профілювання (доповнення і деталізація) моделей для конкретних галузей, яке має супроводжуватися оглядом характеристик і підхарактеристик, що додаються на підставі досвіду розроблення та використання СШІ;
- розроблення метрик і алгоритмів для оцінювання ШІ за кожною з запропонованих характеристик та якості в цілому. Доцільно збирати та аналізувати інформацію про різні метрики задля їх включення до загальної бази даних;
- розроблення інструментальних засобів та кейс-орієнтованих методів оцінювання якості ШІ [39, 40]. Вони можуть базуватися на загальних Assurance Case підходах [41] і підходах, які стосуються оцінювання функційної і кібербезпеки [42]. Відбір інструментальних засобів, зокрема, для оцінювання кібербезпеки є окремою задачею, яка може виконуватися за допомогою засобів ШІ [43].

Робота підтримана проектом ECHO "European network of cybersecurity centres and competence hub for innovation and operations", який отримав фінансування від програми досліджень та інновацій Європейського Союзу Horizon 2020 в рамках грантової угоди № 830943. Автори вдячні колегам з консорціуму, співробітникам кафедри комп'ютерних систем, мереж і кібербезпеки Національного аерокосмічного університету ім. М. Є. Жуковського «Харківський авіаційний інститут» за участь в дискусіях, творчий аналіз результатів і цінні поради впродовж підготовки цієї статті.

Внесок авторів: огляд і аналіз літературних джерел – **Г. В. Фесенко, О. О. Ілляшенко**; формулювання принципів та послідовності досліджень – **В. С. Харченко, Г. В. Фесенко**; гармонізація визна-

чень характеристик ШІ – **В. С. Харченко, Г. В. Фесенко, О. О. Ілляшенко**; розроблення графових форм моделей якості ШІ – **В. С. Харченко, Г. В. Фесенко, О. О. Ілляшенко**; розроблення моделей якості для прикладів СШІ – **Г. В. Фесенко, О. О. Ілляшенко**. Усі автори прочитали та погодилися з опублікованою версією рукопису.

Reference (GOST 7.1:2006)

1. *A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks [Text] / M. R. Islam, M. U. Ahmed, S. Barua, S. Begum // Applied Sciences. – 2022. – Vol. 12. – Article Id: 1353. DOI: 10.3390/app12031353.*

2. *European Commission, High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI [Electronic resource]. – Available at: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf. – 10.03.2022.*

3. *European Commission, High-Level Expert Group on Artificial Intelligence. The Assessment List for Trustworthy Artificial Intelligence (ALTAI) [Electronic resource]. – Available at: https://airegio.emscarsa.com/nfs/programme_5/call_3/call_preparation/ALTAI_final.pdf. – 10.03.2022.*

4. *ISO/IEC TR 24372:2021. Information technology. Artificial intelligence. Overview of computational approaches for AI systems [Electronic resource]. – Available at: https://www.iso.org/standard/78508.html. – 10.03.2022.*

5. *ISO/IEC TR 24028:2020. Information technology. Artificial intelligence. Overview of trustworthiness in artificial intelligence [Electronic resource]. – Available at: https://www.iso.org/standard/77608.html. – 10.03.2022.*

6. *ISO/IEC TR 24027:2021. Information technology. Artificial intelligence. Bias in AI systems and AI aided decision making [Electronic resource]. – Available at: https://www.iso.org/standard/77607.html. – 10.03.2022.*

7. *IEEE 2941-2021. Standard for Artificial Intelligence (AI) Model Representation, Compression, Distribution, and Management [Electronic resource]. – Available at: https://ieeexplore.ieee.org/document/6922153. – 10.03.2022.*

8. *Four Principles of Explainable Artificial Intelligence: Draft NISTIR 8312 [Text] / P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, M. A. Przybocki, C. A. Hahn, P. C. Fontana. – Gaithersburg : National Institute of Standards and Technology, 2020. – 24 p. DOI: 10.6028/NIST.IR.8312.*

9. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence: NIST Special Publication 1270 [Text] / R. Schwartz, L. Down, A. Jonas, E. Tabassi. – Gaithersburg : National Institute of Standards and Technology, 2021. – 77 p. DOI: 10.6028/NIST.SP.1270.*

10. *Trust and Artificial Intelligence: Draft NISTIR 8332 [Text] / B. Stanton, T. Jensen. – Gaithersburg : National Institute of Standards and Technology. – 2021. – 23 p. DOI: 10.6028/NIST.IR.8332-draft.*

11. *OECD. Tools for Trustworthy AI: A Framework to Compare Implementation Tools [Electronic resource]. – Available at: https://www.oecd.org/science/tools-for-trustworthy-ai-008232ec-en.htm. – 10.03.2022.*

12. *UNESCO. Recommendation on the Ethics of Artificial Intelligence [Electronic resource]. – Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137. – 10.03.2022.*

13. *Christoforaki, M. AI Ethics—A Bird's Eye View [Text] / M. Christoforaki, O. Beyan // Applied Sciences. – 2022. – Vol. 12. – Article Id: 4130. DOI: 10.3390/app12094130.*

14. *Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges [Text] / F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu // Natural Language Processing and Chinese Computing: collective monograph ; edited by J. Tang, M. Y. Kan, D. Zhao, S. Li, H. Zan. – Berlin/Heidelberg: Springer International Publishing, 2019. – Vol. 11839. – P. 563-574. DOI: 10.1007/978-3-030-32236-6_51.*

15. *Trustworthy AI [Text] / R. Chatila, V. Dignum, M. Fisher, F. Giannotti, K. Morik, S. Russell, K. Yeung // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): collective monograph, edited by B. Braunschweig, M. Ghallab. – Cham : Springer International Publishing, 2021. – Vol. 12600. – P. 13-39. DOI: 10.1007/978-3-030-69128-8.*

16. *Gordieiev, O. IT-oriented software quality models and evolution of the prevailing characteristics [Text] / O. Gordieiev, V. Kharchenko // Dependable Systems, Services and Technologies (DESSERT): Proceeding of 9th Int. Conf., 2018. – P. 375-380. DOI: 10.1109/DESSERT.2018.8409162.*

17. *Gordieiev, O. Software quality standards and models evolution: greenness and reliability issues [Text] / O. Gordieiev, V. Kharchenko, M. Fusani // Information and communication technologies in education, research, and industrial applications: collective monograph, edited by V. Yakovyna, H. C. Mayr, M. Nikitchenko, G. Zholtkevych, A. Spivakovsky, S. Batsakis. – Berlin/Heidelberg : Springer International Publishing, 2016. – P. 38-55. DOI: 10.1007/978-3-319-30246-1_3.*

18. *Gerstlacher, J. Green and Sustainable Software in the Context of Software Quality Models [Text] / J. Gerstlacher, I. Groher, R. Plösch // HMD Praxis der*

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Wirtschaftsinformatik. – 2021. – Article Id: 554. DOI: 10.1365/s40702-021-00821-0.

19. Software Quality for AI: Where We Are Now? [Text] / V. Lenarduzzi, F. Lomio, S. Moreschini, D. Taibi, D. A. Tamburri // Lecture Notes in Business Information Processing : collective monograph, edited by D. Winkler, S. Biffl, D. Mendez, M. Wimmer, J. Bergsmann. – Cham: Springer International Publishing, 2021. – Vol. 404. – P. 43-53. DOI: 10.1007/978-3-030-65854-0_4.

20. Smith, A. L. Quality characteristics of artificially intelligent systems [Text] / A. L. Smith, R. Clifford // CEUR Workshop Proceedings (CEUR-WS). – 2020. – Vol. 2800. – P. 1-6.

21. ISO/IEC 25010:2011. Systems and software engineering. Systems and software Quality Requirements and Evaluation (SQuaRE). System and software quality models [Electronic resource]. – Available at: https://www.iso.org/standard/35733.html. – 10.03.2022.

22. Gordieiev, O. Software individual requirement quality model [Text] / O. Gordieiev // Radioelectronic and Computer Systems. – 2020. – No. 94. – P. 48-58. DOI: 10.32620/reks.2020.2.04.

23. The Industrial Internet of Things. Trustworthiness Framework Foundations. An Industrial Internet Consortium Foundational Document. Version V1.00 – 2021-07-15 [Electronic resource]. – Available at: https://www.iiconsortium.org/pdf/Trustworthiness_Framework_Foundations.pdf. – 10.03.2022.

24. Cambridge Dictionary. Acceptability. [Electronic resource]. – Available at: https://dictionary.cambridge.org/dictionary/english/acceptability. – 10.03.2022.

25. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI [Text] / A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera // Information Fusion. – 2020. – Vol. 58. – P. 82-115. DOI: 10.1016/j.inffus.2019.12.012.

26. Adadi, A. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) [Text] / A. Adadi, M. Berrada // IEEE Access. – 2018. – Vol. 6. – P. 52138-52160. DOI: 10.1109/ACCESS.2018.2870052.

27. Burciaga, A. Six Essential Elements of a Responsible AI Model [Electronic resource]. – Available at: https://www.forbes.com/sites/forbestechcouncil/2021/09/01/six-essential-elements-of-a-responsible-ai-model/?sh=21ebcb8456cf. – 10.03.2022.

28. Causability and Explainability of Artificial Intelligence in Medicine [Text] / A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller // WIREs Data Mining and Knowledge Discovery. – 2019. – Vol. 9. – P. 1-13. DOI: 10.1002/widm.1312.

29. Cambridge Dictionary. Comprehensibility [Electronic resource]. – Available at: https://dictionary.cambridge.org/dictionary/english/comprehensibility. – 10.03.2022.

30. From "Explainable AI" to "Graspable AI" [Text] / M. Ghajargar, J. Bardzell, A. S. Renner, P. G. Krogh, K. Höök, D. Cuartielles, L. Boer, M. Wiberg // Tangible, Embedded, and Embodied Interaction (TEI) : Proceeding of 15th Int. Conf., 2021. – P. 1-4. DOI: 10.1145/3430524.3442704.

31. From Responsibility to Reason-Giving Explainable Artificial Intelligence [Text] / K. Baum, S. Mantel, E. Schmidt, T. Speith // Philosophy & Technology. – 2022. – Vol. 35. – Article Id: 12. DOI: 10.1007/s13347-022-00510-w.

32. Explaining explanations: An overview of interpretability of machine learning [Text] / L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal // Data Science and Advanced Analytics (DSAA) : Proceeding of 5th Int. Conf., 2018. – P. 80-89. DOI: 10.1109/DSAA.2018.00018.

33. Wright, D. Understanding "Trustworthy" AI: NIST Proposes Model to Measure and Enhance User Trust in AI Systems [Electronic resource] / D. Wright. – Available at: https://www.jdsupra.com/legalnews/understanding-trustworthy-ai-nist-6387341. – 10.03.2022.

34. Traceability for Trustworthy AI: A Review of Models and Tools [Text] / M. Mora-Cantallops, S. Sánchez-Alonso, E. García-Barriocanal, M.-A. Sicilia // Big Data and Cognitive Computing. – 2021. – Vol. 5, Iss. 2. – Article Id: 20. DOI: 10.3390/bdcc5020020.

35. When Autonomous Systems Meet Accuracy and Transferability through AI: A Survey [Text] / C. Zhang, J. Wang, G.G. Yen, C. Zhao, Q. Sun, Y. Tang, F. Qian, J. Kurths // Patterns. – 2020. – Vol. 1, Iss. 4. – P. 1-28. DOI: 10.1016/j.patter.2020.100050.

36. Patil, K. R. Verifiability as a Complement to AI Explainability: A Conceptual Proposal [Preprint] [Electronic resource] / K. R. Patil, B. Heinrichs. – Available at: http://philsci-archive.pitt.edu/20297. – 10.03.2022.

37. Багатоетапний метод глибинного навчання з попереднім самонавчанням для класифікаційного аналізу дефектів стічних труб [Текст] / В. В. Москаленко, М. О. Зарецький, А. С. Москаленко, А. Г. Коробов, Я. Ю. Ковальський // Радіоелектронні і комп'ютерні системи. – 2021. – № 4. – С. 71-81. DOI: 10.32620/reks.2021.4.06.

38. Kuchuk, H. System of license plate recognition considering large camera shooting angles [Text] / H. Kuchuk, A. Podorozhniak, N. Liubchenko, D. Onischenko // Radioelectronic and Computer Sys-

tems. – 2021. – No. 4. – P. 82-91. DOI: 10.32620/reks.2021.4.07.

39. Felderer, M. Quality Assurance for AI-Based Systems: Overview and Challenges (Introduction to Interactive Session) [Text] / M. Felderer, R. Ramler // Lecture Notes in Business Information Processing : collective monograph ; edited by D. Winkler, S. Biffl, D. Mendez, M. Wimmer, J. Bergsmann. – Cham : Springer International Publishing, 2021. – Vol. 404. – P. 33-42. DOI: 10.1007/978-3-030-65854-0_3.

40. Bloomfield, R. Security-informed safety: If it's not secure, it's not safe [Text] / R. Bloomfield, K. Netkachova, R. Stroud // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): collective monograph, edited by S. Tonetta, E. Schoitsch, F. Bitsch. – Cham : Springer International Publishing, 2013. – Vol. 8166. – P. 17-32. DOI: 10.1007/978-3-642-40894-6_2.

41. Potii, O. Advanced security assurance case based on ISO/IEC 15408 [Text] / O. Potii, O. Illiashenko, D. Komin // Advances in Intelligent Systems and Computing : collective monograph ; edited by W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, J. Kacprzyk. – Cham : Springer International Publishing, 2015. – Vol. 365. – P. 391- 401. DOI: 10.1007/978-3-319-19216-1_37.

42. Conception and application of dependable Internet of Things based systems [Text] / O. O. Illiashenko, M. A. Kolisnyk, A. E. Strielkina, I. V. Kotsiuba, V. S. Kharchenko // Radio Electronics, Computer Science, Control. – 2020. – Vol. 4. – P. 139-150. DOI: 10.15588/1607-3274-2020-4-14.

43. Architecture and Model of Neural Network Based Service for Choice of the Penetration Testing Tools [Text] / A. G. Tetskyi, V. S. Kharchenko, D. D. Uzun, A. S. Nechausov // International Journal of Computing. – 2021. – Vol. 20(4). – P. 513-518. DOI: 10.47839/ijc.20.4.2438.

Reference (BSI)

1. Islam, M. R., Ahmed, M. U., Barua, S., Begum, S. A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences*, 2022, vol. 12, article id: 1353. DOI: 10.3390/app12031353.

2. *European Commission, High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI*. Available at: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf. (accessed 10.03.2022).

3. *European Commission, High-Level Expert Group on Artificial Intelligence. The Assessment List for Trustworthy Artificial Intelligence (ALTAI)*. Available at: https://airegio.ems-carsa.com/nfs/programme_5/call_3/call_preparation/ALTAI_final.pdf. (accessed 10.03.2022).

4. *ISO/IEC TR 24372:2021. Information technology. Artificial intelligence. Overview of computational approaches for AI systems*. Available at: https://www.iso.org/standard/78508.html. (accessed 10.03.2022).

5. *ISO/IEC TR 24028:2020. Information technology. Artificial intelligence. Overview of trustworthiness in artificial intelligence*. Available at: https://www.iso.org/standard/77608.html. (accessed 10.03.2022).

6. *ISO/IEC TR 24027:2021. Information technology. Artificial intelligence. Bias in AI systems and AI aided decision making*. Available at: https://www.iso.org/standard/77607.html. (accessed 10.03.2022).

7. *IEEE 2941-2021. Standard for Artificial Intelligence (AI) Model Representation, Compression, Distribution, and Management*. Available at: https://ieeexplore.ieee.org/document/6922153. (accessed 10.03.2022).

8. Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., Przybocki, M. A., Hahn, C. A., Fontana, P. C. *Four Principles of Explainable Artificial Intelligence: Draft NISTIR 8312*. Gaithersburg, National Institute of Standards and Technology Publ., 2020. 24 p. DOI: 10.6028/NIST.IR.8312.

9. Schwartz, R., Down, L., Jonas, A., Tabassi, E. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence: NIST Special Publication 1270*. Gaithersburg, National Institute of Standards and Technology, 2021. 77 p. DOI: 10.6028/NIST.SP.1270.

10. Stanton, B., Jensen, T. *Trust and Artificial Intelligence: Draft NISTIR 8332*. Gaithersburg, National Institute of Standards and Technology, 2022. 23 p. DOI: 10.6028/NIST.IR.8332-draft.

11. *OECD. Tools for Trustworthy AI: A Framework to Compare Implementation Tools*. Available at: https://www.oecd.org/science/tools-for-trustworthy-ai-008232ec-en.htm. (accessed 10.03.2022).

12. *UNESCO. Recommendation on the Ethics of Artificial Intelligence*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137. (accessed 10.03.2022).

13. Christoforaki, M., Beyan, O. AI Ethics—A Bird's Eye View. *Applied Sciences*, 2022, vol. 12, article id: 4130. DOI: 10.3390/app12094130.

14. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. *Natural Language Processing and Chinese Computing* : collective monograph ; edited by J. Tang, M. Y. Kan, D. Zhao, S. Li, H. Zan. Berlin/Heidelberg : Springer Inter-

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

national Publishing, 2019, vol. 11839, pp. 563-574. DOI: 10.1007/978-3-030-32236-6_51.

15. Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., Yeung, K. Trustworthy AI. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: collective monograph, edited by B. Braunschweig, M. Ghallab. Cham: Springer International Publishing, 2021, vol. 12600, pp. 13-39. DOI: 10.1007/978-3-030-69128-8.

16. Gordieiev, O. Kharchenko, V. IT-oriented software quality models and evolution of the prevailing characteristics. *Proc. of 9th Int. Conf. on Dependable Systems, Services and Technologies (DESSERT)*, 2018, pp. 375-380. DOI: 10.1109/DESSERT.2018.8409162.

17. Gordieiev, O., Kharchenko, V., Fusani, M. Software quality standards and models evolution: greenness and reliability issues. *Information and communication technologies in education, research, and industrial applications* : collective monograph ; edited by V. Yakovyna, H. C. Mayr, M. Nikitchenko, G. Zholtkevych, A. Spivakovsky, S. Batsakis. Berlin/Heidelberg: Springer International Publishing, 2016, pp. 38-55. DOI: 10.1007/978-3-319-30246-1_3.

18. Gerstlacher, J., Groher, I., Plösch, R. Green and Sustainable Software in the Context of Software Quality Models. *HMD Praxis der Wirtschaftsinformatik*, 2021, article id: 554. DOI: 10.1365/s40702-021-00821-0.

19. Lenarduzzi, V., Lomio, F., Moreschini, S., Taibi, D., Tamburri, D. A. Software Quality for AI: Where We Are Now? *Lecture Notes in Business Information Processing* : collective monograph ; edited by D. Winkler, S. Biffl, D. Mendez, M. Wimmer, J. Bergsmann. Cham: Springer International Publishing, 2021, vol. 404, pp. 43-53. DOI: 10.1007/978-3-030-65854-0_4.

20. Smith, A. L., Clifford, R. Quality characteristics of artificially intelligent systems. *CEUR Workshop Proceedings (CEUR-WS)*, 2020, vol. 2800, pp. 1-6.

21. *ISO/IEC 25010:2011. Systems and software engineering. Systems and software Quality Requirements and Evaluation (SQuaRE). System and software quality models*. Available at: https://www.iso.org/standard/35733.html. (accessed 10.03.2022).

22. Gordieiev, O. Software individual requirement quality model. *Radioelectronic and Computer Systems*, 2020, vol. 2, pp. 48-58. DOI: 10.32620/reks.2020.2.04.

23. *The Industrial Internet of Things. Trustworthiness Framework Foundations. An Industrial Internet Consortium Foundational Document. Version V1.00 – 2021-07-15*. Available at: https://www.iiconsortium.org/pdf/Trustworthiness_Framework_Foundations.pdf. (accessed 10.03.2022).

24. *Cambridge Dictionary. Acceptability*. Available at: https://dictionary.cambridge.org/dictionary/english/acceptability. (accessed 10.03.2022).

25. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 2020, vol. 58, pp. 82-115. DOI: 10.1016/j.inffus.2019.12.012.

26. Adadi, A., Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 2018, vol. 6, pp. 52138-52160. DOI: 10.1109/ACCESS.2018.2870052.

27. Burciaga, A. *Six Essential Elements of a Responsible AI Model*. Available at: https://www.forbes.com/sites/forbestechcouncil/2021/09/01/six-essential-elements-of-a-responsible-ai-model/?sh=21ebcb8456cf. (accessed 10.03.2022).

28. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H. Causability and Explainability of Artificial Intelligence in Medicine. *WIREs Data Mining and Knowledge Discovery*, 2019, vol. 9, pp. 1-13. DOI: 10.1002/widm.1312.

29. *Cambridge Dictionary. Comprehensibility*. Available at: https://dictionary.cambridge.org/dictionary/english/comprehensibility. (accessed 10.03.2022).

30. Ghajargar, M., Bardzell, J., Renner, A.S., Krogh, P.G., Höök, K., Cuartielles, D., Boer, L., Wiberg, M. From "Explainable AI" to "Graspable AI". *Proc. of 15th Int. Conf. on Tangible, Embedded, and Embodied Interaction (TEI)*, 2021, pp. 1-4. DOI: 10.1145/3430524.3442704.

31. Baum, K., Mantel, S., Schmidt, E., Speith, T. From Responsibility to Reason-Giving Explainable Artificial Intelligence. *Philosophy & Technology*, 2022, vol. 35, article id: 12. DOI: 10.1007/s13347-022-00510-w.

32. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., Kagal, L. Explaining explanations: An overview of interpretability of machine learning. *Proc. of 5th Int. Conf. on Data Science and Advanced Analytics (DSAA)*, 2018, pp. 80-89. DOI: 10.1109/DSAA.2018.00018.

33. Wright, D. *Understanding "Trustworthy" AI: NIST Proposes Model to Measure and Enhance User Trust in AI Systems*. Available at: https://www.jdsupra.com/legalnews/understanding-trustworthy-ai-nist-6387341. (accessed 10.03.2022).

34. Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., Sicilia, M.-A. Traceability for Trustworthy AI: A Review of Models and Tools. *Big Data and Cognitive Computing*, 2021, vol. 5, iss. 2, article id: 20. DOI: 10.3390/bdcc5020020.

35. Zhang, C., Wang, J., Yen, G.G., Zhao, C., Sun, Q., Tang, Y., Qian, F., Kurths, J. When Autonomous

Systems Meet Accuracy and Transferability through AI: A Survey. *Patterns*, 2020, vol. 1, iss. 4, pp. 1-28. DOI: 10.1016/j.patter.2020.100050.

36. Patil, K. R., Heinrichs, B. *Verifiability as a Complement to AI Explainability: A Conceptual Proposal [Preprint]*. Available at: *http://philsci-archive.pitt.edu/20297*. (accessed 10.03.2022).

37. Moskalenko, V., Zaretskyi, M., Moskalenko, A., Korobov, A., Kovalskyi, Y. Bahatoetapnyi metod hlybynnoho navchannia z poperednim samonavchanniam dlia klasyfikatsiinoho analizu defektiv stichnykh trub [Multi-stage deep learning method with self-supervised pretraining for sewer pipe defects classification]. *Radioelectronic and computer systems*, 2021, vol. 4, pp. 71-81. DOI: 10.32620/reks.2021.4.06.

38. Kuchuk, H., Podorozhniak, A., Liubchenko, N., Onischenko, D. System of license plate recognition considering large camera shooting angles. *Radioelectronic and Computer Systems*, 2021, vol. 4, pp. 82-91. DOI: 10.32620/reks.2021.4.07.

39. Felderer, M., Ramler, R. Quality Assurance for AI-Based Systems: Overview and Challenges. *Lecture Notes in Business Information Processing*: collective monograph, edited by D. Winkler, S. Biffl, D. Mendez, M. Wimmer, J. Bergsmann. Cham: Springer International Publishing, 2021, vol. 404, pp. 33-42. DOI: 10.1007/978-3-030-65854-0_3.

40. Bloomfield, R., Netkachova, K., Stroud, R. Security-informed safety: If it's not secure, it's not safe. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: collective monograph, edited by S. Tonetta, E. Schoitsch, F. Bitsch. Cham, Springer International Publishing, 2013, vol. 8166, pp. 17-32. DOI: 10.1007/978-3-642-40894-6_2.

41. Potii, O., Illiashenko, O., Komin. D. Advanced security assurance case based on ISO/IEC 15408. *Advances in Intelligent Systems and Computing*: collective monograph, edited by W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, J. Kacprzyk. Cham, Springer International Publishing, 2015, vol. 365, pp. 391-401. DOI: 10.1007/978-3-319-19216-1_37.

42. Illiashenko, O. O., Kolisnyk, M. A., Strielkina, A. E., Kotsiuba, I. V., Kharchenko, V.S. Conception and application of dependable Internet of Things based systems. *Radio Electronics, Computer Science, Control*, 2020, vol. 4, pp. 139-150. DOI: 10.15588/1607-3274-2020-4-14.

43. Tetskyi, A. G., Kharchenko, V. S., Uzun, D. D., Nechausov, A. S. Architecture and Model of Neural Network Based Service for Choice of the Penetration Testing Tools. *International Journal of Computing*. 2021, vol. 20(4), pp. 513-518. DOI: 10.47839/ijc.20.4.2438.

Надійшла до редакції 16.01.2022, розглянута на редколегії 15.04.2022

BASIC MODEL OF NON-FUNCTIONAL CHARACTERISTICS FOR ASSESSMENT OF ARTIFICIAL INTELLIGENCE QUALITY

Vyacheslav Kharchenko, Herman Fesenko, Oleg Illiashenko

The **subject** of the research is the models of artificial intelligence (AI) quality. The current paper develops an AI quality model based on the definition and ordering of its characteristics. **Objectives**: to develop the principles and justify the sequence of analysis and development of AI quality models as ordered sets of characteristics; to offer models of AI quality for further use, first, the evaluation of individual characteristics and quality in general; to demonstrate the profiling of AI quality models for systems using artificial intelligence. The following **results** were obtained. The sequence of construction of AI quality models is offered. Based on the analysis of references, a list of AI characteristics was formed and their definitions were harmonized. The general model of AI quality is presented with a description of the step-by-step procedure for the realization of its hierarchical construction. A basic model of AI with abbreviated sets of characteristics is proposed due to its importance. Examples of profiling of quality models for two systems - monitoring of engineering communications and recognition of road signs are given. **Conclusions**. The study's main result is the development of a quality model for artificial intelligence, which is based on the analysis and harmonization of definitions and dependencies of quality characteristics specific to AI. The selection of characteristics and the construction of the quality model were carried out in such a way to exclude duplication, ensure the completeness of the presentation, as well as to determine the specific features of each characteristic. It is extremely difficult to create a model that would fully meet such requirements, so the presented options should be supplemented and improved considering the rapid development of technologies and applications of AI. The proposed quality models are open and can be supplemented and detailed according to the specific purpose and scope of AI.

Keywords: artificial intelligence; characteristics of artificial intelligence; artificial intelligence quality model; the profiling of artificial intelligence quality models.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

БАЗОВАЯ МОДЕЛЬ НЕФУНКЦИОНАЛЬНЫХ ХАРАКТЕРИСТИК ДЛЯ ОЦЕНКИ КАЧЕСТВА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

В. С. Харченко, Г. В. Фесенко, О. А. Ильяшенко

Предметом исследования являются модели качества искусственного интеллекта (ИИ). **Целью** статьи является разработка модели качества ИИ на основе определения и упорядочивания характеристик. **Задачи**: сформулировать принципы и обосновать последовательность анализа и разработки моделей качества ИИ как упорядоченных множеств характеристик; предложить модели качества ИИ для дальнейшего использования, прежде всего, оценки отдельных характеристик и качества в целом; продемонстрировать профилирование моделей качества ИИ для систем, где используется искусственный интеллект. Были получены следующие **результаты**. Предложена последовательность построения моделей качества ИИ. На основании анализа литературных источников сформирован список характеристик ИИ и осуществлена гармонизация их определений. Представлена общая модель качества ИИ с представлением описания пошаговой процедуры реализации ее иерархического построения. Предложена базовая модель ИИ с сокращенными множествами характеристик, учитывая их важность. Представлены примеры профилирования моделей качества для двух систем – мониторинга инженерных коммуникаций и распознавания дорожных знаков. **Выводы**. Основным результатом проведенного исследования является разработка модели качества для искусственного интеллекта, которая базируется на анализе и гармонизации определений и зависимостей характеристик качества, которые специфичны для ИИ. Выбор характеристик и построение модели качества производилось таким образом, чтобы исключить повторение, обеспечить полноту представления, а также определить специфические признаки каждой из характеристик. Понятно, что сделать модель, которая полностью отвечала бы таким требованиям крайне трудно, поэтому представленные варианты должны дополняться и совершенствоваться с учетом быстрого развития технологий и применений ИИ. Предлагаемые модели качества открыты и могут дополняться и детализироваться в соответствии со спецификой назначения и сферой использования ИИ.

Ключевые слова: искусственный интеллект; характеристики искусственного интеллекта; модель качества искусственного интеллекта; профилирование моделей качества искусственного интеллекта.

Харченко Вячеслав Сергійович – д-р техн. наук, проф., зав. каф. комп'ютерних систем, мереж і кібербезпеки, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Фесенко Герман Вікторович – д-р техн. наук, доц., доц. каф. комп'ютерних систем, мереж і кібербезпеки, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Ілляшенко Олег Олександрович – канд. техн. наук, доц., доц. каф. комп'ютерних систем, мереж і кібербезпеки, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Vyacheslav Kharchenko – Doctor of Technical Science, Professor, Head of the Computer Systems, Networks and Cybersecurity Department, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: v.kharchenko@csn.khai.edu, ORCID: 0000-0001-5352-077X.

Herman Fesenko – Doctor of Technical Science, Associate Professor, Associate Professor of the Computer Systems, Networks and Cybersecurity Department, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: h.fesenko@csn.khai.edu, ORCID: 0000-0002-4084-2101.

Oleg Illiashenko – PhD, Associate Professor, Associate Professor of the Computer Systems, Networks and Cybersecurity Department, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: o.illiashenko@khai.edu, ORCID: 0000-0002-4672-6400.