УДК 004.8:004.9 **doi: 10.32620/aktt.2026.1.12**

С. Ф. ЧАЛИЙ, І. О. ЛЕЩИНСЬКА

Харківський національний університет радіоелектроніки

ІНТЕГРОВАНА НЕЙРОСИМВОЛЬНА АРХІТЕКТУРА МЕНТАЛЬНИХ МОДЕЛЕЙ КОРИСТУВАЧІВ ДЛЯ ПЕРСОНАЛІЗОВАНИХ ПОЯСНЕНЬ РІШЕНЬ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ

Предметом *дослідження є методологія побудови інтегрованої нейросимвольної архітектури ментальних моделей користувачів з різним рівнем технічної компетентності для генерації персоналізованих пояснень рішень інтелектуальних інформаційних систем.* ***Метою*** *роботи є розробка архітектури, яка забезпечує автоматизоване виявлення індивідуальних ментальних моделей з поведінкових даних користувачів та створення зрозумілих символьних представлень каузальних взаємозв'язків з адаптацією деталізації до рівня підготовки користувача.* ***Завдання****: виконання порівняльного аналізу існуючих підходів до побудови ментальних моделей за критеріями персоналізації та розробка інтегрованої нейросимвольної архітектури з розподілом функціональності між нейромережевим та символьним компонентами; експериментальна перевірка запропонованої архітектури; визначення області застосування розробленої архітектури. Застосовані* ***методи*** *включають варіаційні автокодувачі з багатоканальними механізмами уваги, нейросимвольну трансляцію з багаторівневою абстракцією, генерацію орієнтованих ациклічних графів. Були отримані наступні* ***результати****. Розроблено інтегровану нейросимвольну архітектуру з нейромережевим компонентом для автоматизованого виявлення індивідуальних когнітивних структур через варіаційне кодування та механізми уваги з динамічною пріоритизацією каналів за категорією користувачів та символьним компонентом для перетворення латентних дескрипторів у інтерпретовані каузальні графи з адаптивною деталізацією й темпоральною валідацією для видалення помилкових залежностей.* ***Висновки.*** *Результати дослідження підтвердили ефективність інтегрованого нейросимвольного підходу до побудови персоналізованих ментальних моделей з автоматизованим виявленням латентних когнітивних структур з поведінкових даних без залучення експертних знань.* ***Наукова новизна*** *отриманих результатів полягає у розробці моделі інтегрованої нейросимвольної архітектури, що передбачає взаємодію нейромережного шару, призначеного для проєктування поведінкових траєкторій у латентний простір й селективного відбору значущих ознак через багатоканальну увагу, та символьного шару для нейросимвольної трансляції латентних векторів у багаторівневе символьне представлення з генерацією орієнтованих ациклічних графів та темпоральною валідацією, що дає можливість підвищити зрозумілість рішень інтелектуальних систем через побудову персоналізованих інтерпретованих ментальних моделей, а також підвищити довіру користувачів до рішень інтелектуальних систем за рахунок деталізації пояснень згідно рівня їхньої технічної компетентності.*

Ключові слова: *нейросимвольна архітектура; ментальні моделі; персоналізовані пояснення; пояснювальний штучний інтелект; варіаційні автокодувачі; механізми уваги; каузальні графи; темпоральна валідація; диференціація користувачів; інтерпретованість.*

1. Вступ

Інтелектуальні інформаційні системи (ІІС) поєднують переваги традиційних інформаційних систем та систем штучного інтелекту при вирішенні комплексних задач підтримки прийняття рішень [1]. Однак використання моделей машинного навчання в таких системах утруднює розуміння логіки роботи ІІС через непрозорість внутрішніх механізмів прийняття рішень та знижує довіру до цих рішень [2]. Для забезпечення прозорості роботи ІІС використовуються методи пояснювального штучного інтелекту (XAI).

Перехід до нової парадигми XAI 2.0 характеризується зміщенням акценту від пояснень, що генеруються після прийняття рішення системою, представленою як «чорна скринька», до включення можливості побудови пояснень у структуру моделі. Ключовою вимогою XAI 2.0 є персоналізація пояснень, яка полягає в адаптації змісту, деталізації та форми пояснення до рівня знань, інформаційних потреб та когнітивних можливостей конкретного користувача [3].Проте існуючі підходи до XAI використовують переважно узагальнені пояснення, що не враховують персональні особливості сприйняття користувачів. Останні визначаються індивідуальни-

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

ми ментальними моделями, які представляють собою внутрішні когнітивні структури, що відображають розуміння користувачем логіки роботи ІІС, каузальних залежностей у предметній області та очікувань щодо поведінки системи при вирішенні задач користувача. Різні користувачі можуть мати суттєво відмінні ментальні моделі для однієї й тої ж інтелектуальної системи залежно від їхнього рівня технічної компетентності, професійного досвіду та когнітивних структур. Використання ментальних моделей як основи для персоналізації пояснень дає можливість адаптувати пояснення відповідно до рівня знань користувача, його очікувань щодо можливостей застосування рішення та когнітивних обмежень [4].

Однак побудова персоналізованих ментальних моделей користувачів у системах ХАІ стикається з двома ключовими проблемами. По-перше, ментальні моделі є неявними когнітивними структурами, які не можуть бути безпосередньо зафіксовані. Ці моделі необхідно виявляти на основі аналізу поведінкових даних користувачів, історії взаємодії з ІІС та патернів прийняття рішень [5]. По-друге, виявлені ментальні моделі мають бути пояснюваними та інтерпретованими, тобто вони мають бути представлені у вигляді символьних структур (наприклад, графів, правил, онтологій), що дозволяють користувачу та розробнику системи зрозуміти логіку міркування ІІС [6].

Ці проблеми обумовлюють необхідність розробки нейросимвольних підходів до побудови ментальних моделей, які інтегрують переваги нейромережевих методів (автоматизоване виявлення латентних патернів з поведінкових даних, адаптивність, здатність працювати з неструктурованими даними) та символьних методів (пояснюваність, верифікованість, можливість логічного виводу та інтерпретації каузальних залежностей). Однак існуючі нейросимвольні архітектури, що застосовуються в ХАІ, не враховують диференціацію користувачів за рівнем технічної компетентності, вони генерують типові ментальні моделі [7].

1.1. Мотивація дослідження

Розвиток систем ХАІ стримується відсутністю архітектурних рішень для побудови персоналізованих ментальних моделей користувачів, що забезпечують узгодження між автоматизованим виявленням індивідуальних когнітивних структур та зрозумілими каузальними залежностями для категорій користувачів з різним рівнем технічної компетентності [3]. Нейромережні архітектури, що використовують варіаційні автокодувачі та механізми уваги для автоматизованого виявлення латентних ознак з поведінкових даних користувачів, демонструють високу якість розпізнавання індивідуальних когнітивних патернів [8]. Однак такі архітектури представляють ментальні моделі як латентні вектори у багатовимірному просторі, що залишаються непрозорими для користувачів. Відсутність символьних структур унеможливлює інтерпретацію каузальних залежностей між концептами, що знижує зрозумілість пояснень та робить неможливою верифікацію коректності ментальних моделей користувачами й розробниками ІІС. Крім того, нейромережні архітектури не дають можливість адаптувати складність ментальних моделей відповідно до рівня підготовки користувача, оскільки латентний простір не містить явної структури для відображення рівня деталізації представлень відповідно до когнітивних можливостей та інформаційних потреб користувачів з різним рівнем технічної компетентності. Символьні архітектури, що базуються на експертних правилах та онтологіях предметної області і забезпечують можливість побудови каузальних графів з логічним виводом, забезпечують високу інтерпретованість [9]. Проте символьні архітектури вимагають трудомісткого ручного конструювання правил експертами для кожної предметної області та категорії користувачів, що потребує значних часових ресурсів на розробку онтологічних структур. Тому такий підхід унеможливлює автоматизоване виявлення індивідуальних ментальних моделей безпосередньо з вхідних поведінкових даних. Також символьні архітектури генерують однакові ментальні моделі для всіх користувачів на основі фіксованих правил, ігноруючи індивідуальні відмінності у розумінні логіки функціонування інтелектуальної системи.

Аналіз існуючих підходів до побудови ментальних моделей показує, що вони не орієнтовані на побудову персоналізованих ментальних моделей користувачів з неоднаковим рівнем технічної компетентності. Нейромережеві архітектури забезпечують персоналізацію. Однак такі моделі генерують непрозорі латентні представлення без можливості інтерпретації каузальних залежностей користувачами. Символьні архітектури забезпечують інтерпретованість, але вимагають трудомісткого ручного конструювання правил та не орієнтовні на персоналізацію. Гібридні нейросимвольні архітектури об’єднують персоналізацію та інтерпретованість, однак не враховують диференціацію користувачів за рівнем компетентності для адаптації складності пояснень [10].

Розробка інтегрованої нейросимвольної архітектури, що поєднує переваги нейромережевих методів (варіаційні автокодувачі, багатоголові механізми уваги) та символьні методи (нейросимвольне перетворення з багаторівневою абстракцією, орієнтовані

ациклічні графи, лінійна темпоральна логіка), створює умови для виявлення індивідуальних ментальних моделей з поведінкових даних без необхідності явного формулювання правил експертами [11].

1.2. Аналіз останніх досліджень й публікацій

Підходи до побудови нейросимвольних архітектур для персоналізованих пояснень включають нейромережеві архітектури з глибоким навчанням для виявлення латентних когнітивних структур, символьні архітектури на основі онтологій та логічного виводу для побудови інтерпретованих каузальних графів, гібридні нейросимвольні архітектури з механізмами інтеграції нейромережевих та символьних компонентів, та архітектури з механізмами уваги для селективного відбору значущих ознак з поведінкових даних.

У роботі [1] запропоновано архітектуру на основі варіаційних автокодувачів для виявлення прихованих когнітивних патернів з послідовностей взаємодії користувачів з ІІС. Проте не розглядається символьне представлення для забезпечення пояснюваності ментальних моделей. Дослідження [2] показало, що нейросимвольні архітектури забезпечують значно вищу інтерпретованість порівняно з нейромережними підходами, проте необхідною умовою пояснюваності є темпоральна узгодженість каузальних залежностей. Робота [3] присвячена дослідженню ефекту персоналізації пояснень у системах XAI. Показано, що адаптація рівня абстракції концептів до професійного досвіду користувачів суттєво підвищує задоволеність поясненнями порівняно з універсальними поясненнями для всіх категорій користувачів [4].

Динамічна адаптація механізмів уваги для персоналізації нейросимвольних систем представлена в роботі [5]. Показано, що адаптивна кількість голів залежно від складності завдання забезпечує кращий баланс між точністю та обчислювальною ефективністю.

В дослідженні [6] описано механізм верифікації темпоральних обмежень для каузальних графів у системах пояснення рішень. Верифікація на основі темпоральної логіки суттєво знижує частку хибних залежностей порівняно з простою статистичною кореляцією.

Механізми диференціації користувачів для адаптивних ІІС використовують класифікацію за множинними критеріями компетентності для забезпечення гнучкості персоналізації [7].

Таким чином, існуючі підходи до проєктування нейросимвольних архітектур мають обмеження для побудови персоналізованих ментальних моделей користувачів. Ці обмеження пов'язані із відсутністю підходів до диференціації користувачів за рівнем технічної компетентності з критеріями адаптації складності символьних представлень, а також адаптивної темпоральної верифікації.

1.3. Мета та завдання дослідження

Метою роботи є розробка інтегрованої нейросимвольної архітектури для побудови персоналізованих ментальних моделей користувачів з різним рівнем технічних знань, що поєднує нейромережевий шар латентного представлення індивідуальних когнітивних структур та символьний шар, який містить відображення латентного представлення на направлений ациклічний граф, зрозумілий для користувача.

Для досягнення мети необхідно вирішити такі завдання:

1. Провести порівняльний аналіз існуючих архітектурних підходів до побудови ментальних моделей користувачів і розробити модель інтегрованої нейросимвольної архітектури з розподілом функціональності між нейромережевим компонентом для виявлення латентних когнітивних структур та символьним компонентом для генерації інтерпретованих каузальних графів.

2. Виконати експериментальну перевірку запропонованої архітектури на даних взаємодії фахівців з ІІС.

3. Визначити сферу застосування розробленої інтегрованої архітектури ментальної моделі.

2. Інтегрована нейросимвольна архітектура ментальних моделей користувачів

Розроблена інтегрована нейросимвольна архітектура ментальних моделей користувачів A для створення індивідуалізованих пояснень складається з нейромережевого компонента (шару) з механізмами виявлення прихованих патернів для адаптивної персоналізації, та символьного компонента з генераторами каузальних графів та модулями темпоральної перевірки. Формально архітектуру можна представити як:

$$A = (N, S): B \to M, \qquad (1)$$

де N – нейромережний шар;

S – символьний шар;

$B = \{b_1, b_2, \ldots, b_n\}$ – множина поведінкових послідовностей користувачів;

$M = \{m_1, m_2, \ldots, m_k\}$ – множина персоналізованих ментальних моделей.

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

Нейромережний компонент реалізує відображення:

$$N: B \to Z, \quad (2)$$

де Z – латентний простір розмірності d.

Символьний компонент реалізує відображення

$$S: Z \to G, \quad (3)$$

де G – каузальний граф.

Нейромережний компонент відповідає за автоматизоване розпізнавання неявних когнітивних структур з поведінкових траєкторій користувачів:

– проєктування послідовностей взаємодій користувача з ІІС у латентний простір через варіаційне кодування з регуляризацією для гладкої інтерполяції між індивідуальними ментальними моделями;

– відбір значущих латентних характеристик через механізми фокусування уваги з динамічною пріоритизацією каналів за категорією користувача;

– усунення випадкового поведінкового шуму з фінальних моделей;

– генерація компактних латентних дескрипторів для швидкої обробки.

Використання механізму уваги дає можливість виконати персоналізацію без явного формулювання правил фахівцями, що дає можливість використати модель для різних категорій користувачів.

Символьний компонент активується після нейромережного для перетворення латентних дескрипторів у інтерпретовані конструкції. Він має наступну функціональність, яка забезпечує прозорість ментальних моделей для кінцевих користувачів:

– перетворення латентних векторів у символьні сутності через нейросимвольну трансляцію з багаторівневою абстракцією відповідно до технічної компетентності користувача;

–генерація орієнтованих ациклічних графів для візуалізації каузальних взаємозв’язків між сутностями з контролем деталізації;

– перевірка відсутності темпоральних протиріч у каузальних залежностях з використанням лінійної темпоральної логіку з тим, щоб видалити помилкові залежності;

– динамічне налаштування деталізації графів відповідно до вимог щодо представлення пояснень для різних категорій користувачів.

Символьний компонент забезпечує явне представлення каузальних взаємозв’язків, а також створює умови для перевірки ментальних моделей безпосередньо користувачами й розробниками.

Властивості складових запропонованої архітектури наведено в табл. 1.

Взаємодія компонентів запропонованої інтегрованої нейросимвольної архітектури реалізується через послідовний потік обробки поведінкових даних: нейромережний шар виявляє латентні когнітивні структури з поведінкових послідовностей, нейросимвольний транслятор перетворює латентні представлення у символьні концепти, символьний шар будує в каузальні графи з адаптацією до категорії користувача.

Послідовний процес побудови ментальної моделі забезпечує перетворення вхідних поведінкових даних у персоналізовані інтерпретовані каузальні структури. Модуль збору поведінкових даних виконує накопичення послідовностей взаємодій користувача з інтелектуальною системою, фіксуючи типи дій, параметри функцій та темпоральні інтервали між подіями з формуванням повних профілів взаємодії. Обробник темпоральних послідовностей формує ковзні вікна спостережень, обчислює статистичні характеристики поведінкових патернів та нормалізує дані для забезпечення можливості порівняння для різних користувачів. Підготовлені послідовності поведінки користувача передаються енкодеру варіаційного автокодувача для відображення у латентний простір через обчислення параметрів апостеріорного розподілу латентних змінних. Латентні вектори проходять через багатоголовий механізм уваги для відбору значущих ознак з фільтрацією поведінкового шуму через адаптивні ваги голів в залежності від категорії користувача. Виділені значущі латентні ознаки передаються нейросимвольному транслятору для відображення у символьні концепти з обранням рівня абстракції концептів відповідно до рівня технічної компетентності користувача. Генератор каузальних графів обчислює ваги потенційних дуг між парами концептів на основі умовних ймовірностей з поведінкових послідовностей та застосовує адаптивні порогові значення фільтрації для контролю складності графа. Модуль темпоральної верифікації перевіряє узгодженість каузальних дуг з темпоральними обмеженнями з використанням лінійної темпоральної логіки, відфільтровуючи помилкові залежності. Модуль використовує адаптивні критерії верифікації в залежності від категорії користувачів. Механізм адаптивної деталізації виконує фінальне налаштування кількості концептів та дуг у графі через агрегацію або деталізацію відповідно до інформаційних потреб користувача. Результуюча персоналізована ментальна модель у вигляді каузального графа з символьними концептами може бути використана у ІІС для генерації пояснень рішень користувачу. Послідовна взаємодія між нейромережним та символьним шарами не містить зворотних зв'язків. Однак при накопиченні нових поведінкових даних виконується оновлення моделі шляхом періодичного перенавчання варіаційного

Таблиця 1

Структура інтегрованої нейросимвольної архітектури ментальних моделей

Модуль	Функціональність модуля	Механізми адаптації
Нейромережний шар		
Варіаційний кодувач	Проєктування траєкторій поведінки користувача у латентний простір з регуляризацією для інтерполяції між індивідуальними моделями	Параметри регуляризації адаптуються за варіативністю когнітивних стратегій кожної категорії користувачів
Механізм багатоканальної уваги	Селективний відбір значущих латентних характеристик з усуненням поведінкового шуму через адаптивні ваги каналів	Пріоритизація каналів: темпоральні ознаки для фахівців, категоріальні для новачків
Збирач поведінкових траєкторій	Безперервний збір послідовностей взаємодії користувача з інтелектуальною системою через реєстрацію подій	Частота відбору подій може бути змінена в залежності від швидкості взаємодії для категорії користувачів
Обробник темпоральних послідовностей	Конструювання ковзних вікон спостережень та обчислення статистичних дескрипторів поведінкових патернів	Ширина вікна залежить від типової тривалості когнітивних епізодів категорії
Символьний шар		
Нейросимвольний транслятор	Відображення латентних векторів у багаторівневі символьні сутності за категорією користувача	Рівень абстракції задається в залежності від категорій користувачів: деталізація для фахівців, узагальнення для новачків
Генератор каузальних графів	Генерація орієнтованих ациклічних графів (DAG) для відображення каузальних взаємозв’язків з контролем деталізації	Порогова фільтрація дуг: високі пороги задаються для новачків, а низькі для фахівців
Модуль темпоральної валідації	Перевірка відсутності темпоральних протиріч для каузальних взаємозв’язків з використанням лінійної темпоральної логіки (LTL)	Адаптивні критерії несуперечності, які відрізняються для фахівців та новачків
Механізм динамічної деталізації	Автоматичне налаштування кількості сутностей та дуг у графах згідно вимог категорії користувачів	Кількість сутностей та дуг задається максимальним для фахівців, мінімальним для новачків

автокодувача та нейросимвольного транслятора. Час обробки залежить від довжини поведінкових послідовностей.

3. Експериментальна перевірка

Експериментальну перевірку розробленої інтегрованої нейросимвольної архітектури ментальних моделей виконано на даних із датасету Medical DSS Dataset, який містить 1,247 послідовностей взаємодій 156 медичних фахівців з діагностичною системою підтримки для виявлення серцево-судинних захворювань.

Датасет містить три категорії користувачів: 43 досвідчені лікарі-діагности, 68 лікарів загальної практики та 45 інтернів. Імплементація архітектури реалізована на PyTorch 2.1.0. Варіаційний кодувач побудовано з енкодером на основі двошарової LSTM, латентний простір має розмірність 128.

Для оцінки використано коефіцієнт силуету, індекс Девіса-Болдіна, точність класифікації та F1-міру. Коефіцієнт силуету відображає якість кластеризації через відношення міжкластерної та внутрішньокластерної відстані. Індекс Девіса-Болдіна оцінює якість кластеризації через відношення внутрішньокластерного розсіювання до міжкластерної відстані. Індекс темпоральної несуперечності (Temporal Consistency Index, TCI) вимірює частку каузальних зв'язків, що не порушують темпоральну логіку послідовностей дій.

Результати експериментальної перевірки представлено в табл. 2.

4. Обговорення результатів та перспективи розвитку

Експериментальні результати демонструють ефективність розробленої інтегрованої нейросимвольної архітектури для побудови персоналізованих пояснень в інтелектуальних системах.

Досягнутий коефіцієнт силуету 0,68 підтверджує спроможність нейромережного компонента автоматизовано виявляти та диференціювати корис-

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

Таблиця 2

Результати експериментальної перевірки розробленої моделі

Метод	Коефіцієнт силуету	Індекс Девіса-Болдіна	Accuracy (%)	F1-score	TCI
Запропонована архітектура	0,68	0,58	89,3	0,88	0,82
Виключно нейромережева	0,64	0,63	85,1	0,84	0,78
Символьна система	0,42	0,89	73,2	0,71	0,79
Гібридна без диференціації	0,58	0,71	81,7	0,80	0,80
Без механізму уваги	0,61	0,66	83,4	0,82	0,81
Без темпоральної валідації	0,66	0,6004	87,2	0,86	0,58

тувачів з різним рівнем підготовки без залучення експертних знань. Дійсно, значення коефіцієнту в діапазоні 0,51–0,70 свідчать про хорошу структуризацію латентного простору. Індекс Девіса-Болдіна становить 0,58, що свідчить про достатньо хорошу кластеризацію. Відповідно, розроблена архітектура дає можливість виявити і пояснити три патерни поведінки користувачів.

Варіаційний автокодувач розміщує схожих користувачів поруч у латентному просторі, формуючи компактні кластери. Відповідно, модель точно визначає категорію користувача у 89,3% випадків. Запропонована архітектура має обмеження, пов'язані із об'ємом вхідної вибірки та вимогами до онтологічного опису. Навчання варіаційного автокодувача вимагає достатньо великої кількості поведінкових послідовностей для кожної категорії користувачів (орієнтовно від 30 послідовностей) для формування компактних кластерів у латентному просторі. Нейросимвольна трансляція базується на фіксованій онтології предметної області, що обмежує можливості каузальних графів. У разі появи нових симптомів, діагностичних процедур або захворювань потрібно ручне доповнення онтології та повторне навчання нейросимвольного транслятора.

Отримані експериментальні результати свідчать про можливість застосування моделі у ризикових областях, де пояснення мають критичну роль для здоров'я та добробуту людей – авіаційній галузі, медицині, фінансовій сфері.

5. Висновки

Виконане дослідження дозволило отримати наступні основні результати:

1. Розроблено інтегровану нейросимвольну архітектуру з розподілом функціональності між нейромережевим компонентом для автоматизованого виявлення індивідуальних когнітивних структур з траєкторій взаємодії користувачів з інтелектуальною системою та символьним компонентом, що містить інтерпретований каузальний граф з відсутністю темпоральних протиріч.

2. Виконано експериментальну перевірку на основі даних про послідовності взаємодії користувачів з інтелектуальною системою. Запропонована архітектура досягла високих значень коефіцієнта силуету, низьких значень індексу Девіса-Болдіна, високої точності класифікації категорій користувачів та високих значень F1-міри.

4. Визначено сферу застосування для критичних предметних областей з високими вимогами до пояснюваності рішень та необхідністю персоналізації пояснень відповідно до рівня професійної кваліфікації користувачів. Важливим напрямком застосування є системи авіаційних тренажерів для підготовки льотного складу, де адаптація складності пояснень до кваліфікації курсантів та пілотів безпосередньо впливає на ефективність навчання. Також архітектура може бути використана в навчальних медичних системах та фінансових платформах.

Внесок авторів: концепції та методологія – **С. Ф. Чалий;** формулювання цілей та задач дослідження – **С. Ф. Чалий;** проведення дослідження поточного стану – **І. О. Лещиньска;** розробка інтегрованої нейросимвольної архітектури – **І. О. Лещинська,** проведення експерименту – **І. О. Лещинська,** інтерпретація результатів – **С. Ф. Чалий, І. О. Лещинська.**

Конфлікт інтересів

Автори заявляють, що немає конфлікту інтересів щодо цього дослідження, фінансового, особистого, авторського чи іншого, який міг би вплинути на дослідження та його результати, представлені в цій статті.

Фінансування

Дослідження проводилося без фінансової підтримки.

Доступність даних

Рукопис не містить пов'язаних даних.

Використання штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту під час створення цієї роботи.

Автори прочитали та погодилися з опублікованою версією рукопису.

Література

1. Kingma, D. P. Auto-Encoding Variational Bayes [Text] / D. P. Kingma, & M. Welling // Advances in Neural Information Processing Systems (NeurIPS 2013). – 2013. – Vol. 27. – Available at: https://arxiv.org/abs/1312.6114 (accessed: 12.12.2025).

2. Logic Tensor Networks [Text] / S. Badreddine, A. d'Avila Garcez, L. Serafini, & M. Spranger // Artificial Intelligence. – 2022. – Vol. 303. – Article no. 103649. DOI: 10.1016/j.artint.2021.103649.

3. Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences [Text] / T. Miller // Artificial Intelligence. – 2019. – Vol. 267. – P. 1–38. DOI: 10.1016/j.artint.2018.07.007.

4. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI [Text] / M. Chromik, M. Eiband, F. Buchner, A. Krüger, & A. Butz // Proceedings of the 26th International Conference on Intelligent User Interfaces. – New York : ACM, 2021. – P. 307–317. DOI: 10.1145/3397481.3450644.

5. User Characteristics in Explainable AI: The Rabbit Hole of Personalization? [Text] / R. Nimmo, M. Constantinides, K. Zhou, D. Quercia, & S. Stumpf // Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI 2024). – New York : ACM, 2024. DOI: 10.1145/3613904.3642352.

6. Pearl, J. Causality: Models, Reasoning, and Inference [Text] / J. Pearl. – 2nd ed. – Cambridge: Cambridge University Press, 2009. – 484 p. – ISBN: 978-0521895606.

7. Causability and Explainability of Artificial Intelligence in Medicine [Text] / A. Holzinger, G. Langs, H. Denk, K. Zatloukal, & H. Müller // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. – 2019. – Vol. 9, iss. 4. – Article no. e1312. – DOI: 10.1002/widm.1312.

8. European Commission. Proposal for a Regulation on Artificial Intelligence (AI Act) [Electronic resource] // EUR-Lex. – 2021. – Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri= CELEX:52021PC0206 (accessed: 12.12.2025).

9. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI [Text] / A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, & et al. // Information Fusion. – 2020. – Vol. 58. – P. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

10. Ribeiro, M. T. "Why Should I Trust You?": Explaining the Predictions of Any Classifier [Text] / M. T. Ribeiro, S. Singh, & C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2016. – P. 1135–1144. DOI: 10.1145/2939672.2939778.

11. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications [Text] / W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, & K.-R. Müller // Proceedings of the IEEE. – 2021. – Vol. 109, no. 3. – P. 247–278. DOI: 10.1109/JPROC.2021.3060483.

12. Neuro-symbolic AI for auditable cognitive information extraction from medical reports [Text] / G. A. Prenosil, T. K. Weitzel, A. Afshar-Oromieh, & et al. // Communications Medicine. – 2025. – Vol. 5. – Article no. 491. DOI: 10.1038/s43856-025-01194-x.

References

1. Kingma, D.P., & Welling, M. Auto-Encoding Variational Bayes. In *Advances in Neural Information Processing Systems (NeurIPS 2013)*, vol. 27. Available at: https://arxiv.org/abs/1312.6114 (Accessed: 12 December 2025).

2. Badreddine, S., d'Avila Garcez, A., Serafini, L., & Spranger, M. Logic Tensor Networks. *Artificial Intelligence*, 2022, vol. 303, Article no. 103649. DOI:10.1016/j.artint.2021.103649.

3. Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 2019, vol. 267, pp. 1–38. DOI: 10.1016/j.artint. 2018.07.007.

4. Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI', in *Proceedings of the 26th International Conference on Intelligent User Interfaces*. New York, ACM, 2021, pp. 307–317. DOI:10.1145/3397481.3450644.

5. Nimmo, R., Constantinides, M., Zhou, K., Quercia, D., & Stumpf, S. User Characteristics in Explainable AI: The Rabbit Hole of Personalization? In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI 2024)*. 2024, New York, ACM. DOI: 10.1145/3613904.3642352.

6. Pearl, J. *Causality: Models, Reasoning, and Inference*. 2nd edn. Cambridge, Cambridge University Press, 2009. ISBN 978-0521895606.

7. 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, iss. 4, article no. e1312. DOI:10.1002/widm.1312.

8. European Commission *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. EUR-Lex, 2021. Available at: https://eur-lex.europa.eu/legal-

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

content/EN/TXT/?uri=CELEX:52021PC0206 (Accessed: 12 December 2025).

9. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., & et al. 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.

10. Ribeiro, M. T., Singh, S., & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, ACM, 2016, pp. 1135–1144. DOI:10.1145/2939672.2939778.

11. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, 2021, vol. 109, iss. 3, pp. 247–278. DOI: 10.1109/JPROC.2021.3060483.

12. Prenosil, G. A., Weitzel, T. K., Afshar-Oromieh, A., & et al. Neuro-symbolic AI for auditable cognitive information extraction from medical reports. *Communications Medicine*, 2025, vol. 5, Article no. 491. DOI: 10.1038/s43856-025-01194-x.

Отримано 15.12.2025, отримано у доопрацьованому вигляді 02.01.2026
Дата ухвалення 15.01.2026, дата публікації 22.01.2026

INTEGRATED NEURO-SYMBOLIC ARCHITECTURE OF USER MENTAL MODELS FOR PERSONALIZED EXPLANATIONS OF INTELLIGENT SYSTEMS DECISIONS

Serhii Chalyi, Iryna Leshchynska

The subject of the article is the methodology for building an integrated neuro-symbolic architecture of mental models for users with different levels of technical competence to generate personalized explanations of intelligent information systems decisions. The goal is to develop an architecture that provides automated detection of individual mental models from user behavioral data and the creation of interpretable symbolic representations of causal relationships with the adaptation of the level to detail to the user's skill level. The tasks addressed include: performing a comparative analysis of existing approaches to building mental models according to personalization criteria; developing an integrated neuro-symbolic architecture with functional distribution between neural network and symbolic components; conducting experimental verification of the proposed architecture; determining the scope of application for the developed architecture. The methods used include variational autoencoders with multi-channel attention mechanisms, neuro-symbolic translation with multi-level abstraction, and the generation of directed acyclic graphs. The following results were obtained: an integrated neuro-symbolic architecture was developed featuring a neural network component for the automated detection of individual cognitive structures through variational encoding and attention mechanisms with dynamic channel prioritization by user category, and a symbolic component for transforming latent descriptors into interpretable causal graphs with adaptive detailing and temporal validation to eliminate spurious dependencies. Conclusions. The results of the study confirmed the effectiveness of the integrated neuro-symbolic approach to building personalized mental models with the automated detection of latent cognitive structures from behavioral data without the need for expert knowledge. The scientific novelty of the results obtained lies in the development of an integrated neuro-symbolic architecture model that provides interaction between a neural network layer, designed for projecting behavioral trajectories into latent space and selecting significant features through multi-channel attention, and a symbolic layer for the neuro-symbolic translation of latent vectors into multi-level symbolic representation. This involves the generation of directed acyclic graphs and temporal validation, which improves the comprehensibility of intelligent systems' decisions through the construction of personalized, interpretable mental models. This, in turn, increases user trust in intelligent systems' decisions by tailoring explanations according to their level of technical competence.

Keywords: neuro-symbolic architecture; mental models; personalized explanations; explainable artificial intelligence; variational autoencoders; attention mechanisms; causal graphs; temporal validation; user differentiation; interpretability.

Чалий Сергій Федорович –д-р техн. наук, проф., проф. каф. інформаційних управляючих систем Харківського національного університету радіоелектроніки, Харків, Україна.

Лещинська Ірина Олександрівна –канд. техн. наук, доц., доц. каф. програмної інженерії Харківського національного університету радіоелектроніки, Харків, Україна.

Serhii Chalyi – Doctor of Technical Sciences, Professor, Professor at the Department of Information Control Systems, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine,
e-mail: serhii.chalyi@nure.ua, ORCID: 0000-0002-9982-9091.

Iryna Leshchynska – Candidate of Technical Sciences, Associate Professor, Associate Professor at the Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, e-mail: iryna.leshchynska@nure.ua, ORCID: 0000-0002-8737-4595.