DOI: 10.20998/2079-0023.2023.01.11
УДК 004.891.3

С. Ф.ЧАЛИЙ, доктор технічних наук, професор, Харківський національний університет радіоелектроніки, професор кафедри інформаційних управляючих систем, м. Харків, Україна; e mail: serhii.chalyi@nure.ua, ORCID: https://orcid.org/0000-0002-9982-9091
І. О. ЛЕЩИНСЬКА, кандидат технічних наук (PhD), доцент, Харківський національний університет радіоелектроніки, доцент кафедри програмної інженерії; м. Харків, Україна, ; e–mail: iryna.leshchynska@nure.ua, ORCID: https://orcid.org/0000-0002-8737-4595

КОНЦЕПТУАЛЬНА МЕНТАЛЬНА МОДЕЛЬ ПОЯСНЕННЯ В СИСТЕМІ ШТУЧНОГО ІНТЕЛЕКТУ

Предметом дослідження є процеси формування пояснень щодо отриманих в системах штучного інтелекту рішень. Для вирішення проблеми непрозорості прийняття рішень в таких системах користувачі мають отримати пояснення щодо отриманих рішень. Пояснення дозволяє довіряти цим рішенням та забезпечити їх використання на практиці. Мета роботи полягає у розробці концептуальної ментальної моделі пояснення для визначення базових залежностей, що визначають зв'язок між вхідними даними, а також діями з отримання результату в інтелектуальній системі, та її кінцевим рішенням. Для досягнення мети вирішуються такі задачі: структуризація підходів до побудови ментальних моделей пояснень; побудова концептуальної ментальної моделі пояснення на основі об'єднаного представлення знань користувача.. Висновки. Виконано структуризацію підходів до побудови ментальних моделей пояснень в інтелектуальних системах. Ментальні моделі призначені для відображення сприйняття пояснення користувачем. Виділено каузальний, статистичний, семантичний та концептуальний підходи до побудови ментальних моделей пояснення. Показано, що концептуальна модель задає узагальнені схеми та принципи щодо процесу функціонування інтелектуальної системи. Її подальша деталізація виконується на основі каузального підходу у випадку побудови пояснення для процесів, статистичного підходу при побудові пояснення щодо результату роботи системи, а також семантичного при узгодженні пояснення із базовими знаннями користувача. Запропоновано трирівневу концептуальну ментальну модель пояснення, що містить рівні концепції щодо базових принципів функціонування системи штучного інтелекту, пояснення, що деталізує цю концепцію у прийнятному та зрозумілому для користувача вигляді, а також базових знань про предметну область, які є основою для формування пояснення. У практичному аспекті запропонована модель створює умови для побудови та упорядкування множини узгоджених пояснень, які описують процес та результат роботи інтелектуальної системи з урахуванням можливості їх сприйняття користувачем.

Ключові слова: : пояснення, система штучного інтелекту, зрозумілий штучний інтелект, залежності, ментальна модель, каузальні залежності.

Вступ. У процесі створення сучасних інтелектуальних інформаційних систем знайшли широке застосування методи машинного навчання, які дають можливість сформувати моделі прийняття рішень на основі виявлення закономірностей у великих масивах даних [1]. Отримані складні моделі, що є «ядром» таких систем, стають непрозорими й незрозумілими для користувачів. Тому користувачі можуть не довіряти рішенням, запропонованим такими інтелектуальними системами, що зменшує ефективність використання отриманих результатів [2].

Для вирішення проблеми непрозорості прийняття рішень в системах штучного інтелекту користувачі мають отримати пояснення щодо отриманих рішень. Пояснення дозволяє довіряти цим рішенням та забезпечити їх використання на практиці [3].

Наведені положення вказують на важливість інтерпретації для користувача процесу формування рішень та отриманого в інформаційній системі результату.

Для вирішення цієї проблеми застосовуються два підходи:

– використання моделей, що можуть бути інтерпретовані безпосередньо;

– використання пояснень для «непрозорих» алгоритмів та моделей, що лежать в основі функціонування інтелектуальної системи.

Проблема інтерпретації процесу прийняття рішень в рамках першого підходу виникає лише у випадку комерційних обмежень на інформацію щодо структури моделі.

В рамках другого підходу підсистема пояснень вбудовується на етапі розробки інтелектуальної системи або ж розробляється окремо для вже існуючої систем. При побудові пояснень, згідно з концепцією зрозумілого (пояснювального) штучного інтелекту [3, 4], вирішуються три базових задачі:

– побудова ментальних моделей пояснень на основі узагальнення та розширення психологічних теорій пояснень;

– розробка нових методів машинного навчання та методів побудови тлумачень для створення ефективних пояснень;

– оцінка ефективності методів побудови пояснень при вирішенні задач аналітики даних та підтримки автономних агентів.

Вирішення першої задачі дає можливість формалізувати сприйняття пояснення людиною-користувачем інтелектуальної системи. Також реалізація першої задачі створює умови для розробки нових методів побудови пояснень та оцінки останніх при вирішенні задач аналізу даних та розробки автономних агентів.

Використання ментальних моделей забезпечує умови для представлення пояснення у формі, зрозумілій для користувача, з урахування рівня його підготовки, а також його знань [5]. Зазначене свідчить про актуальність даного напрямку досліджень.

Аналіз останніх досліджень і публікацій.

Дослідження щодо формування пояснень в сучасних системах штучного інтелекту були суттєво інтенсифіковані в рамках програми зрозумілого штучного інтелекту (XAI) [3], де було виділено як окремий напрямок досліджень задачу побудови ментальних моделей пояснень [4]. Можливості моделювання пояснень з урахуванням когнітивних аспектів, що суттєво впливає на сприйняття цих тлумачень людьми, було розглянуто в роботі [6]. Психологічні аспекти тлумачень, які визначають можливість побудови концептуальної ментальної моделі пояснень, розглянуті в роботах [7–9]. В цих роботах виділено структуру та функції пояснень. Переваги побудови пояснень з урахуванням як каузального, так і темпорального аспектів були розглянуті в роботах [10–12].

Виконаний аналіз показав, що на сьогодні головна увага при вирішенні задачі побудови ментальних моделей пояснень приділялась побудові описових психологічних моделей пояснень, а також спеціалізованих ментальних моделей, які відображають сприйняття пояснення щодо окремих задач та процесів в інтелектуальній системі.

Однак для розробки методів побудови тлумачень необхідно також розробити концептуальну модель пояснень, яка може бути доповнена локальними моделями. Така модель дає можливість користувачеві зрозуміти ключові принципи побудови пояснень, що створює для ефективного використання користувачем рішень інтелектуальної системи і свідчить про важливість вирішення задачі побудови такої моделі.

Мета та задачі дослідження. Мета роботи полягає у розробці концептуальної ментальної моделі пояснення для визначення базових залежностей, що визначають зв'язок між вхідними даними, а також діями з отримання результату в інтелектуальній системі, та її кінцевим рішенням.

Для досягнення мети вирішуються такі задачі:

– структуризація підходів до побудови ментальних моделей пояснень;

– побудова концептуальної ментальної моделі пояснення на основі об'єднаного представлення знань користувача.

Структуризація підходів до побудови ментальної моделі пояснень.

Мета розробки ментальної моделі пояснення – забезпечити розуміння користувачем внутрішніх процесів системи штучного інтелекту. Забезпечуючи прозоре та інтерпретоване пояснення, ментальна модель має створювати умови для підвищення довіри користувача, а також забезпечувати користувачам можливість оцінити релевантність рішень системи штучного інтелекту. Модель має сприяти ефективному використанню отриманих в інтелектуальній системі результатів.

Ментальна модель пояснень визначає концептуальне представлення, яке використовує інтелектуальна система для тлумачення своїх рішень або дій з отримання цих рішень. Це адаптоване до задачі побудови пояснень представлення знань щодо внутрішніх процесів та послідовності виводу в системі штучного інтелекту, яке може бути інтерпретовано з тим, щоб допомогти користувачам або зацікавленим сторонам зрозуміти, чому інтелектуальна система прийняла певне рішення.

Ключові підходи до побудови ментальної моделі пояснень базуються на підходах до побудови представлення знань і містять у собі: каузальний; статистичний; семантичний; концептуальний підходи.

Порівняльну характеристику розглянутих підходів наведено у табл. 1.

Вибір форми представлення ментальної моделі залежить від задач, які вирішує система штучного інтелекту, а також контексту вирішення цих задач.

Підхід до побудови ментальної моделі пояснення, оснований на каузальних залежностях (наприклад, правилах), забезпечує формування тлумачень рішень системи штучного інтелекту на основі набору попередньо визначених правил: обмежень або умов. Відмінності між ними полягають у тому, що обмеження є істинними для всіх варіантів процесу формування рішення, а умови – лише для підмножини таких альтернативних варіантів.

Таблиця 1 – Підходи до побудови ментальної моделі пояснень

Підхід	Особливості
1.Каузальний	Використовуються каузальні залежності. Можливість надати пояснення як результату роботи інтелектуальної системи, так і процесу отримання даного результату.
2. Статистичний	Використовуються патерни та залежності, отримані на основі аналізу даних інтелектуальної системи. Орієнтований в першу чергу на пояснення рішення системи штучного інтелекту.
3. Семантичний	Використовуються знання з предметної області (а тому числі темпоральні) щодо процесу прийняття рішення та результату інтелектуальної системи. Можливість адаптації до рівня знань користувача.
4. Концептуальний	Використовуються знання щодо загальних принципів, схеми, технологій та процесу прийняття рішень у предметній області. Можливість побудови високорівневого пояснення, що є узагальненим для різних контекстів прийняття рішень у предметній області. Можливість деталізації на основі каузального, статистичного та семантичного підходів.

Вказані залежності мають охоплювати та узагальнювати весь процес прийняття рішень інтелектуальною системою з тим, щоб надати прозоре пояснення як окремих дій, так і їх послідовностей, що забезпечують досягнення результуючого рішення.

У каузальній ментальній моделі пояснення виводяться шляхом оцінки вхідних та проміжних даних з використанням набору логічних правил. Кожне правило складається з умови або набору умов, які перевіряють значення певних змінних, і пов'язаного пояснення або подальшої дії. Коли умови правила задовольняються, відповідне пояснення надається як результат. Або ж виконується дія із уточнення можливого пояснення. Послідовність таких правил може надати пояснення як для процесу, так і для результату, отриманого в системі штучного інтелекту.

Фактично, правила визначають граничні умови, які впливають на прийняття рішення.

Перевага даного підходу до побудови ментальної моделі полягає у можливості її безпосередньої інтерпретації. Такі пояснення є інтуїтивно зрозумілими, оскільки вони базуються на детермінованих залежностях, які спрощують процес прийняття рішень в системі штучного інтелекту.

Недолік підходу пов'язаний із обмеженнями в представленні складних зв'язків між вхідними даними і отриманим результатом, а також невизначеності яка може виникати в результаті зовнішніх впливів на інтелектуальну систему. Зазвичай такий підхід підходить для випадків, коли можуть бути визначені чіткі умови та обмеження у процесі прийняття рішення.

Побудова ментальної моделі на основі правил передбачає визначення правил, вибір функцій оцінки правил для відбору альтернатив пояснення, а також постійне вдосконалення множини правил з урахуванням змін і доповнень у процесі прийняття рішення у системі штучного інтелекту.

Статистичний підхід до побудови ментальної моделі пояснення в системі штучного інтелекту орієнтований на визначення ключових особливостей або факторів, які вплинули на отримане рішення. Пояснення виводяться на основі аналізу патернів та закономірностей, наявних у даних, які використовує система штучного інтелекту. Ці закономірності можуть бути отримані методами машинного навчання. На відміну від попереднього підходу, в даному випадку ключова увага приділяється встановленню зв'язків між вхідними даними та кінцевим результатом інтелектуальної системи.

Перевага даного підходу полягає у здатності зафіксувати складні нелінійні зв'язки та враховувати невизначеність в даних. Також враховується контекст прийняття рішення.

Недолік підходу полягає базується на його ймовірнісній природі і полягає у відсутності визначення детермінованих каузальних залежностей.

Семантичний підхід до побудови ментальної моделі пояснення базується на використанні семантики предметної області. В даному випадку можуть бути використані онтології, семантичні мережі, графи знань. Розробка представлення знань може бути виконана як за допомогою інженерів знань, так і з використанням методів автоматизованої побудови бази знань.

Перевага даного підходу полягає у можливості надавати чіткі, інтерпретовані та зрозумілі пояснення, спираючись на зрозумілі людині представлення. Такий підхід може відображати складні відношення, в тому числі каузальні та темпоральні, дозволяючи користувачам пояснити розгортання процесу прийняття рішень у часі.

Недолік підходу полягає пов'язаний із обмеженням при формуванні знань та при поясненні зовнішніх впливів, оскільки вони не завжди відображені у відповідній базі знань.

Підхід до побудови абстрактної ментальної моделі пояснення в системі штучного інтелекту базується на використанні загальних принципів, концепцій та моделей, що лежать в основі процесу прийняття рішень. Цей підхід має на меті забезпечити концептуальне розуміння поведінки інтелектуальної замість пояснення конкретних процесів або результатів.

Концептуальна ментальна модель пояснення містить знання щодо структури та властивостей предметної області. Така модель базується на психологічних теоріях пояснення. Останні використовуються при аналізі навчання дітей у людському суспільстві. Наприклад, концептуальне пояснення щодо розпізнавання зображень має охоплювати принципи розпізнавання образів, виділення ознак, класифікації, а також концепції згорткових нейронних мереж, тощо. В даному випадку замість правил або патернів пояснень щодо конкретного результату модель підкреслює загальну схему, процес та принципи розпізнавання зображень у конкретній інтелектуальній системі.

У практичному аспекті концептуальне представлення може визначати ключові елементи та принципи обробки вхідних даних, які інтелектуальна система використовує для прийняття рішення.

Наприклад, для пояснення щодо рішення «людина на зображенні біжить», прийнятого системою розпізнавання зображень, можуть бути виділені зони зображення, що відображають відповідні фрагменти тіла людини у русі. В даному випадку концепція базується на визначенні підмножини фрагментів зображення, які є ключовим для прийнятого в інтелектуальній системі рішення.

Аналогічно, концептуальним поясненням щодо кредитного рішення в інтелектуальній системі, що має відповідати сприйняттю людини, може виступати залежність між кредитним рейтингом та доходом клієнта.

Перевага даного підходу полягає в тому, що концептуальна ментальна модель забезпечує розуміння поведінки інтелектуальної системи на більш високому рівні абстракції, що дає можливість узагальнити контекст прийняття рішення в рамках пояснення. Такий підхід дозволяє користувачам зрозуміти базові принципи та концепції прийняття рішень системою штучного інтелекту, що дає можливість сформувати умови та обмеження щодо застосування цих рішень, визначити семантику цих рішень, а також передбачити додаткові можливості їх використання.

Недолік концептуального підходу до побудови ментальної моделі пояснень полягає у відсутності конкретних деталей щодо конкретних рішень інтелектуальної системи. Тому даний підхід потребує доповнення можливостями інших підходів..

Зв'язок підходів до побудови ментальних моделей пояснення представлено на рис. 1. Даний зв'язок відображає загальну схему побудови системи ментальних моделей в залежності від вимог до пояснень в системі штучного інтелекту.

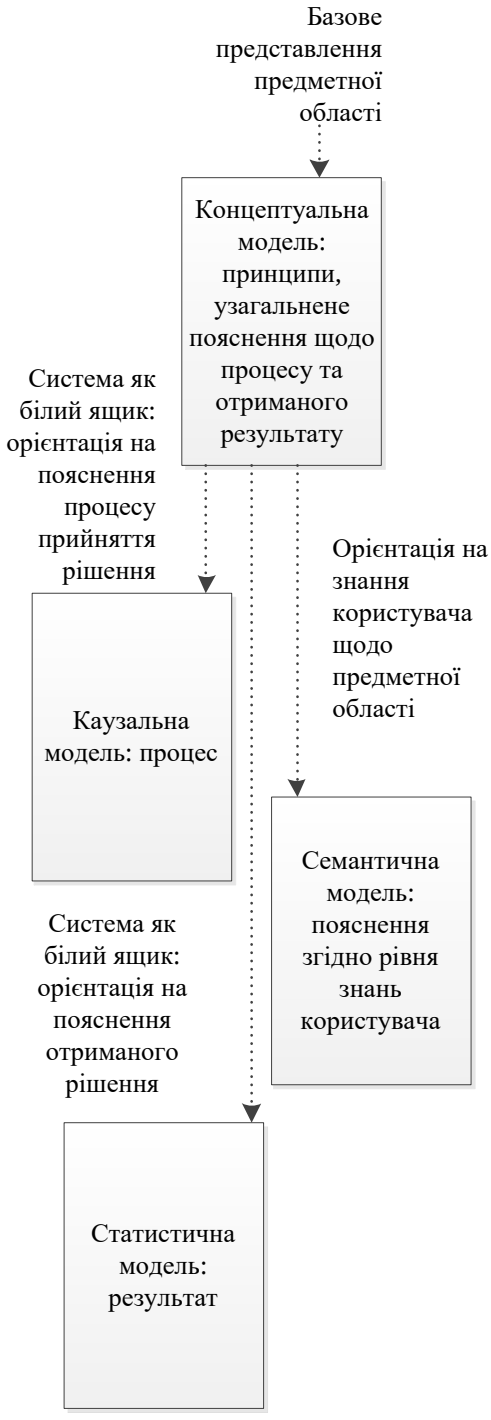

Рис. 1. Послідовність та умови реалізації підходів до побудови ментальних моделей пояснень

Порівняння представлених підходів до побудови ментальної моделі пояснень дає можливість зробити висновок про те, що концептуальна ментальна модель пояснення має розроблятись в першу чергу і в подальшому розширюватись з використанням розглянутих каузального, статистичного та семантичного підходів. Статистичний підхід моє переваги у випадку вирішення задачі пояснення результату роботи інтелектуальної системи, каузальний – для тлумачення процесу отримання результату. Семантичний підхід дає можливість адаптувати пояснення до рівня підготовки користувача та створює умови для пояснення щодо способу використання отриманого в системі результату.

Концептуальна ментальна модель пояснення.

Запропонована концептуальна ментальна модель складається із елементів, пов'язаних структурними та темпоральними відношеннями, враховує онтологію понять в предметній області, а також може якісно описувати впливи на процес прийняття рішень в системі штучного інтелекту.

Ключова особливість концептуальної моделі полягає в тому, що вказані елементи можуть бути неузгодженими. Це створює умови для побудови альтернативних варіантів пояснень. В подальшому кожен варіант пояснення може бути ймовірнісно оцінений.

Найбільш ймовірний варіант розглядається як ключове пояснення. Отримане пояснення має бути пов'язано із концепцією, що узагальнює ситуацію та умови прийняття рішення. В результаті базовим елементом концептуальної моделі є трійка (знання, що деталізують певну концепцію щодо прийняття рішення; безпосередньо концепція; пояснення).

Знання щодо предметної області визначають властивості певної концепції прийняття рішень в інтелектуальній системі і можуть містити:

– опис структури предметної області;

– опис базових залежностей, на яких базується процес прийняття рішень у системі штучного інтелекту;

– опис зовнішніх впливів на процес прийняття рішення в системі:

– показник оцінки пояснень, що дає можливість упорядкувати їх за відповідністю концепції.

Тобто концепція як узагальнене представлення щодо певної категорії або класу об'єктів та процесів, а також їх властивостей і взаємозв'язків між ними пояснюється через ієрархічну структуру предметної області та процесу використання цієї структури для отримання рішення.

Узагальнену структуру концептуальної моделі пояснення наведено на рис. 2.

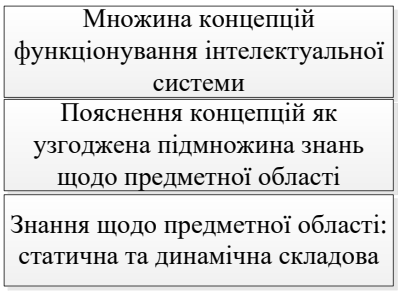

Рис. 2. Структура концептуальної моделі пояснення

Згідно представленої структури, концептуальна модель пояснення для користувача M складається із множини концепцій $C=\{c_i\}$, пояснень $Expl=\{\exp l_{i,j}\}$, а також знань для формування пояснень $K=\{k_m\}$ за умови локальної узгодженості концепції та пояснення:

$$M=\langle C, Expl, K \mid Expl \subset K\rangle. \quad (1)$$

З урахуванням можливих протиріч у знаннях щодо процесів прийняття рішень в системі штучного інтелекту, запропонована модель використовує принцип локальної узгодженості знань.

Згідно даного принципу кожна пара (*концепція*; *пояснення*) має бути узгодженою, тобто має характеризуватись властивостями повноти та несуперечливості.

Іншими словами, із множини знань K щодо предметної області для пояснення $\exp l_{i,j}$ відбираються лише узгоджені з концепцією знання. Узгодженість пояснення і концепції $c_i \sim \exp l_{i,j}$ визначається наступним чином:

$$\begin{aligned} &c_i \sim \exp l_{i,j} \equiv (\forall i)\{\exp l_{i,j}\} \neq \varnothing, \\ &(\forall i, j, m)(\exp l_{i,j} \neg \exp l_{i,m}) = \text{false}. \end{aligned} \quad (2)$$

Концепція в рамках моделі пов'язує об'єкти o_l, o_m відношенням r_m^l, що дає можливість визначити високорівневі зв'язки у предметній області. Відповідно, концепція визначається трійкою:

$$o_l r_m^l o_m, (o_l, o_m) \in O, r_m^l \in R. \quad (3)$$

Представлення концепції у вигляді (3) дає можливість визначити як ключові складові ієрархічної структури предметної області, що впливають на прийняття рішення, так і задати ключові залежності у самому процесі прийняття рішення.

В першому випадку може бути використаний онтологічний підхід, а в другому – каузальні або темпоральні залежності.

Концепція у вигляді (3) формує узагальнені запитання для пояснення щодо результату роботи інтелектуальної системи : Чому об'єкти o_l, o_m мають між собою зв'язок виду r_m^l? Запитання щодо процесу прийняття рішення має вигляд: Чому за об'єктом (або дією) o_l при прийнятті рішення було використано об'єкт o_m способом r_m^l ? Концептуальні запитання щодо семантики прийняття рішення можуть мати вигляд: Який сенс (можливості використання) має отриманий як рішення об'єкт o_m для вхідного об'єкту o_l з урахуванням контексту прийняття рішення r_m^l ?

У наведеному вище прикладі щодо аналізу зображень в якості об'єкту o_l концепції виступає фрагмент зображення людини (зокрема, зображення ніг), який свідчить про її рух, в якості об'єкту o_m виступає характеристика руху людини (біжить, стоїть, йде), а залежність r_m^l визначає принцип розпізнавання руху на зображенні: за станом певних фрагментів тіла людини.

Кожний об'єкт концепції o_l задається через множину його властивостей $v_{l,n}$: $o_l = \{v_{l,n}\}$. Таке представлення об'єкту дає можливість деталізувати концепцію через залежності між окремими властивостями об'єктів. Такі залежності $\rho_{m,q}^{l,n}$ можуть виступати в якості основи для пояснень $\exp l_{i.j}$ щодо концепції c_i :

$$\exp l_{i.j} = v_{l,n} \rho_{m,q}^{l,n} v_{m,q}. \quad (4)$$

Висновки. Виконано структуризацію підходів до побудови ментальних моделей пояснень в інтелектуальних системах. Ментальні моделі призначені для відображення сприйняття пояснення користувачем. Виділено каузальний, статистичний, семантичний та концептуальний підходи до побудови ментальних моделей пояснення. Показано, що концептуальна модель задає узагальнені схеми та принципи щодо процесу функціонування інтелектуальної системи. Її подальша деталізація виконується на основі каузального підходу у випадку побудови пояснення для процесів, статистичного підходу при побудові пояснення щодо результату роботи системи, а також семантичного при узгодженні пояснення із базовими знаннями користувача.

Запропоновано трирівневу концептуальну ментальну модель пояснення, що містить рівні концепції щодо базових принципів функціонування системи штучного інтелекту, пояснення, що деталізує цю концепцію у прийнятному та зрозумілому для користувача вигляді, а також базових знань про предметну область, які є основою для формування пояснення. У практичному аспекті запропонована модель створює умови для побудови та упорядкування множини узгоджених пояснень, які описують процес та результат роботи інтелектуальної системи з урахуванням можливості їх сприйняття користувачем.

Список використаної літератури

1. Engelbrecht Andries P. *Computational Intelligence: An Introduction*. New York: John Wiley & Sons, 2007. 632 p.
2. Tintarev N., Masthoff J. A survey of explanations in recommender systems. *The 3rd international workshop on web personalisation, recommender systems and intelligent user interfaces (WPRSIUI'07).* 2007. P. 801–810.
3. Gunning D., Aha D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*. 2019. Vol. 40(2). P. 44–58.
4. Gunning D., Vorm E., Wang J., Turek M. DARPA's explainable AI (XAI) program: A retrospective, *Applied AI Letters*. Vol. 2, no. 4. 2021. DOI: https://doi.org/10.1002/ail2.61/
5. Gilpin L. H., Bau D., Yuan B. Z., Bajwa A., Specter M., Kagal L. Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv:1806.00069*. 2018.
6. Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial. Intelligence*. 2019. Vol. 267 P. 1–38.
7. Chi M., de Leeuw N., Chiu M., Lavancher C. Eliciting self-explanations improves understanding. *Cognitive Science.* 1994. Vol. 18. P. 439–477.

8. Carey S. *The origin of concepts*. New York: Oxford University Press. 2009. 608 p.
9. Holyoak Keith J., Morrison Robert G. *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press, 2012. 864 p.
10. Чалий С.Ф., Лещинський В.О., Лещинська І.О. Декларативно-темпоральний підхід до побудови пояснень в інтелектуальних інформаційних системах. *Вісник Нац. техн. ун-ту "ХПІ": зб. наук. пр. Темат. вип. Системний аналіз, управління та інформаційні технології*. Харків: НТУ «ХПІ», 2020. № 2(4). С. 51–56.
11. Halpern J. Y., Pearl J. *Causes and explanations: A structural-model approach. Part II: Explanations*. URL: https://arxiv.org/pdf/cs/0208034.pdf (дата звернення: 11.05.202).
12. Chalyi S., Leshchynskyi V. Temporal representation of causality in the construction of explanations in intelligent systems. *Advanced Information Systems*. Kharkiv: NTU "KhPI", 2020. Vol. 4, № 3. P. 113–117.

References (transliterated)

1. Engelbrecht Andries P. *Computational Intelligence: An Introduction*. NJ: John Wiley & Sons, 2007. 632 p.
2. Castelvecchi D. Can we open the black box of AI? *Nature News* 2016. Vol. 538 (7623). P. 20.
3. Tintarev N., Masthoff J. A survey of explanations in recommender systems. *The 3rd International workshop on web personalisation, recommender systems and intelligent user interfaces (WPRSIUI'07)*. 2007, pp. 801–810.
4. Gunning D., Vorm E., Wang J., Turek M. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*. Vol. 2, no. 4, 2021. DOI: https://doi.org/10.1002/ail2.61.
5. Gilpin L. H., Bau D., Yuan B. Z., Bajwa A., Specter M., Kagal L. Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv:1806.00069*. 2018.
6. Miller T. Explanation in artificial intelligence: Insights from the social sciences, *Artificial. Intelligence*. 2019. Vol. 267 P. 1–38.
7. Chi M., de Leeuw N., Chiu M., Lavancher C. Eliciting self-explanations improves understanding. *Cognitive Science*. 1994. Vol.18. P. 439–477.
8. Carey S. *The origin of concepts*. New York: Oxford University Press. 2009. 608 p.
9. Holyoak Keith J., Morrison Robert G. *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press, 2012. 864 p.
10. Chalyi S., Leshchynskyi V., Leshchynska I. Deklaratyvno-temporalnyi pidkhid do pobudovy poiasnen v intelektualnykh informatsiinykh systemakh [Declarative-temporal approach to the construction of explanations in intelligent information systems]. *Visnyk Nats. tekhn. un-tu "KhPI": zb. nauk. pr. Temat. vyp. Systemnyi analiz, upravlinnia ta informatsiini tekhnolohii* [Bulletin of the National Technical University "KhPI": a collection of scientific papers. Thematic issue: System analysis, management and information technology]. Kharkov, NTU "KhPI" Publ., 2020, no. 2(4), pp. 51–56.
11. Halpern J. Y., Pearl J. *Causes and explanations: A structural-model approach. Part II: Explanations*. Available at: https://arxiv.org/pdf/cs/0208034.pdf (accessed 11.05.202).
12. Chalyi S., Leshchynskyi V. Temporal representation of causality in the construction of explanations in intelligent systems. *Advanced Information Systems*. 2020, vol. 4, no 3, pp. 113–117.

Надійшла (received) 14.05.2023

UDC 004.891.3

S. F. CHALYI, Doctor of Technical Sciences, Full Professor, Kharkiv National University of Radio Electronics, Professor of the Department of Information Control System, Kharkiv; e mail: serhii.chalyi@nure.ua, ORCID: https://orcid.org/0000-0002-9982-9091

I. O. LESHCHYNSKA, Candidate of Technical Sciences (PhD), Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor at the Department of Software Engineering доцент кафедри програмної інженерії, Kharkiv; e–mail: iryna.leshchynska@nure.ua, ORCID: https://orcid.org/0000-0002-8737-4595

THE CONCEPTUAL MENTAL MODEL OF EXPLANATION IN AN ARTIFICIAL INTELLIGENCE SYSTEM

The subject of research is the process of formation of explanations in artificial intelligence systems. To solve the problem of the opacity of decision-making in artificial intelligence systems, users should receive an explanation of the decisions made. The explanation allows you to trust these solutions and ensure their use in practice. The purpose of the work is to develop a conceptual mental model of explanation to determine the basic dependencies that determine the relationship between input data, as well as actions to obtain a result in an intelligent system, and its final solution. To achieve the goal, the following tasks are solved: structuring approaches to building mental models of explanations; construction of a conceptual mental model of explanation based on a unified representation of the user's knowledge. Conclusions. The structuring of approaches to the construction of mental models of explanations in intelligent systems has been carried out. Mental models are designed to reflect the user's perception of an explanation. Causal, statistical, semantic, and conceptual approaches to the construction of mental models of explanation are distinguished. It is shown that the conceptual model sets generalized schemes and principles regarding the process of functioning of the intellectual system. Its further detailing is carried out on the basis of a causal approach in the case of constructing an explanation for processes, a statistical approach when constructing an explanation about the result of the system's work, as well as a semantic approach when harmonizing the explanation with the user's basic knowledge. A three-level conceptual mental model of the explanation is proposed, containing levels of concepts regarding the basic principles of the functioning of the artificial intelligence system, an explanation that details this concept in an acceptable and understandable way for the user, as well as basic knowledge about the subject area, which is the basis for the formation of the explanation. In a practical aspect, the proposed model creates conditions for building and organizing a set of agreed explanations that describe the process and result of the intelligent system, considering the possibility of their perception by the user.

Keywords: explanation; artificial intelligence system; understandable artificial intelligence; dependencies; mental model; causal dependence.

Повні імена авторів / Author's full names

Автор 1 / Author 1: Чалий Сергій Федорович**,** Chalyi Serhii Fedorovych
Автор 2 / Author 2: Лещинська Ірина Олександрівна, Leshchynska Irina Oleksandrivna