DOI: 10.20998/2079-0023.2023.02.14
УДК 004.8:004.9

***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
***В. О. ЛЕЩИНСЬКИЙ***, кандидат технічних наук (PhD), доцент, Харківський національний університет радіоелектроніки, доцент кафедри програмної інженерії; м. Харків, Україна, ; volodymyr.leshchynskyi@nure.ua, ORCID: https://orcid.org/0000-0002-8690-5702

# МЕТОД МОЖЛИВІСНОГО ОЦІНЮВАННЯ ПОЯСНЕННЯ В СИСТЕМІ ШТУЧНОГО ІНТЕЛЕКТУ

Предметом дослідження є процеси формування пояснень щодо рішення системи штучного інтелекту. Пояснення використовуються для того, щоб користувач зрозумів процес отримання результату і міг більш ефективно застосовувати інтелектуальну інформаційну систему для формування практично прийнятих для нього рішень. Мета роботи полягає у розробці методу оцінки пояснень з урахуванням відмінностей у вхідних даних та відповідному рішенні системи штучного інтелекту. Вирішення цієї задачі дає можливість оцінити відповідність пояснення щодо внутрішньому механізму прийняття рішення в інтелектуальній інформаційній системі незалежно від рівня знань користувача щодо особливостей формування та використання такого рішення. Для досягнення мети вирішуються такі задачі: структуризація оцінки пояснень в залежності від рівня їх деталізації з урахуванням їх відповідності процесу прийняття рішення в інтелектуальній системі та рівню сприйняття користувача такої системи; розробка методу оцінки пояснень на основі їх відповідності процесу прийняття рішення в інтелектуальній системі. Висновки. Виконано структуризацію оцінки пояснень в залежності від рівня їх деталізації. Виділено рівні асоціативних залежностей, прецедентів, каузальних залежностей та інтерактивний, що визначають різний ступінь деталізації пояснень. Показано, що асоціативний та каузальний рівні деталізації пояснень можуть бути оцінені з використанням числових, ймовірнісних або можливісних показників. Прецедентний та інтерактивний рівні потребують суб'єктивної оцінки на основі опитування користувачів системи штучного інтелекту. Розроблено метод можливісного оцінювання відповідності пояснень процесу прийняття рішень в інтелектуальній системі з урахуванням залежностей між вхідними даними та рішенням інтелектуальної системи. Метод містить етапи оцінювання чутливості, коректності та складності пояснення на основі порівняння значень та кількості використаних у поясненні вхідних даних. Метод дає можливість комплексно оцінити пояснення з позицій стійкості до несуттєвих змін у вхідних даних, відповідності пояснення отриманому результату, а також складності обчислення пояснення. У аспекті практичного застосування метод дає можливість мінімізувати кількість вхідних змінних для пояснення при задоволенні обмеження на чутливість пояснення, що створює умови для більш ефективного формування тлумачення на основі використання підмножини ключових вхідних змінних, які мають суттєвий вплив на отримане в інтелектуальній системі рішення.

**Ключові слова:** пояснення, оцінка пояснення, система штучного інтелекту, інтелектуальна система, зрозумілий штучний інтелект, асоціативна залежність, каузальна залежність, прецедент, інформаційна система, рекомендаційна система.

**Вступ.** Ефективне застосування сучасних систем штучного інтелекту базується на використанні методів машинного навчання, які орієнтовані на побудову алгоритмів прийняття рішень з використанням виявлених у наборах даних патернів та закономірностей [1].

Навчені моделі часто є непрозорими і, відповідно, незрозумілими для користувачів інтелектуальних систем. Це знижує ступінь довіри до запропонованих рішень. В результаті користувача може відмовитись від використання рішень системи штучного інтелекту і в цілому знижується ефективність її застосування [2].

Для вирішення даної проблеми використовуються пояснення, що представляють користувачеві причини прийнятих рішень, а також причини окремих дій із процесу прийняття цих рішень. Такий підхід створює умови для успішного застосування рішень інтелектуальної системи при вирішенні практичних задач користувача [3].

При побудові «прозорої» системи штучного інтелекту використовуються підходи на основі білого та чорного ящиків [4]. В першому випадку використовуються моделі, які можуть бути інтерпретовані користувачем, наприклад дерева рішень [5]. Тому проблема інтерпретації в рамках першого підходу зазвичай є наслідком організаційних обмежень.

У другому випадку паралельно з моделлю прийняття рішення в системі штучного інтелекту формується спрощена модель процесу отримання рішення, яка використовується для пояснення. Ця модель не в повній мірі відповідає базовій моделі інтелектуальної системи [6]. Невідповідність між моделями може привести до некоректних пояснень, що свідчить про актуальність проблематики оцінки пояснення щодо його відповідності процесу прийняття рішення та відповідності потребам практичного застосування отриманого в інтелектуальній системі результату [7–9].

На сьогодні в межах вказаної задачі послідовно вирішуються дві підзадачі.

Перша орієнтована на перевірку відповідності пояснення отриманому результату при відомих вхідних даних. Оцінювання виконується з використанням властивостей вхідних даних та результату.

Друга пов'язана із оцінкою того, як пояснення сприймається користувачем системи штучного інтелекту. Оцінювання виконується на основі опитувань та інтерактивної взаємодії із користувачами системи.

В даній роботі розглядається перша підзадача. Її вирішення є необхідною умовою для ефективного розв'язання другої задачі, оскільки невідповідність між поясненням та процесом прийняття рішення суттєво знижує ефективність осмислення такого процесу користувачем.

Аналіз останніх досліджень і публікацій.

Сучасні напрямки досліджень щодо побудови пояснень в інтелектуальних системах виділено в рамках програми зрозумілого штучного інтелекту (XAI),

реалізованої в останні роки DARPA (Агентство оборонних досліджень проектів з розширених можливостей) [3, 8]. Мета зрозумілого штучного інтелекту полягає в створенні таких інтелектуальних систем, які можуть пояснювати свої висновки користувачем у такій формі, щоб останні могли зрозуміти як логіку їх функціонування, так і послідовність робіт процесу отримання рішення.

Існуючі методи оцінки пояснень спрямовані переважно визначення впливу вхідних даних на процес прийняття рішень [2, 3, 8,10, 11]. Такі підходи є корисними, коли інтелектуальна система визначається як «чорний ящик». Також розглядається сприйняття пояснень самими користувачами [2]. Однак, довіра користувачів до результатів та процесу прийняття рішень в цілому визначається коректністю, правильністю цих висновків. Оцінка коректності рішень інтелектуальної системи розглядається в роботі [10]. В цілому в існуючих роботах основна увага приділяється оцінюванню окремих аспектів пояснення [2, 10, 11], виділяються відповідні показники. Однак комплексній оцінці пояснення з урахуванням відповідності пояснення процесу прийняття рішень, а також складності реалізації пояснення не приділялось достатньо уваги.

Визначення такої комплексної оцінки у загальному випадку суттєво залежить від предметної області. Уніфікація оцінки пояснень з урахуванням вказаних відмінностей може бути виконана на базі теорії можливостей, оскільки остання дозволяє ймовірнісно описати використання вхідних даних різних типів, зв'язок даних з результатом, а також можливості зменшення складності пояснення [12].

Таким чином, важливою є задача формування комплексної оцінки пояснення, оскільки вона дає можливість спростити пояснення, забезпечуючи разом з тим формування коректного тлумачення, що відповідає процесу прийняття рішення в інтелектуальній системі з урахуванням поточних значень вхідних даних.

**Мета та задачі дослідження.** Мета роботи полягає у розробці методу оцінки пояснень з урахуванням відмінностей у вхідних даних та відповідному рішенні системи штучного інтелекту. Вирішення цієї задачі дає можливість оцінити відповідність між пояснення внутрішньому механізму прийняття рішення в інтелектуальній системі незалежно від рівня знань користувача щодо особливостей формування та використання такого рішення.

Для досягнення мети роботи вирішуються такі задачі:

– структуризація оцінки пояснень в залежності від рівня їх деталізації з урахуванням їх відповідності процесу прийняття рішення в інтелектуальній системі та рівню сприйняття користувача такої системи;

– розробка методу оцінки пояснень на основі їх відповідності процесу прийняття рішення в інтелектуальній системі.

Структуризація рівнів деталізації та оцінки пояснень

Типовий процес пояснення в системі штучного інтелекту має такі кроки:

– реалізація процесу побудови пояснень;

– надання згенерованого пояснення користувачеві інтелектуальної системи;

– осмислення пояснення користувачем;

– підвищення ефективності роботи користувача з інтелектуальною системою.

Необхідною умовою реалізації останнього етапу є оцінка пояснення. Така оцінка може бути надана як на етапі його побудови, так і на етапі осмислення пояснення користувачем (рис. 1).

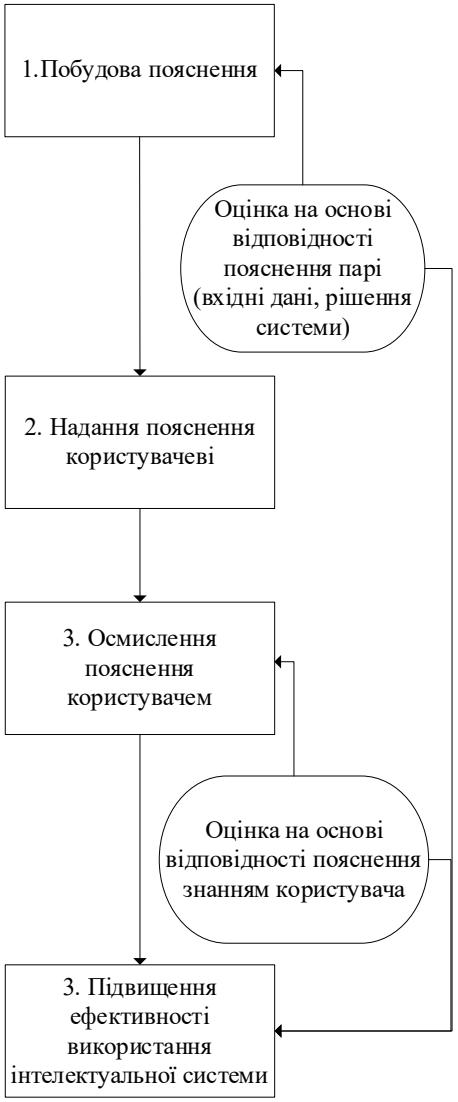


Рис. 1. Оцінка пояснення на першому та третьому етапах формування тлумачення

У першому випадку оцінка дає можливість установити відповідність пояснення вхідним даним та рішенню, отриманого системою штучного інтелекту. У другому випадку оцінка відображає відповідність пояснення поточним знанням користувача щодо предметної області, тобто встановлює прийнятність пояснення для користувача. В обох випадках пояснення може бути представлено з різним рівнем деталізації. Тому для досягнення поставленої мети необхідно структурувати пояснення за рівнем його деталізації та з урахуванням можливостей його суб'єктивної та об'єктивної оцінки.

Структуризація за рівнями деталізації дає можливість врахувати сприйняття пояснення користувачем з урахуванням різного рівня знань останнього щодо предметної області.

Структуризація за суб'єктивними та об'єктивними характеристиками дає можливість окремо оцінити пояснення за його відповідністю вхідними і вихідним даним системи штучного інтелекту, а також за його відповідністю знанням та потребам користувача.

Перша оцінка може бути представлена в числовому вигляді. Друга оцінка потребує розробки ментальної моделі користувача. У поточному дослідженні ключова увага зосереджена на оцінці об'єктивних характеристик пояснення.

Зв'язок між першим та другим аспектами представлено на рис. 1.

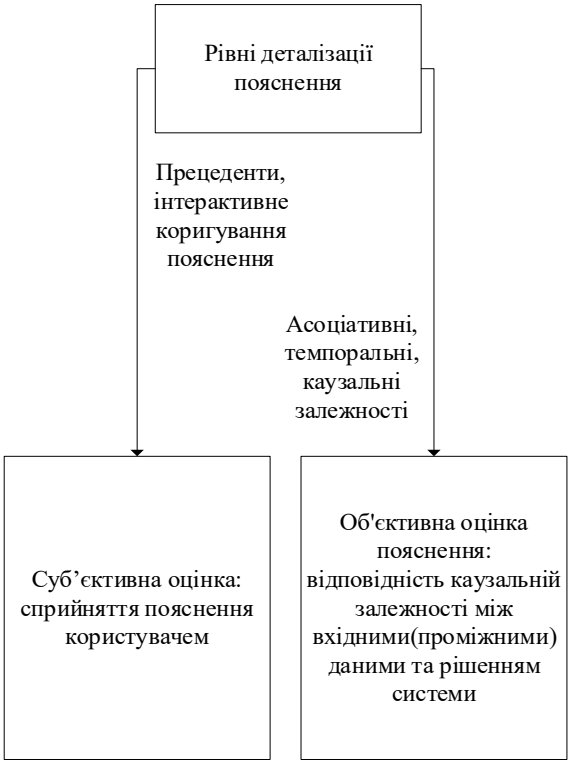


Рис. 2. Оцінка пояснення в залежності від рівня деталізації

Запропонована ієрархія рівнів деталізації пояснення представлена у табл. 1.

Ключова ідея виділення рівнів полягає в тому, щоб відокремити «поверхневі» та «глибинні» знання щодо предметної області.

Перші відображають інтерфейс, зовнішнє представлення роботи інтелектуальної системи.

Другі задають причини наслідкові зв'язки, які обумовлюють процес роботи системи.

Пояснення на першому та другому рівнях деталізації відображають знання з практики застосування рішення інтелектуальної системи. Вони не потребують знання принципів роботи інтелектуальної системи.

Пояснення на 3 та 4 рівнях містять зв'язки, що базуються на фізичних, інформаційних та інших залежностях, які обумовлюють прийняте в систем штучного інтелекту рішення.

Таблиця 1 – Рівні деталізації пояснення

| Рівень | Особливості |
|---|---|
| 1.Асоциативні залежності | 1) Пояснення відповідає на запитання «На що схожі причини прийнятого рішення?»<br>2) Відповідь на вказане запитання представляється у формі діаграм графіків, рамок для фрагментів зображень, тексту, уточнення семантики так, щоб встановити асоціативний зв'язок між властивостями вхідних даних та рішенням інтелектуальної системи. |
| 2. Прецеденти | 1) Пояснення відповідає на запитання «Як на прикладі показати причини отриманого рішення?»<br>2) Відповідь на дане запитання представляється спрощеним зрозумілим прикладом, який однозначно вказує, чому було отримано рішення для поточної категорії вхідних даних.<br>3) Пояснення доповнюється значенням критерія відповідності рішення інтелектуальної системи.<br>4) Низьке значення оцінки відповідності отриманого рішення (наприклад, менше 0.5) свідчить про його непридатність до практичного застосування. У такому випадку пояснення має бути доповнене властивостями вхідних даних, які привели до низької оцінки. |
| 3. Каузальний | 1)Пояснення відповідає на запитання «Якими є безпосередні причини отриманого рішення?»<br>2) Відповідь на дане запитання представляється у вигляді правил продукції, правил логіки, або у текстовій формі. Різні форми представлення правил мають спільну особливість: містять вхідні або проміжні змінні та їх значення, що мають визначальний вплив на отримане рішення.<br>3) Недетерміноване каузальне пояснення доповнюється показником оцінки відповідності причинно-наслідкової залежності. Даний показник може мати ймовірнісний або можливісний характер.<br>4) Низьке значення оцінки відповідності отриманого рішення (наприклад, менше 0.5) свідчить про його непридатність до практичного застосування. |
| 4. Інтерактивний | 1) На даному рівні пояснення уточнюється з використанням знань користувача щодо предметної області. Пояснення відповідає на запитання «Як доповнити пояснення щоб воно стало зрозумілим користувачеві?».<br>2) Уточнення відбувається із урахуванням додаткових вхідних даних, а також знань користувача. |

Ключові відмінності першого та другого рівнів деталізації полягають в такому. Асоціативні залежності відповідають загальним знанням користувача щодо предметної області. Такі залежності можуть

виступати в якості прообразів детермінованих правил виводу рішення. Тому даний рівень може бути оцінений з використанням об'єктивних числових оцінок.

Наприклад, якщо виділені в поясненні ключові елементи зображення містять кінцівки людини у русі, то користувач має зрозуміти, чому система віднесла зображення до класу людей, що біжать.

Рівень прецедентів містить множину асоціативних залежностей, оскільки прецедент представляє собою вхідні та дані типове реалізоване рішення, що є зрозумілим для користувача.

Наприклад, зображення людини, що біжить та результуючий клас цього зображення.

На другому рівні деталізації виконується порівняння прецеденту із знаннями користувача. Тому даний рівень потребує суб'єктивної оцінки на основі опитування користувачів.

Рівень 3 задає детерміновані або оцінювані залежності, що визначають причини рішення або окремих дій процесу формування рішення в системі штучного інтелекту.

Детерміновані залежності для пояснення можуть бути отримані лише у випадку прозорості для користувача моделі прийняття рішень в системі штучного інтелекту. Тобто в даному випадку модель має можливість інтерпретації.

У інших випадках, при представленні моделі прийняття рішень у вигляді «чорного» або «сірого» ящику, каузальні залежності матимуть ймовірний характер. Тому такі залежності мають бути доповнені ймовірнісною або можливісною оцінками. Таким чином, третій рівень деталізації потребує об'єктивної оцінки пояснення.

Четвертий рівень деталізації забезпечує вибір та уточнення пояснення любого з попередніх рівнів. Відповідно, пояснення може бути доповнено асоціативними або каузальними залежностями, що враховують додаткові вхідні змінні. Або ж можуть бути відкинуті відомі з точки зору користувача та тривіальні залежності.

Аналогічно, при використанні прецедентного рівня деталізації пояснення може бути скоригований перелік базових прецедентів. Задача уточнення прецедентів виникає, наприклад при зміні соціальних, юридичних правил, вподобань користувача інтелектуальної системи, тощо.

Наприклад, при зміні демографічних характеристик користувача (освіти, місце проживання, тощо) можуть змінитися і його вподобання щодо товарів та послуг в систем електронної комерції. Відповідна рекомендаційна підсистема потребує налагодження, що може бути вирішено на основі уточнення пояснення щодо рекомендованого переліку товарів та послуг.

Метод можливісного оцінювання пояснень

Метод оцінювання пояснень призначений для перевірки відповідності пояснень парам (набір вхідних даних, рішення інтелектуальної системи) або (набір проміжних даних, проміжний стана інтелектуальної системи). Метод орієнтований на перевірку тлумачень на етапі побудови пояснення (рис. 1).

Метод використовує такі показники оцінки пояснень: чутливість пояснень до властивостей вхідних даних; коректність пояснень; складність формування пояснень.

Перший показник відображає оцінку відмінностей у рішеннях системи при незначних змінах у вхідних даних. Оцінка чутливості виконується шляхом установлення відмінностей між елементами рішення та вхідними даними за умови, що пояснення є схожими або однаковими. Тобто якщо пари (вхідні дані, рішення) відрізняються несуттєво, то і пояснення має бути схожим або ідентичним. У випадку суттєвих відмінностей в даних пояснення також має відрізнятись [12]. Фактично даний показник впливає на точність пояснення: при невисокій чутливості будуть сформовані однакові пояснення для різних пар (вхідні дані, результат), що знижує точність тлумачення. При високій чутливості пояснення має відрізнятись для несуттєвих змін у вхідних даних, що ускладнює сприйняття пояснення користувачем. Таким чином, оцінка чутливості пояснення є підґрунтям для визначення ефективності використання тлумачення користувачем.

Показник коректності є бінарною оцінкою, яка відповідає на питання: «Чи пояснення відображає процес прийняття рішення в інтелектуальній системі?», «Чи пояснення відображає асоціативну або каузальну залежність між вхідними даними та рішенням інтелектуальної системи?». Перше питання є актуальним при представленні системи у вигляді «чорного» або «сірого» ящику, а друге – якщо система представляється як «чорний» ящик. Даний показник визначає можливість використання пояснення при застосуванні системи [10].

Показник складності відображає кількість вхідних даних, які необхідні для побудови пояснення. Побудова пояснень пов'язана із створенням моделі процесу тлумачення, яка є менш складною порівняно із моделлю прийняття рішення. Тому, на відміну від двох попередніх, цей показник визначає можливість реалізації пояснень в інтелектуальній системі.

Зв'язок між показниками відображено на рис. 3.

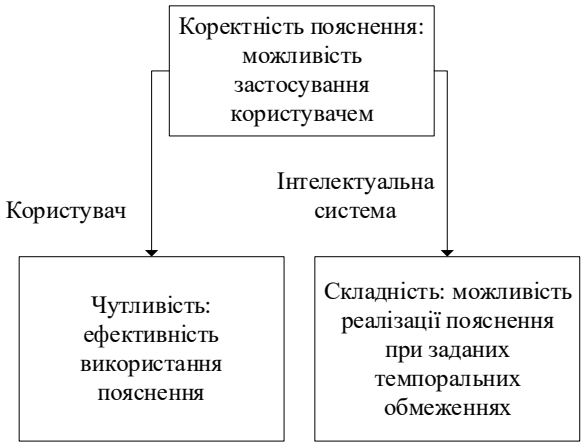


Рис. 3. Зв'язок між показниками оцінки пояснення

Метод оцінювання пояснення містить такі ключові етапи.

Етап 1. Перевірка коректності пояснення.

Результати даного етапу дають можливість установити відповідність пояснення процесу прийняття рішення в інтелектуальній системі.

Етап 2. Оцінка складності пояснення.

Результати даного етапу дають можливість з'ясувати, чи пояснення може бути реалізовано та використано.

Етап 3. Оцінка чутливості пояснення.

Результати даного етапу дають можливість визначити вплив відхилень у значеннях вхідних даних на пояснення.

Етап 4. Зменшення складності при обмеженні по чутливості пояснення.

Реалізація даного етапу дає можливість зменшити вплив неключових вхідних даних на пояснення.

На першому етапі методу розраховується коректність пояснення. В рамках можливісного підходу коректність визначається через необхідність вибору отриманого рішення на заданому наборі вхідних даних. Можливість $\Pi(R_i)$ рішення $R_i$ визначається через найбільшу ймовірність використання $R_i$ в минулому на множині схожих рішень $R^* = \{R_j : (\forall j) R_i \neq R_j\}$. Наприклад, можливість рішення щодо вибору комп'ютера з певним процесором в системі електронної комерції визначається через найбільшу ймовірність вибору комп'ютера з процесором цієї ж фірми такого ж покоління.

Необхідність рішення $N(R_i)$ задається через можливість вибору всіх інших рішень, тобто $1-\Pi(R \setminus R_i)$. В наведеному прикладі необхідність задається через можливість вибору процесорів всіх інших фірм, незалежно від їх покоління. Тобто можливість визначається для цільової підмножини вхідних даних, а необхідність – для повної множини за виключенням цільової підмножини.

Формально, пояснення $Expl(R_i)$ рішення $R_i$ для вхідних даних $X_i$ є коректним, якщо значення необхідності перевищує 0.5 [10]:

$$C(Expl(R_i)) = \text{true} \left| 1-\Pi(R \setminus R_i) > 0.5. \right. \quad (1)$$

Семантика виразу (1) полягає в тому, що для коректного пояснення $Expl(R_i)$ щодо цільового рішення $R_i$ можливість вибору альтернативних рішень $R_j \neq R_i$ на підмножині $X \setminus X_i$ має становити менше, ніж 0.5. Можливість використання даного представлення коректності пояснення базується на тому, що вхідні змінні $X_{i,k}$ мають безпосередній зв'язок з результатом. Наприклад, в якості вхідних даних використовується модель процесора, а вихідних – комп'ютер з процесором даної моделі.

На етапі 2 оцінюється складність імплементації та використання пояснення. Складність пояснення $Expl(R_i)$ оцінюється як кількість вхідних змінних $|X_i|$ для даного пояснення.

На етапі 3 виконується оцінка чутливості пояснення. Чутливість задається через відхилення співвідношень вхідних даних та рішення за умови схожості пояснень [12]. У можливісному аспекті чутливість $S$ визначається через відмінність можливостей використання різних вхідних даних для отримання одного й того ж пояснення для однакового результату:

$$\begin{aligned} &S(Expl) = \max\left(\left|\Pi(X_i) - \Pi(X_j)\right|\right) \\ &\left| Expl_i = Expl_j, R_i = R_j. \right. \end{aligned} \quad (2)$$

Обмеження $R_i = R_j$ ускладнює оцінку чутливості пояснення, оскільки для схожих результатів може бути надано одне й те ж пояснення. Наприклад, в системі електронної комерції ми можемо рекомендувати різні моделі ноутбуків з одним і тим же поясненням: модель пропонується на основі властивостей процесора.

Тому у випадку однакового пояснення для схожих результатів потрібно враховувати відповідність між можливостями вхідних даних та рішення системи:

$$S(Expl) = \max\left(\left|\frac{\Pi(X_i)}{\Pi(R_i)} - \frac{\Pi(X_j)}{\Pi(R_j)}\right|\right) \left| Expl_i = Expl_j. \right. \quad (3)$$

В основі можливісної оцінки (3) лежить ідея про те, що чутливість відображає максимальне відхилення між можливостями впливу вхідних даних на рішення системи штучного інтелекту.

На етапі 4 мінімізується кількість вхідних даних, що використовується для побудови пояснення. В основі такої оцінки лежить ідея про те, що необхідно виділити лише суттєві для використання пояснення вхідні дані. Суттєвими є дані, які впливають на чутливість пояснення. Ці данні потрібно враховувати, оскільки їх вилучення може привести до зниження точності пояснення, тобто до невідповідності пояснення реальному процесу прийняття рішення. Відповідно, якщо частина даних не впливає на чутливість пояснення, то вони можуть бути виключені з останнього.

Постановка задачі мінімізації складності має вигляд:

$$\begin{aligned} &\textit{Дано}: \\ &C(Expl(R_i)) = \text{true}. \\ &\\ &\textit{Необхідно}: \\ &|X_i| \to \min. \\ &\\ &\textit{За умови}: \\ &(\forall i) S(Expl(R_i)) = S(Expl). \end{aligned} \quad (4)$$

На даному етапі циклічно на кожному n-кроці виконується вилучення складових $X_{i,k}$ із поточної множини даних $X_i^{(n)}$ для пояснення за умови, що

$S\left(Expl\left(R_i\right)\right) = S\left(Expl\right)$, тобто чутливість пояснення не зменшується:

$$\left|X_i^{(n+1)}\right| = \left|X_i^{(n)} \setminus X_{i,k}\right| \left|S\left(Expl\left(R_i\right)\right) = S\left(Expl\right). \quad (5)$$

Використання методу дає можливість зменшити кількість вхідних даних для пояснення, представленого асоціативною або каузальною залежністю згідно деталізації пояснень у табл. 1.

Розглянемо приклад формування вхідних даних для запропонованого методу на основі аналізу журналу продажів у системі електронної комерції. Вхідними даними є пояснення у формі залежності

*ключові комплектуючі комп'ютера* $\rightarrow$
*пояснення щодо вибору комп'ютера* .

Ці дані необхідно доповнити значеннями $\Pi\left(X_i\right)$, $\Pi\left(R_i\right)$, $\left|X_i\right|$. У журналі продажів міститься інформація щодо проданих комп'ютерів різних типів. Інформація про продажі дає можливість побудувати множини ймовірностей вибору цих комп'ютерів. На основі обрахованих ймовірностей формується можливість результату $\Pi\left(R_i\right)$, тобто можливість покупки з проданих комп'ютерів, наприклад із процесорами з однієї серії. Для кожного з комп'ютерів відомі ключові комплектуючі, характеристики яких впливають на вибір користувача. Це дає можливість розрахувати ймовірності їх використання в куплених продуктах і відповідні можливості $\Pi\left(X_i\right)$. Наприклад, розраховується можливість використання процесорів певної серії з урахуванням ймовірності їх використання в комп'ютерах різних торгових марок. Значення $\left|X_i\right|$ розраховується для відповідних підмножин комплектуючих, наприклад, для серії процесорів, один із яких є у рекомендованому комп'ютері.

Експериментальна перевірка методу була виконана на основі журналу продажів святкових товарів мережі супермаркетів. Пояснення мало вигляд: призначення групи товарів, час покупки, ключові слова як асоціативні причини вибору. За результатами експериментальної перевірки встановлено, що пояснення на основі групи товарів має показник коректності $C = 0.63$ лише на окремих інтервалах часу, зазвичай в межах тижнів від дати свят. Тобто мінімальний набір елементів, який забезпечував незмінну чутливість пояснення становить 2. Значення чутливості становило $S = 0.44$.

**Висновки.** Виконано структуризацію оцінки пояснень в залежності від рівня їх деталізації. Виділено рівні асоціативних залежностей, прецедентів, каузальних залежностей та інтерактивний, що визначають різний ступінь деталізації пояснень. Показано, що асоціативний та каузальний рівні деталізації пояснень можуть бути оцінені з використанням числових, ймовірнісних або можливісних показників. Прецедентний та інтерактивний рівні потребують суб'єктивної оцінки на основі опитування користувачів системи штучного інтелекту.

Запропоновано метод можливісного оцінювання відповідності пояснень процесу прийняття рішень в інтелектуальній системі з урахуванням залежностей між вхідними даними та рішенням інтелектуальної системи. Метод містить етапи оцінювання чутливості, коректності та складності пояснення на основі порівняння значень та кількості використаних у поясненні вхідних даних. Метод дає можливість комплексно оцінити пояснення як у аспектах стійкості до несуттєвих змін у вхідних даних, відповідності пояснення отриманому результату, а також складності обчислення пояснення.

У аспекті практичного застосування метод дає можливість підвищити ефективність формування пояснення шляхом відбору підмножини ключових вхідних змінних, які мають суттєвий вплив на отримане в інтелектуальній системі рішення.

**Список використаної літератури**


1. Engelbrecht Andries P. *Computational Intelligence: An Introduction*. NJ: John Wiley & Sons, 2007. 632 p.
2. Alonso J. M., Castiello C., Mencar C. A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field. In: Medina, J., et al. Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations. *IPMU. Communications in Computer and Information Science*. 2018. Vol 853. P. 3–15.
3. Gunning D., Aha D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*. 2019. Vol. 40 (2). P. 44-58.
4. 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.
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. Camburu O. M, Giunchiglia E., Foerster J., Lukasiewicz T., Blunsom P. Can I trust the explainer? Verifying post-hoc explanatory methods. 2019. *arXiv*:1910.02065.
8. Gunning D., Vorm E., Wang J., Turek M., "Darpa's Explainableai (XAI) Program: a Retrospective", *Applied AI Letters*. 2021. Vol. 2, no. 4. DOI: https://doi.org/10.1002/ail2.61.
9. Chalyi S., Leshchynskyi V. Temporal-oriented model of causal relationship for constructing explanations for decision-making process. *Advanced Information Systems.* 2022. № 6 (3). P. 60–65.
10. Chalyi S., Leshchynskyi V. Possible evaluation of the correctness of explanations to the end user in an artificial intelligence system. *A.I.S.* 2023. № 7. P. 75–79.
11. Chalyi S, Leshchynskyi V. Probabilistic counterfactual causal model for a single input variable in explainability task. *Advanced Information Systems.* 2022. №7(3), P. 54–59. DOI: https://doi.org/10.20998/2522-9052.2023.3.08.
12. Чалий С. Ф., Лещинський В. О. Оцінка чутливості пояснень в інтелектуальній інформаційній системі. *Системи управління, навігації та зв'язку. Збірник наукових праць.* 2023. № 2. С. 165–169.


**References (transliterated)**


1. Engelbrecht Andries P. *Computational Intelligence: An Introduction*. NJ: John Wiley & Sons, 2007. 632 p.
2. Alonso J.M., Castiello C., Mencar C. A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field. In: Medina, J., et al. Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations. *IPMU. Communications in Computer and Information Science.* 2018, vol. 853, pp. 3–15.

3. Gunning D., Aha D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*. 2019, vol. 40 (2), pp. 44–58.
4. 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.
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, pp. 1–38.
7. Camburu O.M, Giunchiglia E., Foerster J., Lukasiewicz T., Blunsom P. Can I trust the explainer? Verifying post-hoc explanatory methods. 2019. *arXiv*:1910.02065.
8. Gunning D., Vorm E., Wang J., Turek M. Darpa's Explainableai (XAI). Program: a Retrospective. *Applied AI Letters*. 2021, vol. 2, no. 4. DOI: https://doi.org/10.1002/ail2.61.
9. Chalyi S., Leshchynskyi V. Temporal-oriented model of causal relationship for constructing explanations for decision-making process. *Advanced Information Systems.* 2022, no. 6 (3), pp. 60–65.
10. Chalyi S., Leshchynskyi V. Possible evaluation of the correctness of explanations to the end user in an artificial intelligence system. *A.I.S.* 2023, no. 7, pp.75–79.
11. Chalyi S, Leshchynskyi V. Probabilistic counterfactual causal model for a single input variable in explainability task. *Advanced Information Systems.* 2022, no. 7 (3), pp. 54–59. DOI: https://doi.org/10.20998/2522-9052.2023.3.08.
12. Chalyi S. Leshchynskyi V. Otsinka chutlyvosti poiasnen v intelektualnii informatsiinii systemi [Evaluation of the sensitivity of explanations in the intelligent information system]. *Systemy upravlinnia, navihatsii ta zviazku. Zbirnyk naukovykh prats* [Control, navigation and communication systems. Collection of scientific papers]. 2023, no. 2, pp. 165–169.

*Надійшла (received) 05.10.2023*

UDC 004.8:004.9

***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

***V. O. LESHCHYNSKYI***, Candidate of Technical Sciences (PhD), Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor at the Department of Software Engineering, Kharkiv; e–mail: volodymyr.leshchynskyi@nure.ua, ORCID: https://orcid.org/0000-0002-8690-5702

## A METHOD FOR EVALUATING EXPLANATIONS IN AN ARTIFICIAL INTELLIGENCE SYSTEM USING POSSIBILITY THEORY

The subject of the research is the process of generating explanations for the decision of an artificial intelligence system. Explanations are used to help the user understand the process of reaching the result and to be able to use an intelligent information system more effectively to make practical decisions for him or her. The purpose of this paper is to develop a method for evaluating explanations taking into account differences in input data and the corresponding decision of an artificial intelligence system. The solution of this problem makes it possible to evaluate the relevance of the explanation for the internal decision-making mechanism in an intelligent information system, regardless of the user's level of knowledge about the peculiarities of making and using such a decision. To achieve this goal, the following tasks are solved: structuring the evaluation of explanations depending on their level of detail, taking into account their compliance with the decision-making process in an intelligent system and the level of perception of the user of such a system; developing a method for evaluating explanations based on their compliance with the decision-making process in an intelligent system. Conclusions. The article structures the evaluation of explanations according to their level of detail. The levels of associative dependencies, precedents, causal dependencies and interactive dependencies are identified, which determine different levels of detail of explanations. It is shown that the associative and causal levels of detail of explanations can be assessed using numerical, probabilistic, or possibilistic indicators. The precedent and interactive levels require a subjective assessment based on a survey of users of the artificial intelligence system. The article develops a method for the possible assessment of the relevance of explanations for the decision-making process in an intelligent system, taking into account the dependencies between the input data and the decision of the intelligent system. The method includes the stages of assessing the sensitivity, correctness and complexity of the explanation based on a comparison of the values and quantity of the input data used in the explanation. The method makes it possible to comprehensively evaluate the explanation in terms of resistance to insignificant changes in the input data, relevance of the explanation to the result obtained, and complexity of the explanation calculation. In terms of practical application, the method makes it possible to minimize the number of input variables for the explanation while satisfying the sensitivity constraint of the explanation, which creates conditions for more efficient formation of the interpretation based on the use of a subset of key input variables that have a significant impact on the decision obtained by the intelligent system.

**Keywords:** explanation, evaluation of explanation, artificial intelligence system, intelligent system, comprehensible artificial intelligence, associative dependence, causal dependence, precedent, information system, recommendation system.

*Повні імена авторів / Author's full names*

**Автор 1 / Author 1:** Чалий Сергій Федорович**,** Chalyi Serhii Fedorovych
**Автор 2 / Author 2:** Лещинський Володимир Олександрович, Leshchynskyi Volodymyr Oleksandrovich