DOI: 10.20998/2079-0023.2026.01.13
УДК 004.8:004.9

І. О. ЛЕЩИНСЬКА, кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, доцент кафедри програмної інженерії, м. Харків, Україна; ORCID: https://orcid.org/0000-0002-8737-4595; e-mail: iryna.leshchynska@nure.ua

МЕТОД ВЕРИФІКАЦІЇ ЗБАЛАНСОВАНОСТІ МЕНТАЛЬНИХ МОДЕЛЕЙ РІШЕННЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ

Предметом дослідження є процес верифікації ментальних моделей рішення інтелектуальної системи. Метою роботи є розробка підходу до оцінювання збалансованості пояснень рішень інтелектуальних систем щодо їх негативних та позитивних аспектів. Відповідно до цієї мети вирішуються такі основні задачі: розробити підхід до оцінювання збалансованості пояснень рішень інтелектуальних систем на основі пропорційності подання негативних і позитивних характеристик у поясненні; розробити загальний метод верифікації збалансованості ментальних моделей рішення інтелектуальної системи, який враховує структурне й вагове покриття множини негативних аспектів ментальними моделями рішення; провести експериментальну перевірку запропонованого методу на прикладі набору користувацьких відгуків про роботу рекомендаційної системи, що містять інформацію ментальні моделі запропонованих рішень. Запропоновано підхід до оцінювання збалансованості пояснень рішень інтелектуальних систем з урахуванням негативних аспектів, який передбачає формування еталонної зваженої множини суттєвих негативних аспектів рішення, вилучення негативних елементів пояснення та розрахунок показників структурного і вагового покриття, а також оцінки пропорційності подання негативної інформації. Такий підхід дозволяє отримати кількісну оцінку, наскільки пояснення відображає обмеження та можливі негативні наслідки застосування рішення у співвідношенні до позитивних його властивостей. Запропоновано метод верифікації збалансованості ментальних моделей рішення інтелектуальної системи. Метод включає етапи формування еталонної множини негативних аспектів рішення на основі аналізу користувацьких відгуків, вилучення негативної компоненти ментальних моделей рішення, обчислення показників структурного та вагового покриття в ментальних моделях, оцінювання пропорційності й релевантності подання негативних аспектів, формування інтегральної оцінки щодо збалансованості ментальних моделей. Метод забезпечує можливість уточнення ментальних моделей з урахуванням недоліків практичного застосування рішення користувача внаслідок неповноти моделей. Експериментальна перевірка методу на основі набору користувацьких відгуків показала, що використання відгуків як джерела інформації щодо ментальних моделей користувачів дає можливість сформувати еталонну множину негативних аспектів рішення інтелектуальної системи, яка відображає проблеми використання та ризики, важливі для користувачів.

Ключові слова: пояснення; система штучного інтелекту; зрозумілий штучний інтелект; залежності; ментальна модель; верифікація.

Вступ. Побудова пояснюваних інтелектуальних систем, які можуть надавати користувачам інтерпретовані тлумачення своїх рішень, є одним із актуальних напрямів досліджень у галузі штучного інтелекту [1]. При використанні цих рішень на практиці користувачі мають розуміти і враховувати як їх переваги, так і ризики, а також обмеження щодо можливостей застосування [2]. Однак значна частина існуючих підходів до побудови пояснень в інтелектуальних системах (ІС) орієнтована переважно на обґрунтування позитивних аспектів рекомендацій, тоді як негативні властивості рішень часто представлені фрагментарно або не відображаються у процесі взаємодії з користувачем [3]. Дане обмеження призводить до формування незбалансованих ментальних моделей рішень у користувачів, що, в свою чергу, знижує довіру до ІС та ускладнює практичне застосування отриманих результатів [4].

Ментальна модель рішення ІС представляє собою внутрішнє представлення сукупності причинно-наслідкових зв’язків між властивостями рішення, а також сценаріїв використання та обмежень, на основі якого формується пояснення щодо його корисності для користувача [4]. Якщо ментальна модель не містить або не повністю враховує суттєві негативні аспекти рішення ІС, то навіть обґрунтоване для користувача рішення може бути сприйняте як упереджене [5]. Тому важливим завданням в рамках пояснювального штучного інтелекту є розробка методів верифікації збалансованості ментальних моделей рішень інтелектуальних систем з урахуванням відображення негативних аспектів отриманого результату [6].

Негативні аспекти рішень мають прояви не лише на рівні внутрішніх ментальних моделей, але й, за результатами досвіду використання цих рішень користувачами, проявляються як відгуки, скарги, коментарі, а також позитивні й негативні оцінки [7]. Ці джерела інформації містять неформалізовані, фрагменти ментальних моделей користувачів, які відображають проблемні аспекти рішень, а також користувацьку інтерпретацію поведінки інтелектуальної системи [7]. Інтеграція результатів такого емпіричного зворотного зв’язку із формальними моделями рішень відкриває можливість верифікації відповідності ментальних моделей та реального досвіду користувачів. Невідповідність моделей і досвіду практичного застосування рішень свідчить про суттєві обмеження у процесі побудови пояснень в інтелектуальній системі, а також незбалансованість пояснення з урахування негативних аспектів рішення ІС [8].

Таким чином, перевірка збалансованості ментальних моделей рішення ІС має базуватись на формуванні базової множини суттєвих негативних аспектів рішення на основі, наприклад, аналізу корпусу користувацьких відгуків, з подальшою верифікацією ментальних моделей щодо покриття та пропорційності позитивних та негативних властивостей рішення [6, 8]. Таке співставлення дає можливість встановити можливу перевагу позитивних аспектів рішення у поточній ментальній моделі й невідповідність негативних аспектів тому представленню, яке формують користувачі у своїх ментальних моделях [8]. Як наслідок, можуть

бути виявлені і потім уточнені «сліпі зони» у ментальних моделях рішень інтелектуальних систем [6].

Таким чином, верифікація збалансованості ментальних моделей рішення ІС створює умови для адекватного використання цих рішень при вирішенні практичних задач користувачів.

Аналіз останніх досліджень і публікацій

Дослідження у сфері пояснюваного штучного інтелекту (XAI) приділяють значну увагу розробці методів локальної та глобальної інтерпретації моделей, таких як LIME, SHAP, а також атрибутивних підходів. Сукупність цих підходів дає можливість виявити найбільш важливі властивості або компоненти вхідних даних [9, 10], що підвищує прозорість моделей. Проте ці методи призначені для пояснення того, чому рішення ІС можуть бути застосовані на практиці. Оцінюванню того, наскільки повно відображені потенційні ризики, обмеження або негативні наслідки рішення ІС у вказаних методах і підходах не приділяється достатньо уваги [3, 10]. Тому поняття збалансованості пояснення зазвичай розглядається як неявне, що обмежує розробку критеріїв та процедур верифікації [6].

Дослідження щодо моделювання та аналізу ментальних моделей користувачів у системах підтримки прийняття рішень орієнтовані в першу чергу на виявлення способів формування внутрішнього представлення про роботу системи, її мету, можливі обмеження та очікувану поведінку [11, 12]. У таких дослідженнях ментальна модель розглядається у аспекті довіри з позиції прийняття або відхилення рішень системи, а також як об'єкт адаптації зі сторони системи, зокрема з використанням персоналізованих пояснень [11]. Проте питанням верифікації збалансованості ментальних моделей не приділяється достатньо уваги [6].

В рамках побудови пояснень в ІС суттєвого розвитку набули нейросимвольні підходи. Особливість таких підходів полягає у поєднанні нейронних мереж із символьним представленнями знань у вигляді логічних правил, онтологій або ж графових структур та сценарних моделей. Таке поєднання створює умови для підвищення прозорості роботи інтелектуальних систем [13, 14]. Водночас навіть у межах нейросимвольних систем основна увага часто зосереджується на правильності та повноті знань, узгодженості логічних правил чи відповідності онтологічних зв'язків, тоді як досягненню балансу між позитивними та негативними аспектами рішень не приділяється достатньо уваги [6, 14].

Використання відгуків користувачів, оцінок та коментарів як джерела знань про реальний досвід взаємодії з інтелектуальними системами розглядається в [7, 15]. Аналіз тональності та тематики відгуків дає можливість виявити позитивні й негативні сторони продуктів та сервісів, що дозволяє формувати профілі якості та задоволеності користувачів [15]. Однак такі емпіричні дані зазвичай не використовуються для виявлення відмінностей між внутрішніми моделями системи та ментальними моделями користувачів [6, 8, 15].

Таким чином, аналіз останніх досліджень і публікацій свідчить, що проблема верифікації збалансованості ментальних моделей рішень інтелектуальних систем з урахуванням негативних аспектів на основі аналізу даних з користувацьких відгуків потребує свого вирішення. Це обумовлює актуальність даного дослідження.

Мета та задачі дослідження. Метою роботи є розробка підходу до оцінювання збалансованості пояснень рішень інтелектуальних систем щодо їх негативних та позитивних аспектів.

Відповідно до цієї мети вирішуються такі основні задачі:

– розробити підхід до оцінювання збалансованості пояснень рішень інтелектуальних систем на основі пропорційності подання негативних і позитивних характеристик у поясненні;

– розробити загальний метод верифікації збалансованості ментальних моделей рішення інтелектуальної системи, який враховує структурне й вагове покриття множини негативних аспектів ментальними моделями рішення;

– провести експериментальну перевірку запропонованого методу на прикладі набору користувацьких відгуків про роботу рекомендаційної системи, що містять інформацію ментальні моделі запропонованих рішень.

Підхід до оцінювання збалансованості пояснень рішень інтелектуальних систем.

Підхід до оцінювання збалансованості пояснень рішень інтелектуальних систем щодо негативних аспектів визначає критерії та процедуру оцінювання для одного пояснення. Підхід включає чотири кроки: формування еталонної множини негативних аспектів рішення; вилучення негативних аспектів рішення із пояснення; обчислення показників покриття та пропорційності; формування інтегральної оцінка збалансованості пояснення.

На першому кроці формується множина суттєвих негативних аспектів рішення, яка використовується для оцінювання конкретних пояснень. Джерелами для побудови цієї множини є набір даних про досвід використання рішення користувачами, наприклад, текстові відгуки. Із даного набору виділяються суттєві (зазвичай повторювані) негативні фактори, які впливають на експлуатацію рішення. Для кожного негативного аспекту задається нормована вага, яка відображає його значущість для застосування рішення. Зокрема, суттєві збої мають більшу вагу, ніж недоліки інтерфейсу. Наприклад, для телефона такими негативними аспектами можуть бути реальна ємність та автономність акумуляторної батареї, відсутність гарантії, перегрів корпусу, дискомфорт від екрана, проблемами з авторизацією та конфіденційністю.

Другий крок розробленого підходу пов'язаний із аналізом конкретного пояснення щодо рішення ІС. Із тексту пояснення виділяється множина негативних елементів, які в явній або неявній формі описують недоліки, небажані наслідки застосування або обмеження даного рішення. При вилученні ознак фактично формується відображення пояснення як упорядкованої множини властивостей рішення на множину негативних аспектів. Кожен елемент пояснення, який містить опис певної проблеми, відображається на один чи кілька елементів еталонної множини негативних

аспекті. Якщо пояснення містить конкретний опис, наприклад, згадку про швидкий розряд батареї, перегрів або можливі помилки авторизації, то відповідні аспекти вважаються покритими поясненням. В іншому випадку відсутні у поясненні суттєві негативні аспекти фіксую

На третьому кроці виконується в обчислення показників структурного і вагового покриття, які відображають, наскільки повно й адекватно пояснення покриває еталонну множину негативних аспектів.

Коефіцієнт структурного покриття $C_{\text{struct}}^{\text{neg}}$ характеризує частку L^{neg} еталонних негативних аспектів L^{etalon}, представлених у поясненні:

$$C_{\text{struct}}^{\text{neg}} = \frac{\left|L^{\text{neg}}\right|}{\left|L^{\text{etalon}}\right|}. \quad (1)$$

Значення $C_{\text{struct}}^{\text{neg}}$, близькі до 1, свідчать про те, що пояснення містить більшість суттєвих негативних аспектів. Низькі значення вказують про замовчування частини проблем, тобто про структурну незбалансованість пояснення.

Коефіцієнт вагового покриття відображає частку сумарної ваги негативних аспектів l_i, яку перекриває пояснення:

$$C_{\text{weight}}^{\text{neg}} = \frac{\sum_{l_i \in L^{\text{neg}}} w(l_i)}{\sum_{l_i \in L^{\text{etalon}}} w(l_i)}, \quad (2)$$

де $w(l_i)$ – вага негативного аспекту.

Даний показник враховує в першу чергу найбільш значущі аспекти з високою вагою.

Збалансованості пояснення крім показників (1) та (2) важливими є значення пропорційності та релевантності подання негативних аспектів.

Пропорційність означає, що негативні аспекти не лише згадуються, але й займають у поясненні таку частку змісту, яка відповідає їхній значущості порівняно з позитивними властивостями рішення. Тобто якщо важливий негативний аспект згаданий коротко на фоні великого позитивного опису, то це свідчить про непропорційне відображення даного аспекту.

Релевантність означає, що формулювання негативних аспектів повинні відповідати рівню підготовки та інформаційним потребам користувачів. Тобто складні аспекти рішення мають бути представлені у спрощеній формі для кінцевого користувача, або детально для експерта у предметній галузі.

Пропорційність і релевантність оцінюються за експертними шкалами або на основі таких критеріїв, як:

– відсоток тексту пояснення, присвячений негативним аспектам;

– відсоток тексту пояснення для користувачів з відповідним рівнем знань;

– використання/відсутність відповідної термінології.

Результати попередніх етапів об'єднуються у інтегральну оцінку збалансованості пояснення щодо негативних аспектів. Така оцінка може бути представлена згорткою коефіцієнта структурного та вагового покриття, а також пропорційності й релевантності, або ж визначенням порогових значень цих показників.

Метод верифікації збалансованості ментальних моделей рішення інтелектуальної системи.

Метод призначений для перевірки узгодженості ментальних моделей з еталонною множиною негативних аспектів рішення.

Метод містить такі етапи.

Етап 1. Формування еталонної множини негативних аспектів рішення.

Результуюча множина L^{etalon} має вигляд:

$$L^{\text{etalon}} = \left\{\left(l_i, w(l_i)\right) : i = \overline{1, I}\right\}. \quad (3)$$

Етап 2. Виділення негативної компоненти з ментальних моделей рішення.

Інтелектуальна система може мати кілька ментальних моделей $Mm^{(n)}$ для одного рішення, в залежності від рівня знань користувача та різних варіантів практичного використання.

Для кожної ментальної моделі проводиться:

– структурний аналіз з метою ідентифікувати елементи, що відповідають негативним властивостям рішення;

– зіставлення отриманих елементів з негативними аспектами з L^{etalon}.

У результаті для кожної моделі $Mm^{(n)}$ формується підмножина $L^{Mm^{(n)}} \subseteq L^{\text{etalon}}$, яка містить ті негативні аспекти з L^{etalon}, що присутні у відповідній ментальній моделі рішення

Етап 3. Обчислення показників покриття негативних аспектів ментальними моделями.

На третьому етапі для кожної ментальної моделі обчислюються коефіцієнт структурного покриття та коефіцієнт вагового покриття.

Етап 4. Оцінка пропорційності подання негативних аспектів у ментальних моделях.

Навіть за високих значень структурного та вагового покриття ментальна модель може бути незбалансованою у випадку негативні аспекти представлені у мінімальному вигляді на фоні детальної позитивної частини моделі.

Тому для кожної ментальної моделі рішення визначається:

– відсоток структурних елементів, пов'язаних із негативними аспектами;

– співвідношення між деталізацією негативних та позитивних аспектів, зокрема враховується чи мають негативні аспекти однаковий рівень формалізації у порівнянні з позитивними аспектами (умови та сценарії використання тощо);

– наявність обмежень щодо негативних аспектів, наприклад, захисних дій, попереджень тощо.

Етап 5. Оцінка релевантності негативних аспектів до профілю користувачів.

Ментальна модель на даному етапі оцінюється з позиції, наскільки її негативна компонента відповідає профілю цільових користувачів, для яких призначені пояснення.

Для кожного рівня ментальної моделі (рівня підготовки користувача) аналізується:

– подано негативні аспекти у зрозумілій для відповідної аудиторії формі, чи навпаки форма не відповідає рівню знань користувача;

– представлено складні або критичні ризики застосування рішення, чи навпаки дані ризики закамуфльовані довгими технічними описами.

Результатом етапу є оцінка релевантності представлення негативних аспектів у кожній ментальній моделі відповідно до профілю користувача. Низьке значення релевантності може свідчити про важливість адаптації ментальної моделі, зокрема її спрощення для користувачів з низьким рівнем підготовки.

Етап 6. Формування інтегральної оцінки збалансованості ментальних моделей.

Виконується згортка результатів попередніх етапів у єдиний інтегральний показник.

Додатково може виконуватися інтеграція показників для всіх ментальних моделей $Mm^{(n)}$ з метою оцінки загальної збалансованості ментального представлення користувачів щодо рішення інтелектуальної системи.

Експериментальна перевірка розробленого методу.

Перевірка виконана з використанням множини текстових відгуків про рекомендовану в системі електронної комерції модель смартфона. Відгуки розміщені покупцями на сайті системи електронної комерції і містять як позитивні характеристики, так і опис типових проблем та негативних вражень. Така множина може бути інтерпретована як сукупність фрагментів індивідуальних ментальних моделей користувачів, які відображають їхнє суб'єктивне сприйняття властивостей продукту та якості пов'язаного з цим продуктом сервісу.

На першому етапі експерименту набір відгуків було проаналізовано з метою виділення еталонної множини ключових негативних аспектів продукту та сервісу. На основі узагальнення критичних відгуків було сформовано еталонну множину негативних аспектів з десяти елементів (корпоративне блокування, фактична ємність батареї, швидкий розряд, проблеми з гарантією, перегрів, дискомфорт від екрана/ШІМ, глюки/повільна робота, проблеми з авторизацією, конфіденційність, неякісні аксесуари). Кожному аспекту було призначено вагу у діапазоні від 0,5 до 0,99. Вага відображає відносну важливість недоліка для користувача. Наприклад, корпоративне блокування та критичні проблеми з батареєю оцінювалися як більш значущі, ніж недоліки плівки.

На другому етапі було змодельовано три варіанти ментальної моделі рішення системи, які по-різному відображають еталонну множину негативних аспектів:

– модель М1, збалансована ментальна модель, що включає всі 10 негативних елементів потенційно здатна враховувати весь спектр негативних аспектів, відображених у користувацьких відгуках;

– модель М2, частково збалансована модель, яка охоплює приблизно половину еталонних негативних аспектів і включає найкритичніші властивості (корпоративне блокування, проблеми з ємністю батареї та автономністю, питання гарантії, перегрів);

– модель М3, незбалансована модель, яка враховує лише один із еталонних негативних аспектів (швидкий розряд батареї) та ігнорує інші суттєві проблеми, описані користувачами.

На третьому етапі для трьох ментальних моделей М1, М2, М3 було розраховано показники структурного покриття та вагової відповідності. Результати етапу наведено в табл. 1. Як видно з табл. 1, модель М1 покриває всі негативні аспекти з відгуків як за кількістю, так і за сумарною вагою. Модель М2 демонструє покриття близько 50 % аспектів за кількістю та приблизно 56 % за сумарною вагою цих аспектів, тобто враховує головні і ігнорує другорядні недоліки. Модель М3 має низькі показники як за структурним, так і за ваговим покриттям (близько 11–13 %), що відображає суттєвий розрив між ментальною моделлю рішення системи та сприйняттям цього рішення користувачами.

Розроблені підхід та метод забезпечують можливість автоматизації процесу верифікації ментальних моделей рішень рекомендаційних систем на основі зворотного зв'язку від користувачів.

Якщо ментальна модель не містить частину негативних аспектів, які регулярно фіксуються у відгуках споживачів, то це означає подальше зниження довіри користувачів, оскільки вони не отримують отримувати повну інформацію про ризики використання рішення.

Таблиця 1 – Показники ментальних моделей

Модель	Коефіцієнт структурного покриття	Коефіцієнт вагового покриття
М1 – збалансована	1,00	1,00
М2 – частково збалансована	0,51	0,56
М3 – незбалансована	0,11	0,13

Перспективними напрямами подальших досліджень є розширення експерименту на кілька різнорідних продуктів і доменів, поєднання автоматизованого семантичного аналізу відгуків із формальним виділенням ментальних моделей користувачів.

Висновки. Запропоновано підхід до оцінювання збалансованості пояснень рішень інтелектуальних систем з урахуванням негативних аспектів, який передбачає формування еталонної зваженої множини суттєвих негативних аспектів рішення, вилучення негативних елементів пояснення та розрахунок показників структурного і вагового покриття, а також оцінки пропорційності подання негативної інформації. Такий підхід дозволяє отримати кількісну оцінку, наскільки пояснення відображає обмеження та можливі негативні наслідки застосування рішення у співвідношенні до позитивних його властивостей.

Запропоновано метод верифікації збалансованості ментальних моделей рішення інтелектуальної системи. Метод включає етапи формування еталонної множини негативних аспектів рішення на основі аналізу користувацьких відгуків,

вилучення негативної компоненти ментальних моделей рішення, обчислення показників структурного та вагового покриття, оцінювання пропорційності й релевантності подання негативних аспектів, формування інтегральної оцінки щодо збалансованості ментальних моделей.

Метод забезпечує можливість уточнення ментальних моделей з урахуванням недоліків практичного застосування рішення користувача внаслідок неповноти моделей.

Експериментальна перевірка методу на основі набору користувацьких відгуків показала, що використання відгуків як джерела інформації щодо ментальних моделей користувачів дає можливість сформувати еталонну множину негативних аспектів рішення інтелектуальної системи, яка відображає проблеми використання та ризики, важливі для користувачів.

Декларація про використання генеративного штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при написанні тексту цієї роботи.

References

1. Goyal P., Ferrara E. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 2018, vol. 151, pp. 78–94. DOI: 10.1016/j.knosys.2018.03.022.
2. Brown D. E. Introduction to data mining for medical informatics. *Clinics in Laboratory Medicine*, 2008, vol. 28, no. 1, pp. 17–26. DOI: 10.1016/j.cll.2007.10.005.
3. Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D. A survey of methods for explaining black box models. *ACM Computing Surveys*, 2019, vol. 51, no. 5, art. 93. DOI: 10.1145/3236009.
4. 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.
5. Doshi-Velez F., Kortz M., Budish R. et al. Accountability of AI under the law: The role of explanation. *arXiv preprint*, 2017, arXiv:1711.01134.
6. 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*. San Francisco, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.
7. Zhang Y., Chen X. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval*, 2020, vol. 14, no. 1, pp. 1–101. DOI: 10.1561/1500000066.
8. Tintarev N., Masthoff J. Explaining recommendations: Design and evaluation. In: *Recommender Systems Handbook*, 2nd ed. Boston, Springer, 2015, pp. 353–382. DOI: 10.1007/978-1-4899-7637-6_10.
9. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Long Beach, 2017, pp. 4765–4774.
10. Samek W., Montavon G., Vedaldi A., Hansen L. K., Müller K.-R., eds. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham, Springer, 2019. 433 p. DOI: 10.1007/978-3-030-28954-6.
11. Kulesza T., Stumpf S., Wong W.-K. et al. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In: *Proceedings of the 2011 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. Pittsburgh, 2011, pp. 41–48. DOI: 10.1109/VLHCC.2011.6070395.
12. Eiband M., Schneider H., Bilandzic M., Fazekas-Con Z., Haug M., Hussmann H. Bringing transparency design into practice: User interfaces for explainable AI. In: *Proceedings of the 2018 Conference on Human Factors in Computing Systems (CHI'18 Workshop on Explainable Smart Systems)*. Montreal, 2018.
13. Garcez A. d'Avila, Besold T. R., De Raedt L. et al. Neural-symbolic learning and reasoning: A survey and interpretation. *Neurocomputing*, 2019, vol. 339, pp. 3–13. DOI: 10.1016/j.neucom.2019.01.034.
14. Besold T. R., d'Avila Garcez A., Bader S. et al. Neural-symbolic learning and reasoning: Contributions and challenges. In: *Proceedings of the AAAI Spring Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches*. Stanford, 2017.
15. Zhang L., Wang S., Liu B. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018, vol. 8, no. 4, e1253. DOI: 10.1002/widm.1253.

Надійшла (received) 17.03.2026
Прийнята (accepted) 10.04.2026
Оприлюднена (published) 20.05.2026

UDC 004.8:004.9

I. O. LESHCHYNSKA, PhD in Technical Sciences, Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor of the Department of Software Engineering, Kharkiv, Ukraine; ORCID: https://orcid.org/0000-0002-8737-4595; e-mail: iryna.leshchynska@nure.ua

METHOD FOR VERIFYING THE BALANCE OF MENTAL MODELS OF AN INTELLIGENT SYSTEM'S DECISION

The subject of the study is the process of verifying mental models of an intelligent system's decision. The aim of the work is to develop an approach for assessing the balance of explanations of intelligent systems' decisions with respect to their negative and positive aspects. In accordance with this aim, the following main tasks are addressed: to develop an approach to assessing the balance of explanations of intelligent systems' decisions based on the proportional representation of negative and positive characteristics in an explanation; to develop a general method for verifying the balance of mental models of an intelligent system's decision that takes into account the structural and weighted coverage of the set of negative aspects by the decision's mental models; to carry out an experimental evaluation of the proposed method using a set of user reviews of the operation of a recommender system that contain information about mental models of the proposed decisions. An approach to assessing the balance of explanations of intelligent systems' decisions that accounts for negative aspects is proposed. The approach involves constructing a reference weighted set of essential negative aspects of a decision, extracting negative elements of the explanation, computing indicators of structural and weighted coverage, and assessing the proportionality of the presentation of negative information. This approach provides a quantitative estimate of the extent to which an explanation reflects the limitations and potential negative consequences of applying a decision in relation to its positive properties. A method for verifying the balance of mental models of an intelligent system's decision is proposed. The method includes the stages of constructing a reference set of negative aspects of a decision based on the analysis of user reviews, extracting the negative component of the decision's mental models, computing indicators of structural and weighted coverage, assessing the proportionality and relevance of the presentation of negative aspects, and forming an integral measure of the balance of mental models. The method enables refinement of a mental model taking into account shortcomings of the practical application of a decision by the user that arise due to incompleteness of the models. Experimental evaluation of the method based on a set of user reviews has shown that using reviews as a source of information about users' mental models makes it possible to construct a reference set of negative aspects of an intelligent system's decision that reflects usage problems and risks important to users.

Keywords: explanation; artificial intelligence system; explainable artificial intelligence; dependencies; mental model; verification.

Повне ім'я автора / Author's full name
Автор 1 / Author 1: Ліщинська Ірина Олександрівна / Leshchynska Irina Oleksandrivna