

18. Gamification in Education: Benefits, Risks, Examples. *KeenEthics*. 2023. URL: <https://keenethics.com/blog/simulation-gamification-of-education> (дата звернення: 28.09.2024).
19. Paired Comparison (Methods, Examples, Tools). *OpinionX*. 2020. URL: <https://www.opinionx.co/research-method-guides/paired-comparison> (дата звернення: 28.09.2024).
20. Пономарьов А.В. Встановлення узгодженості результатів експертного оцінювання / А.В. Пономарьов, науковий керівник – доцент Козлов Ю.В. *Радіоелектроніка і молодь у XXI столітті : матеріали 25-го Міжнар. молодіжн. форуму, 20-22 квітня 2021 р.* Харків : ХНУРЕ, 2021. Т. 4. С. 165–166.
21. Labijak-Kowalska A., Kadziński M. Experimental comparison of results provided by ranking methods in Data Envelopment Analysis. *Expert Systems with Applications*. 2021. Vol. 173, № 114739. DOI: <https://doi.org/10.1016/j.eswa.2021.114739>

Надійшла до редколегії 24.06.2024 р.

Васильцова Наталія Володимирівна, кандидат технічних наук, доцент, професор кафедри ІУС ХНУРЕ, м. Харків, Україна, e-mail: natalia.vasytsova@nure.ua, ORCID: <https://orcid.org/0000-0002-4043-487X> (науковий керівник здобувачки вищої освіти Бурковської Анни Сергіївни).

Бурковська Анна Сергіївна, здобувачка вищої освіти, група ІУСТМ-22-1, факультет комп'ютерних наук, ХНУРЕ, м. Харків, Україна, e-mail: anna.burkovska@nure.ua

УДК 004.8:004.9

DOI: 10.30837/0135-1710.2024.182.066

С.Ф. ЧАЛИЙ, І.О. ЛЕЩИНСЬКА

УТОЧНЕННЯ МЕНТАЛЬНОЇ МОДЕЛІ РІШЕННЯ НА ОСНОВІ ДОПОВНЕННЯ ВХІДНИХ ДАНИХ В ЗАДАЧІ ФОРМУВАННЯ ПОЯСНЕНЬ В ІНТЕЛЕКТУАЛЬНІЙ СИСТЕМІ

Предметом дослідження є процес побудови ментальної моделі рішення для користувача інтелектуальної системи. Метою є розробка підходу до уточнення ментальної моделі користувача на основі принципу доповнення вхідних даних для того, щоб врахувати як позитивні, так і негативні аспекти рішення. Структуровано принцип доповнення вхідних даних з урахуванням особливостей формування пояснень в інтелектуальних системах. Запропоновано метод уточнення ментальної моделі рішення на основі доповнення вхідних даних. Метод включає етапи виявлення властивостей рішення, оцінки їхньої важливості, доповнення моделі та балансування позитивних і негативних аспектів рішення. Метод орієнтований на побудову пояснень, які враховують як переваги, так і обмеження рішення, що підвищує довіру користувачів до рішень інтелектуальних систем.

1. Вступ

Інтелектуальні інформаційні системи (ІС) використовують методи машинного навчання для аналізу даних та прийняття рішень. Проте через складність і непрозорість алгоритмів машинного навчання користувачі не завжди можуть зрозуміти, яким чином система дійшла кінцевого рішення. Це може привести до зниження довіри до ІС [1]. Для вирішення даної проблеми в галузі штучного інтелекту активно розвивається науковий напрямок пояснювального штучного інтелекту (Explainable Artificial Intelligence – XAI), орієнтований на розробку моделей і методів, що забезпечують прозорість та зрозумілість рішень для користувачів [2].

Однією із задач, що вирішуються в рамках побудови пояснень у XAI, є узгодження пояснення з ментальною моделлю користувача. Ментальна модель – це відображення зовнішньої реальності в свідомості людини. Таке відображення допомагає людині зрозуміти та інтерпретувати навколишні події [3]. У контексті ІС ментальна модель має відображати вхідні дані та рішення системи таким чином, щоб користувач міг їх коректно інтерпретувати та зрозуміти, як їх використовувати для вирішення практичних задач у

своїй професійній сфері діяльності.

Проте існуючі підходи до побудови ментальних моделей часто не враховують специфічні потреби зовнішніх користувачів ІС. Зокрема, вони не завжди відображають обмеження або негативні аспекти рішення, які можуть бути важливими для користувача. Така невизначеність може призводити до неповних або некоректних висновків при використанні інтелектуальної системи [4]. Тому розробка ментальних моделей, що враховують негативні аспекти рішення ІС, є ключовою умовою для створення зрозумілих пояснень в інтелектуальних системах.

2. Аналіз літературних даних і постановка проблеми дослідження

Напрямок ХАІ побудови пояснень у системах штучного інтелекту був започаткований у агентстві DARPA. ХАІ орієнтований на розробку методів, які дають можливість розуміти причини рішення ІС з тим, щоб підвищити довіру користувачів до таких систем [1]. Дослідження в галузі ХАІ зосереджені на створенні інтерпретованих моделей, розробці методів пост-пояснень та побудові інтерфейсів пояснень [2].

У напрямку досліджень ХАІ розроблено ряд підходів до побудови пояснень щодо рішень інтелектуальних систем. Метод контрфактуальних пояснень [3] забезпечує для користувача можливість отримати відповідь на питання: «Що потрібно змінити у вхідних даних, щоб отримати інше рішення?». Метод відображає альтернативні сценарії, що допомагає користувачам оцінити вплив різних значень вхідних даних на рішення ІС. Однак контрфактуальні пояснення часто не враховують негативні аспекти рішення.

Метод LIME (Local Interpretable Model-agnostic Explanations) [5] формує локальну інтерпретовану модель формування рішення для конкретного прикладу. Проте LIME фокусується лише на локальних аспектах рішення, що не дає можливості відобразити загальні закономірності побудови рішення, в тому числі обмеження на вхідні дані.

Метод SHAP (SHapley Additive exPlanations) використовує теорію ігор для розподілу важливості кожної ознаки у прогнозі моделі [6]. Метод забезпечує узгоджені пояснення на локальному рівні. Проте SHAP може бути обчислювально складним для великих моделей.

Метод пошарового розподілу релевантності (LRP) [7] розподіляє релевантність від виходу моделі до її вхідних даних. Перевагою LRP є можливість отримання детальних пояснень про внесок кожного шару мережі у кінцевий результат. Однак LRP не завжди враховує негативні аспекти рішення або його обмеження.

Каузальні методи побудови пояснень з можливістю безпосередньої інтерпретації користувачем було розроблено у [8]-[10]. Можливість безпосередньої інтерпретації каузальних залежностей було показано в [11].

Проте в розглянутих методах не приділяється достатньо уваги узгодженню пояснень із ментальними моделями рішень, які використовують користувачі ІС.

Ментальні моделі відіграють ключову роль у розумінні та інтерпретації рішень інтелектуальних систем користувачами [3]. Основи теорії ментальних моделей було закладено в [12], [13]. Ментальні моделі було визначено як внутрішні репрезентації зовнішнього світу, які люди використовують для розуміння, міркування та прогнозування. Принципи побудови ментальних моделей, розроблені в цих роботах, було розширено принципом доповнення вхідних даних в [14].

Роль ментальних моделей у взаємодії людини з технологіями проаналізовано в [15]. В даній роботі було обґрунтовано важливість розуміння ментальних моделей користувачів для проєктування інтуїтивно зрозумілих інтерфейсів. Особливості представлення ментальної моделі для побудови пояснень в ІС розглянуто в [16].

Існуючі методи побудови ментальних моделей, такі як метод структурного відображення [17] та метод каузального моделювання [18], можуть представляти

структуру та причинно-наслідкові зв'язки у рішеннях ІС. Однак ці методи не враховують обмеження або негативні аспекти рішення. Метод побудови ментальних моделей на основі спільного розуміння [19] розглядає ІС як члена команди людей, що створює умови для ефективної взаємодії між людиною та інтелектуальною системою. Однак побудова точної ментальної моделі ІІІ може бути складною для користувачів, особливо коли йдеться про врахування потенційних негативних аспектів рішень системи.

Аналіз існуючих методів побудови пояснень та ментальних моделей у ХАІ дає можливість зробити висновок про те, що в існуючих підходах, зазвичай, не приділяється достатньо уваги доповненню ментальної моделі відображенням негативних аспектів рішення ІС. Це обмеження може призвести до формування неповної або некоректної ментальної моделі.

Таким чином, актуальною є задача розробки методу уточнення ментальної моделі, який би враховував як позитивні, так і негативні аспекти рішень ІС. Вирішення даної задачі дає можливість отримати збалансоване представлення рішення, включаючи його потенційні обмеження та ризики. Формування узгодженого із потребами користувача рішення підвищує довіру до інтелектуальних систем та ефективність їх використання [20].

3. Мета і задачі дослідження

Метою даного дослідження є розробка підходу до уточнення ментальної моделі користувача інтелектуальної системи на основі принципу доповнення вхідних даних з тим, щоб враховувати не лише позитивні властивості рішення, але й можливі обмеження або негативні аспекти для підвищення рівня довіри до пояснень в інтелектуальних системах.

Для досягнення поставленої мети у роботі вирішуються такі задачі: структуризація принципу доповнення вхідних даних; розробка методу уточнення ментальної моделі користувача інтелектуальної системи на основі доповнення вхідних даних.

4. Принцип доповнення вхідних даних ментальної моделі користувача інтелектуальної системи

Принципи побудови ментальних моделей, що розглядались в [12]-[14], включають в себе такі: відповідності структури ментальної моделі та рішення; множинності ментальних моделей, що дає можливість врахувати окремі варіанти рішення; неповноти даних для прийняття рішення; доповнення вхідних даних ментальної моделі.

Принцип доповнення вхідних даних ментальної моделі користувача інтелектуальної системи дає можливість користувачеві уточнити бажане рішення з урахуванням властивостей, які не були враховані в початковій ментальній моделі.

Користувачі можуть формувати неповні ментальні моделі рішень, враховуючи лише їхні позитивні (бажані) властивості. Такий підхід відображає когнітивні упередження користувачів та може призвести до прийняття нерациональних рішень внаслідок ігнорування потенційних обмежень. Наприклад, при виборі товару у системі e-commerce користувач може зосередитися на окремих технічних характеристиках товару, таких як розмір пам'яті або ціна, але не звернути увагу на такі важливі обмеження, як тривалість гарантії або існуючі проблеми з обслуговуванням.

Принцип доповнення вхідних даних спрямований на подолання цього обмеження шляхом включення додаткової інформації, яка може бути важливою для повного розуміння рішення. Згідно з цим принципом, необхідно провести додаткову оцінку рішення з урахуванням обмежень, що представляють значення властивостей, які раніше не були включені до ментальної моделі користувача.

Даний принцип доповнює принцип відповідності структури, забезпечуючи повніше відображення структури рішення в ментальній моделі. Принцип множинності ментальних моделей розширюється за рахунок включення альтернативних моделей, які враховують

різні комбінації позитивних та негативних аспектів рішення.

Принцип доповнення вхідних даних є комплементарним до принципу неповноти вхідних даних. Якщо останній принцип призводить до спрощення ментальної моделі шляхом виключення «неважливої» інформації, то принцип доповнення вхідних даних компенсує це спрощення, включаючи в модель важливі, але раніше проігноровані аспекти рішення.

У практичному плані застосування принципу доповнення вхідних даних може базуватися на аналізі відгуків інших користувачів інтелектуальної системи. Такі відгуки часто містять інформацію про властивості рішення, які не були очевидними або важливими для користувача на етапі початкового формування ментальної моделі.

5. Метод уточнення ментальної моделі користувача інтелектуальної системи на основі доповнення вхідних даних

Повна ментальна модель користувача M , згідно з принципами множинності, доповнення та неповноти вхідних даних, складається з множини локальних ментальних моделей M_i : $M = \{M_i\}$. Кожна з M_i використовується для визначеної підмножини вхідних даних. Наприклад, при використанні рекомендаційної підсистеми в системі електронної комерції повна ментальна модель смартфона може складатись з локальних ментальних моделей окремих моделей смартфонів, які рекомендовано користувачеві (або які переглядав користувач). Згідно з принципом неповноти вхідних даних, початкова модель M_i містить лише бажані для користувача властивості:

$$M_i = \{V_i^+\}. \quad (1)$$

Згідно з принципом доповнення вхідних даних, модель має характеризуватись неперетинними наборами позитивних V_i^+ та негативних V_i^- властивостей:

$$M_i = \{V_i^+, V_i^- \mid V_i^+ \cap V_i^- = \emptyset\}. \quad (2)$$

Ключова ідея методу уточнення ментальної моделі користувача полягає у визначенні та оцінці позитивних та негативних властивостей рішення ІС на основі аналізу відгуків користувачів такої системи. В подальшому негативні властивості, вага яких перевищує заданий поріг, включаються до складу ментальної моделі.

Метод уточнення ментальної моделі користувача містить такі етапи.

Етап 1. Аналіз відгуків користувачів щодо реалізованих рішень ІС.

Крок 1.1. Виявлення підмножини позитивних властивостей V_i^+ рішення.

Крок 1.2. Виявлення підмножини негативних властивостей V_i^- рішення.

Етап 2. Оцінка важливості властивостей рішення.

Крок 2.1. Для кожної властивості $v_{i,j}^+ \in V_i^+$ або $v_{i,j}^- \in V_i^-$ розрахунок ваги $w_{i,j}^+$ на основі кількості $n_{i,j}$ згадувань про цю властивість у відгуках. Розрахунок ваги i для позитивних, i для негативних властивостей має вигляд:

$$w_{i,j} = \frac{n_{i,j}}{|V_i^+ \cup V_i^-|}. \quad (3)$$

Таким чином, вага відображає частоту згадувань про властивість у відгуках

користувача.

Крок 2.2. Ранжування негативних властивостей за зменшенням ваги.

Етап 3. Доповнення ментальної моделі негативними властивостями рішення.

На даному етапі до ментальної моделі, крім позитивних властивостей V_i^+ рішення, включаються властивості з вагою, більшою порогового значення ε :

$$M_i = \left\{ V_i^+, V_i^- \mid w_{i,j}^- > \varepsilon \right\}. \quad (4)$$

Етап 4. Балансування позитивних і негативних властивостей рішення.

Крок 4.1. Обчислення сумарних ваг $\sum_j w_{i,j}^+$ та $\sum_j w_{i,j}^-$ відповідно позитивних і негативних властивостей доповненої ментальної моделі.

Крок 4.2. Перевірка умови балансу ваг:

$$\sum_j w_{i,j}^+ > \sum_j w_{i,j}^-. \quad (5)$$

Дана умова визначає переваги рішення для користувача, тобто можливість використання рішення на практиці. Сенс принципу доповнення полягає в тому, щоб користувач враховував негативні властивості рішення, а не відкидав рішення повністю. Тому сумарна вага позитивних властивостей має переважати вагу негативних.

Крок 4.3. Якщо умова (5) не виконується, то збільшення порога на етапі 3, після чого – повторне виконання етапу 4.

Результатом застосування методу є ментальна модель користувача, що містить набори позитивних (бажаних для користувача) та найсуттєвіших негативних властивостей, вага яких перевищує порогове значення. Використання даної моделі для пояснень щодо рішення ПС дає можливість представити користувачеві не лише переваги рішення, а й можливі обмеження при його використанні.

6. Експериментальна перевірка методу доповнення ментальної моделі користувача інтелектуальної системи

Експериментальну перевірку методу орієнтовано на доповнення локальних ментальних моделей M_1 та M_2 двох смартфонів вищого цінового сегменту для однієї цільової аудиторії. Використано відгуки покупців цих смартфонів на платформі електронної комерції. Властивості базової ментальної моделі (без негативних властивостей) наведено в табл. 1.

Таблиця 1

Позитивні властивості двох моделей смартфонів M_1 та M_2

Властивість	Моделі	
	M_1	M_2
Кількість відгуків користувачів	231	66
Кількість позитивних властивостей	7	6
Сумарна вага позитивних властивостей	0,8398	0,7121

Ці моделі мають схожі набори позитивних властивостей: співпадають 5 базових властивостей і не співпадають 3 другорядні властивості.

Зважені негативні властивості смартфонів представлено в табл. 2. Вага враховує кількість відгуків за кожною позицією.

Таблиця 2

Зважені негативні властивості двох моделей смартфонів

Властивість	Вага	
	M_1	M_2
Вартість	0,0303	0,1060
Значне нагрівання корпусу	0,0389	–
Час (автономність) роботи	0,0346	–
Завеликі вага та розмір	0,0259	–
Не задовольняє якість звуку	0,0173	–
Проблеми з прийомом сигналу	0,0129	–
Відсутність або висока вартість аксесуарів	–	0,0757
Проблеми з автоматичними оновленнями	–	0,0606
Якість матеріалів корпусу	–	0,0454
Сумарна вага	0,1601	0,2878

Сумарна вага позитивних властивостей значно перевищує вагу негативних. Тому поріг $\varepsilon = 0$ і всі негативні властивості включені до складу обох ментальних моделей.

6. Обговорення результатів

Результати експериментальної перевірки показали, що ментальні моделі для схожих смартфонів мають аналогічні набори позитивних властивостей, але суттєво різні підмножини негативних властивостей. Тобто доповнення ментальних моделей негативними властивостями дає можливість обґрунтувати вибір конкретного рішення ПС в рамках пояснення.

Експериментальна перевірка показала, що пояснення має відповідати повній ментальній моделі та представляти позитивні й негативні властивості рішення за умови більшої ваги позитивних властивостей. В позитивному аспекті пояснення обґрунтовує відповідність вимогам користувача до рішення ПС, а в негативному – визначає можливість використання рішення при вирішенні задач користувача у відповідній предметній галузі.

7. Висновки

У дослідженні запропоновано підхід до удосконалення ментальної моделі рішення інтелектуальної системи шляхом урахування негативних властивостей даного рішення.

З метою обґрунтування запропонованого підходу виконано структурування принципу доповнення вхідних даних та визначено його зв'язки з іншими принципами побудови ментальних моделей, що визначають їхню структуру, множинність та неповноту.

Запропоновано метод доповнення вхідних даних ментальної моделі рішення інтелектуальної системи, що базується на принципі доповнення вхідних даних та включає етапи аналізу відгуків користувачів, оцінки важливості виявлених властивостей рішення, доповнення ментальної моделі з урахуванням негативних властивостей та балансування позитивних і негативних аспектів рішення. Метод дає можливість формувати пояснення, які враховують як переваги, так і обмеження рішення інтелектуальної системи, що створює умови для вилучення типових когнітивних упереджень, пов'язаних із ігноруванням негативних аспектів рішення. Експериментальна перевірка запропонованого методу показала можливість автоматизованого доповнення ментальної моделі рішення на основі аналізу відгуків користувачів інтелектуальної системи.

Перелік посилань:

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160. <https://doi.org/10.1109/ACCESS.2018.2870052>.

2. Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence (XAI) program: A retrospective. *AI Magazine*, 40(2), 44-58. <https://doi.org/10.1609/aimag.v40i2.2850>.
3. Rook, L. (2021). Mental models: A robust definition. *The Learning Organization*, 28(1), 6-17. <https://doi.org/10.1108/TLO-09-2019-0136>.
4. Byrne, R. M. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences*, 6(10), 426-431. [https://doi.org/10.1016/S1364-6613\(02\)01974-5](https://doi.org/10.1016/S1364-6613(02)01974-5).
5. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). «Why should I trust you?»: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.
6. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.
7. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2022). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3), 247-278.
8. Чалий С. Ф., Лещинський В. О. Темпорально-каузальні методи побудови пояснень в системах штучного інтелекту. *АСУ та прилади автоматики*. Вип. 181. 2024. С. 91-99. DOI: 10.30837/0135-1710.2024.181.091.
9. Чалий С. Ф., Лещинський В. О. Побудова пояснень на локальному рівні представлення процесу формування рішень для внутрішніх користувачів інтелектуальної системи. *Біоніка інтелекту*, 2024. Т. 1 (100). С. 58-64. DOI 10.30837/bi.2024.1(100).08.
10. Чалий, С., Лещинський, В., & Лещинська, І. (2022). Каузальна модель процесу побудови пояснень в інформаційній системі. *Системи управління, навігації та зв'язку*, 3(69), 99-103. <https://doi.org/10.26906/SUNZ.2022.3.099>.
11. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38.
12. Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.
13. Johnson-Laird, P. N. (2006). *How we reason*. Oxford University Press/
14. Чалий, С. Ф., & Лещинська, І. О. (2024). Принципи побудови ментальних моделей рішення для зовнішнього користувача в задачі формування пояснень в інтелектуальній системі. *АСУ і прилади автоматики*, (181), 17-25.
15. Norman, D. A. (2014). Some Observations on Mental Models. In D. Gentner & A. L. Stevens (Eds.), *Mental Models* (pp. 7-14). Psychology Press.
16. Чалий, С., & Лещинська, І. (2023). Концептуальна ментальна модель пояснення в системі штучного інтелекту. *Вісник Національного технічного університету «ХПІ»*. Серія: Системний аналіз, управління та інформаційні технології, (1 (9), 70–75. <https://doi.org/10.20998/2079-0023.2023.01.11>
17. Gentner, D., & Stevens, A. L. (2014). *Mental models*. Psychology Press.
18. Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.
19. Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1), 2-11.
20. Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31-57.

Надійшла до редколегії 25.09.2024 р.

Чалий Сергій Федорович, доктор технічних наук, професор, професор кафедри ІУС ХНУРЕ, м. Харків, Україна, e-mail: serhii.chalyi@nure.ua; ORCID: 0000-0002-9982-9091
Лещинська Ірина Олександрівна, кандидат технічних наук, доцент, доцент кафедри програмної інженерії ХНУРЕ, м. Харків, Україна, e-mail: iryna.leshchynska@nure.ua, ORCID: <https://orcid.org/0000-0002-8737-4595>