УДК 004.891.3 **doi: 10.26906/SUNZ.2023.3.138**

С. Ф. Чалий, В. О. Лещинський

Харківський національний університет радіоелектроніки, Харків, Україна

МОЖЛИВІСНА МОДЕЛЬ КАУЗАЛЬНОГО ЗВ'ЯЗКУ ПО ВХІДНІЙ ЗМІННІЙ ДЛЯ ПОБУДОВИ ПОЯСНЕННЯ В ІНТЕЛЕКТУАЛЬНІЙ СИСТЕМІ

Анотація. Предметом вивчення в статті є процеси побудови пояснень для отриманих в інтелектуальній інформаційній системі рішень. **Метою** побудова моделі причинно-наслідкових зав'язків для побудови пояснень в умовах невизначеності щодо станів інтелектуальної інформаційної системи, якщо остання представляється у вигляді чорного ящика. **Завдання**: структуризація пояснення з урахуванням особливостей когнітивної діяльності людини; формування необхідної та достатньої умови щодо каузальної залежності як складової пояснення з використанням теорії можливостей; розробка можливісної моделі каузальної залежності для одної вхідної змінної, яка враховує невизначеність щодо станів інтелектуальної системи. Використовуваними **підходами** є: підходи до побудови пояснень у когнітивній діяльності людини, підходи до побудови пояснень у пояснювальному штучному інтелекті. Отримані наступні **результати**. Структуровано пояснення як елемент когнітивної діяльності людини. Показано, що пояснення може бути представлено в двох аспектах: концептуальному, шляхом порівняння вхідної інформації із існуючою системою знань людини; тлумачному, шляхом порівнянні властивостей вхідних об'єктів. Запропоновано можливісні необхідна та достатня умови для каузальної залежності на базі однієї вхідної змінної, що лежить в основі пояснення. Запропоновано модель можливісної каузальної залежності для побудови пояснення в інтелектуальній системі. **Висновки**. Наукова новизна отриманих результатів полягає в наступному. Запропоновано модель можливісної каузальної залежності між вхідною змінною та результатом роботи інтелектуальної системи, що поєднує необхідну умову каузальності у вигляді рівня довіри до впливу вхідної змінної на результат та достатню умову каузальності у вигляді максимальної можливості впливу значення вхідної змінної на результат інтелектуальної системи. Модель дає можливість сформувати каузально-орієнтоване пояснення на основі зв'язку вхідної змінної і отриманого результату в умовах неповноти знань модель щодо стану інтелектуальної системи.

Ключові слова: інтелектуальна інформаційна система, пояснення, причинно-наслідковий зв'язок, каузальна залежність, когнітивна діяльність.

Вступ

Сучасні інтелектуальні системи широко використовуються для підтримки рішень при вирішенні складних задач пошуку інформації, страхування, лікування, у банківській та військовій справі, тощо.

Використання в інтелектуальних системах непрозорих алгоритмів, що базуються на машинному навчанні, забезпечує формування комплексних результатів, які не завжди є зрозумілими для користувачів таких систем. Непрозорість результатів, у свою чергу, може породжувати сумніви та недовіру користувачів до рішень інтелектуальної системи, і, як наслідок, обмежувати реальне використання останніх.

Тому прозорість та зрозумілість рішень в цих системах є важливою умовою їх широкого практичного застосування [1, 2].

Зрозумілість рішень для користувачів систем штучного інтелекту забезпечується використанням пояснень [3].

Пояснення розглядається в двох аспектах: як кінцевий продукт та як процес.

Пояснення як кінцевий продукт представляє користувачеві необхідну інформацію для розуміння прийнятих інтелектуальною інформаційною системою рішень. Фактично таке пояснення встановлює зв'язки між вхідною інформацією та виходом інформаційною системи. Побудова такого пояснення базується на представленні системи у вигляді чорного ящику.

Тобто інформація щодо внутрішніх станів системи у даному випадку є неповною або відсутньою.

Задача формування пояснення як кінцевого продукту полягає у визначенні каузальних залежностей між вхідними змінними та отриманим результатом.

Дана задача вирішується в умовах невизначеності щодо конкретної реалізації процесу прийняття рішення в інтелектуальній системі [4, 5].

Пояснення як процес включає в себе послідовність кроків, що обґрунтовують, як система прийшла до поточного рішення. Цей підхід потребує інформації щодо внутрішніх станів інтелектуальної системи.

В загальному випадку пояснення як процес не завжди орієнтовано на обґрунтування отриманого результату.

Даний підхід зв’язує дії із прийняття рішення, встановлюючи залежності між ними.

Таким чином, пояснення має представити каузальну залежність або послідовність каузальних залежностей, що обґрунтовують зв'язок між входом та виходом інтелектуальної системи в умовах неповноти інформації щодо станів системи у процесі прийняття рішення.

Сучасні підходи до побудови пояснення було сформовано в рамках програми Explainable Artificial Intelligence [6].

Такі підходи вирішують дві задачі: моделювання сприйняття пояснень людиною та побудова пояснень.

В рамках вирішення першої задачі визначаються особливості та принципи побудови пояснень людиною у процесі пізнання навколишнього світу [6, 7].

При вирішенні другої задачі згідно визначених раніше принципів розробляються методи побудови пояснень.

Пояснення базуються або на безпосередній інтерпретації процесу прийняття рішення [4, 8], або на явному чи неявному представленні темпоральних [9- 11] та каузальних [12, 13] залежностей, що пов'язують вхідні дані та результат із знаннями щодо використання отриманого рішення, а також використовують порівняння із альтернативами [6, 14].

Однак існуючі підходи не приділяють достатньо уваги побудові каузальних залежностей в умовах невизначеності, при багатоваріантності можливих рішень і, відповідно, багатоваріантності пояснень щодо цих результатів.

В той же час, при представленні інтелектуальної системи у вигляді чорного ящика, пояснення оперує із множиною вхідних змінних, значення яких впливають на результат.

Тому актуальною є задача побудови таких каузальних зав'язків, які враховують можливості впливу кожного значення змінної на отриманий інтелектуальною системою результат.

Метою статті є побудова моделі причинно-наслідкових зав'язків для побудови пояснень в умовах невизначеності щодо станів інтелектуальної інформаційної системи, якщо остання представляється у вигляді чорного ящика.

Для досягнення поставленої мети вирішуються такі задачі:

– структуризація пояснення з урахуванням особливостей когнітивної діяльності людини;

– формування необхідної та достатньої умови щодо каузальної залежності як складової пояснення з використанням теорії можливостей;

– розробка можливісної моделі каузальної залежності для одної вхідної змінної, яка враховує невизначеність щодо станів інтелектуальної системи.

Структуризація пояснення як елементу когнітивної діяльності

Моделювання пояснення в умовах невизначеності базується на виділенні його властивостей як елементу когнітивної діяльності людини.

Згідно досліджень у сфері когнітивної психології, пояснення є невід'ємною складовою процесу пізнання.

Останній містить такі ключові фази:

– отримання / сприйняття інформації;

– обробка й зберігання інформації та знань;

– використання інформації та знань.

На першій фазі зовнішня інформація трансформується у внутрішнє представлення.

На другій фазі інформація перетворюється в знання, які визначають залежності між отриманими на першій фазі даними.

На даній фазі для розуміння сутності вхідної інформації шляхом встановлення залежностей між елементами даних використовуються пояснення.

На третій фазі отримані та збережені знання використовуються для прийняття рішень та вибору відповідних дій.

Виходячи з представленої структуризації процесу пізнання, останній представляє собою процес отримання, сприйняття, розуміння, зберігання, пошуку, обробки та використання інформації й знань людиною.

Тому пояснення забезпечує здатність людини отримувати, обробляти інформацію й знання та приймати й реалізовувати відповідні рішення. Використання пояснень у процесі пізнання створює умови для осмислення навколишнього світу, а також формування знань і поведінки людини.

Відповідно, пояснення відіграє одну з ключових ролей при відкритті нових знань та є важливим елементом розуміння об'єктів або явищ у предметній області [15].

Дослідження в сфері когнітивної психології встановлюють зв'язок між пізнанням та поясненнями у двох аспектах: концептуальному та тлумачному. Відмінності цих аспектів пояснення представлено в табл. 1.

Таблиця 1 – **Порівняння структурних елементів пояснення як складової когнітивної діяльності людини**

Аспект пояснення	Структурні елементи пояснення	Відмінності та вимоги
Концептуальний, що базується на відповідності пояснення фоновим знанням людини	Концепція пояснення як залежність між вхідними даними	Використовується набір допустимих пояснень; відбирається пояснення з цього набору, що відповідає фоновим знанням
	Базові (фонові) знання щодо предметної області	Потребує постійного уточнення системи базових знань
Тлумачний, що використовує порівняння властивостей екземплярів вхідних даних	Інформація про властивості вхідних об'єктів	Встановлюється типовість, різноманітність та подібність властивостей об'єктів
	Причинно-наслідкові зв'язки для категоризації об'єктів	Відбирається «найкраще» пояснення

У першому аспекті пізнання людини розглядається як процес побудови концептуальних представлень навколишнього світу. В даному аспекті пояснення є основою навчання та міркувань і, відповідно, умовою побудови концепцій щодо поточної предметної області.

Зокрема, в роботі [15] при розгляді сприйняття концепцій людьми показано, що необхідною умовою для розуміння предметів та явищ є відповідність концепцій базовим (або фоновим) знанням людей щодо навколишнього світу. Така відповідність дає можливість обґрунтувати кожну концепцію шляхом використання одного із набору допустимих пояснень. Розвиток таких концепцій базується на побудові та реструктуризації системи знань людини [16].

У другому аспекті ключова увага приділяється «тлумачному» мисленню, заснованому на побудові каузальних залежностей при порівнянні властивостей вхідних об'єктів у процесі категоризації знань. Безпосередньо категоризація відбувається згідно найкращого пояснення [17].

Проведені в роботі [17] експерименти показали, що використання причинно-наслідкових пояснень при категоризації дає можливість врахувати типовість, різноманітність та подібність властивостей вибраних категорій. Також узагальнення категорій містить ознаки причинно-наслідкових зв’язків.

Тобто в даному аспекті досліджується пізнання на основі побудови причинно-наслідкових зв'язків, що пояснюють відому послідовність подій і є упорядкованими у часі.

Таким чином, в концептуальному аспекті пояснення розкривають зв'язки між новими концепціями та існуючими знанням щодо предметної області, що дає можливість обґрунтувати нові концепції та інтегрувати їх із існуючими знанням. В тлумачному аспекті пояснення забезпечує категоризацію вхідної інформації на основі використання каузальних залежностей, що задають причинно-наслідкові зв'язки між властивостями вхідних об'єктів та категоріями цих об'єктів.

В обох випадках виконується вибір найкращого пояснення із множини можливих на основі їх відповідності фоновим знанням про предметну галузь або ж знанням щодо властивостей та категоризації об'єктів.

В цілому можна зробити висновок, що пояснення задають причинно-наслідкові зв’язки між елементами знань людини в умовах невизначеності щодо знань про предметну область в цілому або про окремі об’єкти в рамках предметної області. Тобто пояснення через визначають причини для наслідків, що розглядаються людиною. Такий каузальний зв'язок може бути встановлено [18]:

– шляхом відповіді на питання типу як або чому було отримано поточний результат (стан, рішення, тощо)?;

– як судження, яке деталізує, чому отримано той чи інший результат;

– безпосередньо через схему причини-наслідки.

Також при формуванні пояснень розділяють представлення пояснення як важливого для людини результату, що може бути оціненим людиною, та пояснення як процес [19].

Пояснення як результат це правило, що визначає причини отриманого рішення. Тобто пояснення як результат становить безпосередню цінність для людини і людина робить оцінку цього пояснення.

Пояснення як процес це когнітивна діяльність, яка формує пояснення як результат. В рамках діяльності може бути сформовано одне або декілька пояснень - результатів.

Таким чином, пояснення в когнітивній діяльності людини базується на множині залежностей, що пов'язують відому вхідну інформацію, знання людини про предметну область, а також рішення, що ця людина приймає. З цією множини залежностей формується каузальна залежність, яка і є основою для результуючого тлумачення. В умовах невизначеності попередньо необхідно сформулювати необхідну й достатню умови, що обумовлюють можливість побудови такої залежності.

Можливісна модель каузальної залежності як основи пояснення в інтелектуальній системі

Пояснення щодо результату інтелектуальної системи є узагальненням послідовності причинно-наслідкових зв'язків, що відображають процес його отримання, і пов'язує властивості вхідних даних із отриманим в інтелектуальній системі рішенням.

Пояснення на основі властивостей вхідних даних визначає, в якій мірі значення кожної із вхідних змінних впливає на отримане в системі рішення.

Розглянемо визначення необхідної й достатньої умов для каузальної залежності, що визначає пояснення щодо результату, для однієї вхідної змінної. Дана змінна $X = \{x_i\}$ має множину можливих значень x_i.

Оскільки при формуванні пояснення щодо результату інтелектуальна інформаційна система представляється у вигляді «чорного ящику», то інформація щодо внутрішніх станів цієї системи є недоступною.

Відповідно, вплив кожного із вхідних значень на кінцевий результат інтелектуальної інформаційної задається через нормовану оцінку

$$0 \le \pi(x_i) \le 1.$$

На практиці оцінка $\pi(x_i)$, як правило, відображає ймовірність використання значення x_i.

Але в загальному випадку дана оцінка є функцією, що відображає множину значень x_i на інтервал $[0,1]$.

Зокрема, якщо при формуванні пояснення щодо запропонованого рекомендаційною системою ноутбука змінна X містить значення типів та поколінь процесора цього ноутбука, то оцінка $\pi(x_i)$ може задавати ймовірність вибору ноутбуків різних моделей саме з визначеним процесором x_i.

Очевидно, що в ситуації, коли у рекомендованому комп’ютері є процесор, який найчастіше фігурував у інших куплених ноутбуках, пояснення може спиратись на модель цього процесора.

Тобто пояснення спирається на відмінності у властивостях елементів категорії «процесори», що відповідає розглянутому вище тлумачному аспекту пояснення у когнітивному процесі людини.

Наприклад, пояснення в рекомендаційній системі на основі моделі процесора задає відмінності для множини процесорів, які були використані у куплених раніше комп'ютерах:

$$\{x_1 = i7\_13 \textit{покоління}, x_2 = i7\_12 \textit{покоління},$$
$$x_3 = i7\_11 \textit{покоління}, x_4 = i7\_10 \textit{покоління}\}$$

в тому випадку, якщо рекомендований список містить комп'ютерів упорядковується з урахуванням моделі процесора за показником $\pi(x_i)$.

Однак, при наявності у користувача рекомендаційної системи додаткових структурованих знань щодо елементів комп'ютерної техніки, таке пояснення може враховувати і концептуальний аспект пояснення.

Наприклад, при появі у рекомендованому списку комп'ютера з процесором

$$i7\_11 \textit{покоління}$$

та додаткових знаннях користувача щодо достатньо високої потужності, раннього року випуску й, відповідно, помірної ціни даного процесора, пояснення на основі моделі процесора обґрунтовує концепцію прагматичного вибору:

«потужний комп'ютер за розумну ціну».

Таким чином, пояснення визначає можливість рекомендації комп'ютера з на основі значення x_i.

Такий підхід передбачає одночасне використання множини значень $\{x_i\}$ для визначення можливісного опису пояснення.

Зазначене свідчить про актуальність використання теорії можливостей для формування необхідної й достатньої умови для каузальної залежності між вхідною змінною і отриманим в системі рішенням.

Розглянемо формально можливісну оцінку вхідної змінної $\pi(x_i)$ з урахуванням темпорального фактору:

– множина X складається із підмножин X_j, які визначають всі можливі значення змінної на різних інтервалах часу T_j:

$$X_j = \{x_{j,i}\} \subseteq X : \forall j \exists T_j = \left[t_j^{beg}, t_j^{end}\right], \quad (1)$$

– інтервали часу T_j в загальному випадку перетинаються:

$$(\forall m \neq j) T_j \bigcap T_m \neq \varnothing, \quad (2)$$

– ідентичні значення $x_{j,i}$ можуть зустрічатись на різних інтервалах і належати різним підмножинам X_j:

$$(\exists m \neq j) : x_{j,i} = x_{m,i}. \quad (3)$$

Розподіл оцінок $\pi(x_i)$ для кожної множини X_j має вигляд:

$$P_j = \{\pi(x_{j,1}) \geq \pi(x_{j,2}) \geq ... \geq \pi(s_{j,I})\}. \quad (4)$$

Можливість Π_j для кожної підмножини X_j визначається як її верхня грань:

$$\Pi_j = \sup_i \pi(x_{j,i}). \quad (5)$$

Можливісна оцінка Π для множини X визначається як відображення сукупності можливих підмножин X_j на інтервал $[0,1]$:

$$\Pi : \{X_j\} \to [0,1]. \quad (6)$$

за умов:

$$\Pi(\varnothing) = 0,$$
$$\Pi\left(\bigcup_j X_i\right) = \sup_i \Pi(X_i). \quad (7)$$

Каузальна залежність має встановлювати зв'язок між значенням вхідної змінної та отриманим в інтелектуальній системі результатом.

Згідно (7), чим вище значення можливості, тим більше ймовірність використання відповідного значення $x_{j,i}$, такого що:

$$x_{j,i} : \forall (j, i \neq m)\ \pi(x_{j,i}) \geq \pi(x_{j,m}). \quad (8)$$

Тоді достатня умова для наявності каузальної залежності між значенням $x_{j,i}$ та отриманим результатом полягає в тому, щоб значення $\pi(x_{j,i})$ на поточному інтервалі часу T_j було максимальним:

$$\Pi(X_j) = \max_i \pi(x_{j,i}). \quad (9)$$

Згідно (9), достатня умова визначається лише для поточної підмножини X_j, тобто для актуальних на поточному інтервалі T_j даних.

Необхідна умова для наявності каузальної залежності з використанням теорії можливостей пов'язана із довірою до того, що значення вхідних даних впливають на отриманий в інтелектуальній системі результат.

Така довіра визначається тим, як часто (з якою оцінкою) була використані значення даної змінної для отримання результату.

Наприклад, як часто в найгіршому випадку були використані дані про процесор для побудови персоналізованого переліку комп'ютерів в рекомендаційній системі.

Ця умова може бути сформована з використанням оцінки необхідності N в теорії можливостей.

Дана оцінка на множині можливих підмножин X_j визначається наступним чином. N є відображенням елементів можливих підмножин X_j на інтервал $[0,1]$:

$$T:\{X_j\}\to[0,1]. \tag{10}$$

за умов:

$$N(\varnothing)=0,$$
$$N\left(\bigcap_j X_j\right)=\inf_i N\left(\bigcap_j X_j\right)\Big|\bigcap_j X_j\neq\varnothing. \tag{11}$$

Згідно (11), оцінка необхідності задається для перетину відомих підмножин X_j. Такий підхід при побудові пояснення дає можливість врахувати історію значень змінних для отримання результату в інтелектуальній системі.

Вираз (11) свідчить про те, що рівень довіри можливого каузального зв'язку визначається через мінімальне значення показника $\pi(x_{j,i})$ за умови, що значення $x_{j,i}$ було використано на поточному інтервалі T_j, тобто дане значення є актуальним.

Тоді для каузальної залежності необхідно, щоб рівень довіри для поточної підмножини значень X_j був більшим, ніж оцінка необхідності, тобто рівень довіри до всіх відомих підмножин. Відповідно, необхідна умова має вигляд:

$$(\forall i)\ \pi(x_{j,i})\geq N\left(\bigcap_j X_j\right). \tag{12}$$

З урахуванням необхідної (12) та достатньої (9) умов каузальності, можливісна модель каузальної залежності пов'язує результат Y інтелектуальної системи із значенням вхідної змінної $x_{j,i}$ таким чином:

$$C(x_{j,i},Y)=true$$
$$iff$$
$$1.(\forall i)\ \pi(x_{j,i})\geq N\left(\bigcap_j X_j\right), \tag{13}$$
$$2.\pi(x_{j,i})=\Pi(X_j).$$

Семантика представленої моделі полягає у виділенні максимально можливого (найбільш ймовірного) значення $x_{j,i}$ вхідної змінної за умови, довіри до змінної.

Остання полягає в тому, що найменша ймовірність вибору значення $x_{j,i}$ цієї змінної на поточному інтервалі T_j перевищує мінімальну ймовірності цього значення на попередніх інтервалах часу.

Наприклад, якщо значення

$$x_{j,2}=i7\_12\textit{покоління}$$

процесору має найбільшу ймовірність $\pi(x_{j,i})$ за умови, що для всіх інших значень цього процесору

$$x_{j,1}=i7\_13\textit{покоління},$$
$$x_{j,3}=i7\_11\textit{покоління},$$
$$x_{j,4}=i7\_10\textit{покоління}$$

поточна ймовірність вибору комп'ютеру вища, ніж на всіх попередніх інтервалах $T_{j-1},T_{j-2},...$

Висновки

Виконано структуризацію пояснення як елементу когнітивної діяльності людини. Показано, що пояснення може бути сформовано на основі порівняння вхідної інформації із існуючою системою знань людини, або ж на основі порівнянні інформації щодо властивостей вхідних об'єктів.

Запропоновано можливісні необхідна та достатня умови для каузальної залежності на базі однієї вхідної змінної, що лежить в основі пояснення.

Необхідна умова задає рівень довіри до залежності між вхідною змінною та отриманим результатом. Дана умова визначається на основі порівняння мінімальної ймовірності використання довільного значення вхідної змінної для отримання результату у поточній підмножині значень та у всіх можливих підмножинах значень.

Тобто необхідна умова показує, що дана змінна є суттєвою для отримання результату в інтелектуальній системі.

Достатня умова визначається через можливісну оцінку значення змінної. Тобто дана умова показує, що представлене значення є найбільш ймовірною умовою отримання результату в інтелектуальній системі.

Запропоновано модель можливісної каузальної залежності для побудови пояснення в інтелектуальній системі.

Дана залежність визначається через можливісні необхідну й достатню умови наявності каузального зв'язку, що дає можливість сформувати каузально-орієнтоване пояснення щодо зв'язку вхідної змінної і отриманого результату в умовах неповноти знань щодо стану інтелектуальної системи.

Подальший розвиток даного підходу пов'язаний із визначенням каузальної залежності для декількох вхідних змінних з тим, щоб отримати упорядковане представлення щодо впливу цих змінних на отриманий в інтелектуальній системі результат.

СПИСОК ЛІТЕРАТУРИ

1. Engelbrecht Andries P. *Computational Intelligence: An Introduction*. NJ: John Wiley & Sons, 2007. 632 p.
2. Castelvecchi D. (2016), "Can we open the black box of AI?" Nature, Vol. 538 (7623), pp. 20-23.
3. Miller T. (2019), "Explanation in artificial intelligence: Insights from the social sciences", Artificial Intelligence, vol. 267, pp.1-38, DOI: https://doi.org/10.1016/j.artint.2018.07.007
4. Q. Zhang, Y. Nian Wu, S.-C. Zhu, Interpretable convolutional neural networks, IEEE Conference on Computer Vision and Pattern Recognition. 2018. pp. 8827–8836.
5. Chalyi, S., Leshchynskyi, V., Leshchynska, I. (2019). Method of forming recommendations using temporal constraints in a

situation of cyclic cold start of the recommender system. EUREKA: Physics and Engineering, 4, 34–40. doi: https://doi.org/10.21303/2461-4262.2019.00952 . Available at: http://eu-jr.eu/engineering/article/view/952/934.
6. Adadi, A., Berrada, M. (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138– 52160.
7. Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, Finale Doshi-Velez. (2019) Human evaluation of models built for interpretability. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol.7, pp 59–67.
8. Gunning i D. Aha, (2019) "DARPA's Explainable Artificial Intelligence (XAI) Program", AI Magazine, Vol. 40(2), pp.44-58, doi: 10.1609/aimag.v40i2.2850.
9. Chalyi, S., Leshchynskyi, V. (2020). Method of constructing explanations for recommender systems based on the temporal dynamics of user preferences. EUREKA: Physics and Engineering, 3, 43-50. doi: 10.21303/2461-4262.2020.001228. Available at: http://journal.eu-jr.eu/engineering/article/view/14.
10. Чалий С. Ф. Реляційно-темпоральна модель набору сутностей предметної області для процесу формування рішення в інтелектуальній інформаційній системі / С. Ф. Чалий, В. О. Лещинський, І. О. Лещинська // Вісник Національного технічного університету "ХПІ". Сер. : Системний аналіз, управління та інформаційні технології = Bulletin of the National Technical University "KhPI". Ser. : System analysis, control and information technology : зб. наук. пр. – Харків : НТУ "ХПІ", 2022. – № 1 (7). – С. 84-89.
11. 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.
12. Halpern J. Y., Pearl J. Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*. 2005. № 56 (4). P. 843-887.
13. 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.
14. Чалий С. Ф., Лещинський В. О., Лещинська І. О. (2021) Контрфактуальна темпоральна модель причинно-наслідкових зв'язків для побудови пояснень в інтелектуальних системах,/ Вісник Національного технічного університету "ХПІ". Сер. : Системний аналіз, управління та інформаційні технології = Bulletin of the National Technical University "KhPI". Ser. : System analysis, control and information technology : зб. наук. пр. – Харків : НТУ "ХПІ", № 2 (6), С. 41-46.
15. Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. Psychological Review, 92(3), 289–316. https://doi.org/10.1037/0033-295X.92.3.289
16. Carey, S. (1985). Conceptual change in childhood. Cambridge, MA: MIT Press.
17. Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 21–59). Cambridge University Press. https://doi.org/10.1017/CBO9780511529863.004
18. Thagard, P. (2006). Evaluating explanations in science, law, and everyday life. *Current* Directions in Psychological Science, 15, 141–145.
19. Chin-Parker S, Bradner A. A contrastive account of explanation generation. Psychon Bull Rev. 2017 Oct;24(5):1387-1397. doi: 10.3758/s13423-017-1349-x. PMID: 28762030.

Received (Надійшла) 12.06.2023
Accepted for publication (Прийнята до друку) 23.08.2023

A possibility-based model of causal relation for input variable in explanation construction within an intelligent system

Serhii Chalyi, Volodymyr Leshchynskyi

Abstract. The **article's subject matter** is the processes of constructing explanations for the decisions made by an intelligent information system. The **goal** is to build a model of causal relationships for explanation construction under conditions of uncertainty regarding the states of the intelligent information system, especially when it is considered as a black box. **The tasks**: structuring explanations considering the specifics of human cognitive activity; establishing necessary and sufficient conditions for causal dependence as a component of explanations using possibility theory; developing a possibility model of causal dependence for a single input variable that considers the uncertainty regarding the states of the intelligent system. **The used approaches**: approaches to explanation construction in human cognitive activity and approaches to explanation construction in explainable artificial intelligence. The obtained **results** are as follows: explanations have been structured as an element of human cognitive activity. It has been demonstrated that explanations can be represented in two aspects: conceptual, through comparing input information with the existing human knowledge system; interpretive, through comparing the properties of input objects. Possibility-based necessary and sufficient conditions for causal dependence based on a single input variable, which forms the basis of explanation, have been proposed. A possibility model of causal dependence for explanation construction in an intelligent system has been suggested. **Conclusions**. The scientific novelty of the obtained results lies in the following: a possibility model of causal dependence between an input variable and the outcome of an intelligent system's operation has been proposed, which combines the necessary condition of causality in the form of confidence level in the impact of the input variable on the outcome and the sufficient condition of causality in the form of the maximum possible influence of the input variable's value on the outcome of the intelligent system. The model enables the formation of causal-oriented explanations based on the connection between the input variable and the obtained result in conditions of incomplete knowledge regarding the state of the intelligent system.

Keywords: intelligent information system, explanation, causal relationship, causal dependence, cognitive activity.