УДК 004. 891.3 **doi: 10.26906/SUNZ.2023.2.165**

С. Ф. Чалий, В. О. Лещинський

Харківський національний університет радіоелектроніки, Харків, Україна

# ОЦІНКА ЧУТЛИВОСТІ ПОЯСНЕНЬ В ІНТЕЛЕКТУАЛЬНІЙ ІНФОРМАЦІЙНІЙ СИСТЕМІ

**Анотація. Предметом** вивчення в статті є процеси побудови пояснень щодо отриманих рішень в інтелектуальній інформаційній системі. **Метою** є оцінка чутливості пояснень на основі аналізу властивостей вхідних даних та відповідних рішень в інтелектуальній інформаційній системі для підтримки вибору найкращого пояснення з позицій задоволення інтересів користувача. **Завдання**: структуризація критеріїв кількісної оцінки пояснень при представленні інтелектуальної системи у вигляді чорного ящику; розробка методу оцінки чутливості пояснень в інтелектуальній інформаційній системі. Використовуваними **підходами** є: підходи до побудови пояснень, підходи оцінки пояснень в інтелектуальних інформаційних системах. Отримані наступні **результати**. Структуровано критерії оцінки пояснень для інтелектуальних систем, представлених згідно принципу чорного ящику. Вказані критерії враховують вплив на пояснення вхідних та вихідних даних інтелектуальної системи, відповідність пояснення процесу прийняття рішення в інтелектуальній системі, а також відповідність пояснення і розуміння результатів інтелектуальної системи користувачем. На базі виконаної структуризації запропоновано метод оцінки чутливості пояснень для інтелектуальної системи, представленої згідно принципу чорного ящика. **Висновки**. Наукова новизна отриманих результатів полягає в наступному. Запропоновано метод оцінки чутливості пояснень для інтелектуальної системи, представленої згідно принципу чорного ящика. Метод містить етапи, пов'язані із перевіркою та визначенням схожості вхідних даних і результатів для альтернативних моделей інтелектуальних систем за кількісними та якісними показниками, а також кількісної оцінки вхідних даних та визначення чутливості пояснення. Запропонований метод дає можливість порівняти та вибрати пояснення з урахуванням властивостей та важливості вхідних даних з тим, щоб визначити можливості застосування альтернативних підходів до побудови пояснень щодо результатів інтелектуальної інформаційної системи. Подальший розвиток запропонованого підходу орієнтований на визначення і реалізацію метрик оцінки точності і прозорості пояснень.

**Ключові слова:** інтелектуальна система, пояснення, процес прийняття рішення, причинно-наслідковий зв'язок, оцінка пояснень.

## Вступ

Важливість використання пояснень щодо отриманих в інтелектуальних системах рішень пов'язана із суттєвим підвищенням складності таких систем і використанням складних, непрозорих для користувача алгоритмів при отриманні результатів в цих системах.

Незрозумілість алгоритмів прийняття рішень для користувача може привести до недовіри щодо отриманих результатів і, як наслідок, до невикористання або затримок використання запропонованого інтелектуальною системою рішення. Така невідповідність на практиці може приводити до суттєвих збитків. Наприклад, недовіра користувача до результатів розпізнавання об'єктів в охоронній системі знижує безпекові можливсті відповідного підприємства, недовіра до рекомендацій в системі електронної комерції знижує кількість та суму покупок товарів або послуг, тощо [1-3]. Для вирішення цієї проблеми використовується механізм пояснення отриманих в інтелектуальній системі результатів [4]. Пояснення дає можливість користувачеві визначити причинно-наслідкові залежності між вхідними даними і результатом роботи інтелектуальної системи, або ж між окремими діями і отриманими проміжними результатами [5].

В першому випадку при побудові пояснень робота інтелектуальної системи представляється як задача, вирішення якої має вхідні дані і результат, й не представлено проміжними станами. Тобто інтелектуальна система розглядається згідно принципу чорного ящику.

В другому випадку функціонування системи розглядається як процес, який містить проміжні стани. Кожен стан є результатом виконання однієї з дій процесу прийняття рішення. Тобто в другому випадку функціонування інтелектуальної інформаційної системи розглядається на різних рівнях деталізації, у відповідності до принципу білого ящику.

Згідно принципів білого та чорного ящиків можна виділити два напрямки побудови пояснень.

За принципом білого ящику, пояснення є безпосередньо однією із функцій інтелектуальної системи. В даному випадку для функції пояснення є доступними всі внутрішні дані системи, що забезпечують формування рішення. Перевага підходу полягає у можливості формування точних пояснень на основі повного набору внутрішніх даних. Суттєвим обмеженням даного підходу є необхідність включення функціональності пояснення безпосередньо на етапі проектування інтелектуальної системи. Доповнення існуючої системи вказаною функціональністю потребує значних витрат на її перепроектування.

Побудова пояснень за принципом чорного ящику передбачає розробку окремої підсистеми, яка використовує вхідні данні, результат і, можливо, доступні проміжні дані для формування пояснень. На базі цих даних для користувача формуються відповідні каузальні залежності.

Основні методи побудови пояснень розроблялись згідно програми XAI (Explainable Artificial intelligence), запропонованої DARPA [6]. Ключова увага за цією програмою приділялась визначенню зрозумілих для користувача причинно-наслідкових зв'язків, що

 

відображають залежності між входами та виходом інтелектуальної системи. Однак проблемі порівняння та обґрунтованого вибору методу побудови пояснень при представленні інтелектуальної системи у вигляді чорного ящику не приділялось достатньо уваги. Для практичного застосування методів побудови пояснень важливо мати критерій їх порівняння. Такий критерій має давати можливість порівняти альтернативних підходів та визначити, який метод точніше визначає залежності для пояснень між характеристиками вхідних даних та отриманим в інтелектуальній системі кінцевим або проміжними результатами.

Таким чином, актуальною задачею при побудові пояснень є формування та використання критеріїв для порівняння їх ефективності при однакових вхідних даних, а також отриманих в інтелектуальній інформаційній системі рішеннях.

Однак існуючі підходи до оцінки ефективності пояснень в інтелектуальних системах орієнтовані в першу чергу на якісне їх порівняння з точки зору людей-користувачів [7]. Також, з використанням принципу білого ящику, використовуються окремі, локальні підходи до перевірки залежностей між властивостями вхідних даних та отриманим в інтелектуальній системі рішенням [8], враховуються особливості моделі прийняття рішень та відповідні властивості вхідних даних інтелектуальної [9].

Таким чином, обґрунтування вибору методу побудови пояснень на основі кількісного критерію оцінки їх ефективності є актуальною задачею.

Такий критерій має враховувати узагальнені властивості пояснень, які забезпечують прозорість останніх для користувача. Ці властивості в першу чергу мають враховувати відповідність пояснення набору вхідних даних, тобто чутливість пояснення до помилок і змінах в таких даних. Також необхідно врахувати зрозумілість отриманого пояснення для користувача, та відповідність пояснення отриманим результатам.

**Метою статті** є оцінка чутливості пояснень на основі аналізу властивостей вхідних даних та відповідних рішень в інтелектуальній інформаційній системі для підтримки вибору найкращого пояснення з позицій задоволення інтересів користувача.

Для досягнення поставленої мети вирішуються такі задачі:

– структуризація критеріїв кількісної оцінки пояснень при представленні інтелектуальної системи у вигляді чорного ящику;

– розробка методу оцінки чутливості пояснень в інтелектуальній інформаційній системі, яка представлена як чорний ящик.

## Критерії оцінки пояснень

Для реалізації кількісної оцінки пояснень доцільно врахувати такі їх загальні властивості:

– чутливість до змін;

– точність;

– прозорість представлення користувачеві.

Обґрунтування вибору вказаних властивостей полягає в наступному. Відповідно до першої властивості, пояснення має, з одного боку, змінюватись при зміні процесу функціонування інтелектуальної системи, а з іншого -бути нечутливим до випадкових або раптових змін у вхідних даних, що не впливають на кінцевий результат роботи системи [10].

Згідно другої властивості, пояснення має максимально точно відображати процес роботи інтелектуальної системи, у загальному вигляді представленої як чорний ящик.

Третя властивість враховує відомі обмеження оперативної пам'яті людини, яка у більшості випадків забезпечує ефективне оперування сімома предметами одночасно. Тому пояснення має бути мінімальним з позицій кількості об'єктів, з якими воно оперує.

Перелік особливостей наведених властивостей пояснень представлено в табл. 1.

*Таблиця 1* – **Відмінності базових властивостей пояснень**

| Властивість | Особливості | Якісна/кількісна оцінка |
|---|---|---|
| Чутливість до змін | Уточнення пояснення відповідно до розвитку інтелектуальної системи, нечутливість до випадкових змін у вхідних даних внаслідок зовнішніх впливів | Якісна та кількісна (з урахуванням властивостей даних та результатів) |
| Точність | Відповідність пояснення реальному процесу функціонування інтелектуальної системи: каузальні залежності пояснення мають відповідати каузальним залежностям процесу прийняття рішення | Кількісна |
| Прозорість представлення користувачеві | Представлення з мінімальною складністю, з урахуванням обмежень сприйняття людини-користувача | Кількісна, за представленими елементами |

Згідно підходу «чорного ящика», пояснення $Expl$ в інтелектуальній системі представляється як функція, яка задає каузальні залежності між заданим набором вхідних даних $V = \{v_i\}$ та відповіддю системи $R$ за умови, що ці дані характеризуються їх якісними властивостями $a_i$ та кількісними властивостями $w_i$:

$$Expl = f(V, R). \quad (1)$$

$$v_i = (a_i, w_i). \quad (2)$$

Таке визначення пояснення як функції показує, що пояснення може бути представлено у вигляді залежності між властивостями елементів набору вхідних даних $v_i$ та отриманим в інтелектуальній системі результатом. Вхідні дані можуть мати як якісні, так і кількісні властивості. Наприклад, при якісній оцінці ми можемо виділити підмножину вхідних даних, яка визначає результат роботи інтелектуальної системи. Кількісна оцінка задає вплив цих даних на результат через їх вагу. Іншими словами, кількісна оцінка може бути пов'язана з якісною через апріорно задане

порогове значення ваги. Наприклад, вплив на результат елемента вхідних даних вважається суттєвим у випадку, якщо його вага перевищує задане порогове значення *w**. Таким чином, якісна оцінка може бути отримана на основі кількісної оцінки властивостей вхідних даних. Для вибору пояснень нам необхідно оцінити їх кількісно й якісно. При такій оцінці необхідно враховувати суб'єктивний та об'єктивний аспекти пояснень. Суб'єктивний аспект пояснення пов'язаний із якісною оцінкою. По-перше, необхідно отримати оцінку задоволеності користувачів цим поясненням. А по друге - ступінь розуміння пояснення користувачами. Структуризацію якісного аспекту пояснень представлено в табл. 2.

*Таблиця 2* – **Якісна оцінка пояснень**

| Характеристика | Особливості | Можливість переходу до кількісної оцінки |
|---|---|---|
| Задоволеність користувачів | Фактично відображає ступінь довіри користувачів до результатів інтелектуальної інформаційної системи | Через підтвердження вибору запропонованого системою рішення, наприклад рекомендованого товару або послуги |
| Розуміння | Відображення процесу прийнятті рішення в інтелектуальній системі у вигляді множини причинно-наслідкових зв'язків | Через аналіз текстів відгуків на результати використання системи (за наявності), через анкетування |

Однак якісна характеристика не дає можливості порівнювати пояснення між собою, оскільки при виборі краще пояснення повинно мати кращу кількісну оцінку. Можливість вибору забезпечує удосконалення підходів до побудови пояснень. Таким чином, об'єктивна характеристика пояснення базується на його кількісні оцінці. Наявність такої оцінки дає можливість перевірити, чи відповідає процес пояснення заданим значенням узагальнених властивостей чутливості до змін, точності та прозорості представлення користувачеві. Властивість точності дає можливість при побудові пояснення визначити підмножину вхідних даних, які значною мірою впливають на отриманий в інтелектуальній системі результат. Тоді встановлення залежності між цією підмножиною властивостей та рішенням системи забезпечує умови для встановлення причинно-наслідкового зв'язку як основи пояснення. Альтернативний підхід полягає у визначенні зв'язку між зваженою сумою властивостей та отриманим результатом. В подальшому, на основі уточнення ваг, можна виділити підмножину даних, що мають властивості з найбільшими значеннями ваг. Такі дані є визначальними для отримання рішення в інтелектуальній системі. Таким чином, з використанням результатів даного підходу також можна отримати підмножину вхідних даних, що суттєво впливають на результат системи.

## Метод оцінки чутливості пояснень

Визначимо кількісну оцінку чутливості пояснення на основі такого підходу.

Нехай маємо дві схожі моделі роботи інтелектуальної системи $M_1$ та $M_2$, які вирішують одну й ту ж задачу. При використанні цими моделями схожих вхідних даних $V_1$ та $V_2$ й отриманні схожих або однакових результатів $R_1$ та $R_2$ пояснення для користувача $Expl_1$ та $Expl_2$ мають бути схожими:

$$Expl_1 \approx Expl_2 \,|\, V_1 \approx V_2 \Rightarrow R_1 = R_2 = R. \qquad (3)$$

Тоді кількісна чутливість пояснення $S$ визначається через порівняння двох схожих (або однакових) пояснень $S(Expl_1, Expl_2)$, для яких виконується умова (3), з використанням метрики $D$:

$$S(Expl_1, Expl_2) = D\big(M_1(V_1, R), M_2(V_2, R)\big). \qquad (4)$$

Схожість між вхідними даними визначимо через врахування відмінності між нормованими значеннями ваг $w_{1,i}$ та $w_{2,i}$ відповідних властивостей вхідних даних $v_{1,i}$ та $v_{2,i}$:

$$\Delta w_i = \left| w_{1,i,} - w_{2,i} \right|, 0 \le w_{k,i} \le 1. \qquad (5)$$

Відповідно до (5), показник $\Delta V$ схожості/відмінності вхідних даних для двох схожих моделей $M_1$ та $M_2$ прийняття рішень в інтелектуальній системі можна визначити як усереднену різницю між вагами властивостей вхідних даних за умови, що ці дані є схожими в якісному і кількісному вимірах:

$$\Delta V_{1.2} = \frac{1}{I}\sum_{i=1}^{I} \Delta w_i \,|\, (\forall i)\exists a_{1,i}, \exists a_{2,i}. \qquad (6)$$

Схожість у якісному вимірі полягає в тому, що в обох випадках використовуються однакові набори вхідних даних:

$$(\forall i)\, a_{1,i} = a_{2,i}. \qquad (7)$$

Згідно виразу (7), властивості $v_i$ можуть, наприклад, мати однакові назви $a_i$. Схожість у кількісному вимірі полягає у незначній, в межах заданого порогу $\varepsilon$, відмінності між вагами вхідних даних:

$$(\forall i)\Delta w_i \le \varepsilon. \qquad (8)$$

Наприклад, в рекомендаційній системі в якості вхідних даних може бути використана інформація про покупки товарів, про їх рейтинги, про дату покупок та виставлення рейтингу товарів, про групи товарів, про їх популярність, демографічна інформація щодо користувача, тощо. Очевидно, що ми не можемо порівнювати пояснення для систем, перша з яких використовує для побудови рекомендацій інформацію про покупки, а інша – про рейтинги. Така ситуація свідчить про невиконання умови (7) щодо схожості вхідних даних у якісному вимірі.

Також важко порівнювати пояснення у системах, перша з яких враховує в першу чергу дані про товари, а друга – дані про користувачів (з відповід-

ною вагою), оскільки може не виконуватись умова (8). Тобто в першій системі будуть мінімальні ваги даних щодо користувачів і максимальні – щодо товарів, а в другій – навпаки. Схожість результатів $R = \{r_j\}$, отриманих в інтелектуальній системі, обчислюється лише у якісному вимірі аналогічно (7):

$$(R_1 = R_2) \equiv (\forall i) r_{1,j} = r_{2,j}. \quad (9)$$

Кількісний вимір елементів рішення є необов'язковим. Наприклад, він може бути відсутнім при вирішенні задачі бінарної класифікації зображень. У випадку, якщо рішення містить кількісну складову $\omega_i$, то визначення еквівалентності рішень є аналогічним (8):

$$R_1 = R_2 \left| (\forall j) r_{1,i} = r_{2,i}, \Delta\omega_j \le \varepsilon. \right. \quad (10)$$

Якщо ж відхилення кількісної складової перевищує поріг $\varepsilon$, то відмінність рішень $\Delta R$ є такою:

$$\Delta R = \sum_{j=1}^{J} \Delta\omega_J. \quad (11)$$

У відповідності до виразів (4) - (11), відхилення вхідних даних та результату є нормованими і тому чутливість пояснення може бути розрахована через співвідношення відхилень результату роботи інтелектуальної системи та її вхідних даних за умови однакових або дуже близьких пояснень $Expl_1 = Expl_2$ :

$$S_2^{(1)} = \left| \frac{\Delta V_1}{\Delta R_1} - \frac{\Delta V_2}{\Delta R_2} \right|. \quad (12)$$

При відсутності кількісної складової рішення згідно (10) вираз (12) спрощується і приймає вигляд:

$$S_2^{(1)} = |\Delta V_1 - \Delta V_2|. \quad (13)$$

Таким чином, для розрахунку чутливості пояснення необхідно визначити відмінності між складовими рішення та між вхідними даними за умови еквівалентності або схожості пояснень. Перші можуть бути отримані з урахуванням доступної із рішення кількісної складової. Для того, щоб отримати кількісну складову для вхідних даних, необхідно знайти латентні фактори, що пов'язують вхідні дані та отримані результати. Постановка цієї задачі має такий вигляд.

Дано розріджену матрицю розміру $I \times J \times N$, де $N$ - кількість значень вхідних даних. В інших вимірах матриця задає елементи вхідних даних, та елементи результату, отриманого в інтелектуальній системі. Оскільки система представлена чорним ящиком, то відомо лише, який результат був для заданих значень вхідних даних.

Ілюстрація даної задачі для одного елементу вхідних даних і одного елементу результату представлена на рис. 1. Стовпчики матриці задають можливі значення результату, а рядки – можливі значення вхідних даних. Хрестики задають відповідність – для яких вхідних даних був отриманий результат.

Потрібно знайти ваги впливу вхідних даних на рішення системи. Наприклад, для рекомендаційної системи на вході маємо перелік користувачів, а на виході – рекомендацію з одного товару. Маємо знайти і представити у кількісному вимірі зв'язок між користувачем та рекомендованим товаром.

Задача у даній інтерпретації є аналогічною задачі побудови рекомендацій і може бути вирішена відповідними методами. Так, виявлення латентних факторів забезпечується на основі використання методу колаборативної фільтрації. Тому для побудови пояснень на одному із етапів методу оцінки чутливості пояснень доцільно використати колаборативну фільтрацію.


**Рис. 1.** Вхідні залежності для виявлення ваг вхідних даних для оцінки пояснення

Запропонований метод оцінки чутливості містить у собі такі етапи.

Етап 1. Перевірка схожості вхідних даних за якісним критерієм (7). У випадку несхожості виконання методу завершується, оскільки невідповідність вхідних даних не дає можливість оцінити пояснення.

Етап 2. Перевірка схожості результатів за якісним критерієм (9). У випадку несхожості виконання методу завершується.

Етап 3. Визначення схожості результатів за кількісним критерієм (11). Етап виконується за наявності кількісних даних щодо результату робот інтелектуальної системи.

Етап 4. Формування матриці вхідних даних для методу колаборативної фільтрації. Матриця формується аналогічно прикладу на рис. 1, але має додаткові виміри, які відображають множину вхідних даних та складових результату.

Етап 5. Виявлення ваг вхідних даних за допомогою колаборативної фільтрації.

Етап 6. Визначення чутливості пояснень згідно (12) у випадку кількісної складової результату або згідно (13) у випадку відсутності такої складової.

Експериментальну перевірку методу виконано на наборі даних, що містить інформацію про продажі подарункових товарів у мережі супермаркетів Лондона. Пояснення були сформовані на основі темпоральних залежностей, що відображають динаміку користувацького попиту на вказані товари.

Були відібрані товари з однаковими поясненнями (ріст попиту та продажів у визначений період). Використовувалась інформація про продажі товарів, групована по тижням за період. Чутливість становила від 72 до 79 відсотків в залежності від періоду часу

покупок товарів. Аналіз ваг показав, що ключовим вхідним параметром є період часу.

Таким чином, експериментальна перевірка показала можливість обґрунтованого переходу від кількісної до якісної оцінки чутливості пояснення в інтелектуальній системі.

## Висновки

Структуровано критерії оцінки пояснень для інтелектуальних систем, представлених згідно принципу чорного ящику. Ці критерії враховують можливість побудови релевантного пояснення на основі вхідних та вихідних даних інтелектуальної системи (чутливість), відповідність пояснення і процесу роботи інтелектуальної системи (точність), а також ступінь сприйняття пояснення користувачем (прозорість). Сукупність цих критеріїв дає можливість комплексно оцінити весь процес побудови пояснення, що створює умови для раціонального вибору пояснення з урахуванням вхідних даних, процесу прийняття рішення в інтелектуальній системі, а також вимог користувача.

Запропоновано метод оцінки чутливості пояснень для інтелектуальної системи, представленої згідно принципу чорного ящика. Метод містить етапи перевірки та визначення схожості вхідних даних і результатів для альтернативних моделей інтелектуальних систем, а також кількісної оцінки вхідних даних і, на цій основі, визначення чутливості пояснення.

В практичному плані запропонований метод дає можливість порівняти пояснення з урахуванням властивостей вхідних даних і, на цій основі, визначити сферу застосування альтернативних підходів до побудови пояснень щодо роботи інтелектуальної інформаційної системи.

Подальший розвиток запропонованого підходу пов'язаний із визначенням метрик для оцінки точності і прозорості пояснень для користувача.

REFERENCES


1. Castelvecchi D. (2016), “Can we open the black box of AI?” Nature, Vol. 538 (7623), pp. 20-23.
2. Tintarev N., Masthoff J. (2012), “Evaluating the effectiveness of explanations for recommender systems”, User Model User-Adap Inter., Vol. 22, pp. 399– 439, https://doi.org/10.1007/s11257-011-9117-5.
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. Adadi, A., Berrada, M. (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138– 52160.
5. 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.
6. 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.
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. Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, Phil Blunsom.( 2019) Can I trust the explainer? Verifying post-hoc explanatory methods. arXiv:1910.02065.
9. Mengjiao Yang and Been Kim (2019). BIM: Towards quantitative evaluation of interpretability methodswith ground truth. arXiv:1907.09701.
10. Peter Lipton. Inference to the best explanation.Routledge, 2003.


Received (Надійшла) 11.04.2023
Accepted for publication (Прийнята до друку) 07.06.2023

**Evaluation of the sensitivity of explanations in the intelligent information system**

S. Chalyi, V. Leshchynskyi

**Abstract.** The **article’s subject matter** is the process of constructing explanations for the received decisions in the intellectual information system. The **goal** is to evaluate the sensitivity of the explanations based on the analysis of the properties of the input data and the corresponding decisions in the intelligent information system to support the selection of the best explanation from the standpoint of satisfying the user's interests. **Task:** structuring of criteria for quantitative assessment of explanations when presenting an intelligent system in the form of a black box; development of a method for assessing the sensitivity of explanations in an intelligent information system. The used approaches are approaches to constructing explanations and approaches to evaluating explanations in intelligent information systems. The following results were obtained. Criteria for evaluating explanations for intelligent systems presented according to the black box principle are structured. The specified criteria consider the impact on the explanation of the input and output data of the intelligent system, the appropriateness of the explanation of the decision-making process in the intelligent system, as well as the appropriateness of the explanation and understanding of the results of the intelligent system by the user. On the basis of the performed structuring, a method for assessing the sensitivity of explanations for an intelligent system presented according to the black box principle is proposed. Conclusions. The scientific novelty of the obtained results is as follows. A method for assessing the sensitivity of explanations for an intelligent system presented according to the black box principle is proposed. The method includes steps related to testing and determining the similarity of input data and results for alternative models of intelligent systems by quantitative and qualitative indicators, as well as quantifying the input data and determining the sensitivity of the explanation. The proposed method makes it possible to compare and select explanations considering the properties and importance of input data in order to determine the possibilities of applying alternative approaches to constructing explanations for the results of an intelligent information system. Further development of the proposed approach is focused on the definition and implementation of metrics for assessing the accuracy and transparency of explanations.

**Keywords:** intellectual system, explanation, decision-making process, causal relationship, evaluation of explanations.