UDC 004.8:004.9 doi: https://doi.org/10.20998/2522-9052.2023.4.10

Serhii Chalyi, Volodymyr Leshchynskyi

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

POSSIBLE EVALUATION OF THE CORRECTNESS OF EXPLANATIONS TO THE END USER IN AN ARTIFICIAL INTELLIGENCE SYSTEM

Abstract. The **subject** of this paper is the process of evaluation of explanations in an artificial intelligence system. The aim is to develop a method for forming a possible evaluation of the correctness of explanations for the end user in an artificial intelligence system. The evaluation of the correctness of explanations makes it possible to increase the user's confidence in the solution of an artificial intelligence system and, as a result, to create conditions for the effective use of this solution. Aims: to structure explanations according to the user's needs; to develop an indicator of the correctness of explanations using the theory of possibilities; to develop a method for evaluating the correctness of explanations using the possibilities approach. **The approaches used are** a set-theoretic approach to describe the elements of explanations in an artificial intelligence system; a possibility approach to provide a representation of the criterion for evaluating explanations in an intelligent system; a probabilistic approach to describe the probabilistic component of the evaluation of explanations. The following **results** are obtained. The explanations are structured according to the needs of the user. It is shown that the explanation of the decision process is used by specialists in the development of intelligent systems. Such an explanation represents a complete or partial sequence of steps to derive a decision in an artificial intelligence system. End users mostly use explanations of the result presented by an intelligent system. Such explanations usually define the relationship between the values of input variables and the resulting prediction. The article discusses the requirements for evaluating explanations, considering the needs of internal and external users of an artificial intelligence system. It is shown that it is advisable to use explanation fidelity evaluation for specialists in the development of such systems, and explanation correctness evaluation for external users. An explanation correctness assessment is proposed that uses the necessity indicator in the theory of possibilities. A method for evaluation of explanation fidelity is developed. **Conclusions**. The scientific novelty of the obtained results is as follows. A possible method for assessing the correctness of an explanation in an artificial intelligence system using the indicators of possibility and necessity is proposed. The method calculates the necessity of using the target value of the input variable in the explanation, taking into account the possibility of choosing alternative values of the variables, which makes it possible to ensure that the target value of the input variable is necessary for the explanation and that the explanation is correct.

Keywords: intellectual system; explanation; decision-making process; temporality; causality.

Introduction

Explainable AI (XAI) is a research area focused on building transparent and user-friendly artificial intelligence systems. XAI focuses on solving the "black box" problem in representing intelligent systems. This problem is associated with the use of machine learning models where the decision-making process is opaque and difficult for the user to interpret [1]. XAI allows users to get an idea of the decision-making algorithms at a certain level of detail, and to understand the reasons for those decisions. This increases confidence in the results of intelligent systems [2].

The concept of explainability is used in computer science, mathematics, physics, engineering and psychology. The concept of "explainability" is different from the concept of "interpretability". Interpretability is a property of a prediction algorithm in an intelligent system that makes that algorithm directly understandable to the user [3].

Explainability is an acquired property of the decision process, usually implemented by external means. Interpretability reveals the internal structure of a machine learning model. The interpretation itself usually requires knowledge in the field of artificial intelligence. Unlike interpretability, explainability is primarily focused on external users of an AI system. A system with this property should explicitly present to the user the factors that had the greatest impact on the resulting decision. Such factors can be related to both input data and actions to form a decision in an intelligent system. The system then becomes comprehensible to the user [4].

Trust is key to increasing user confidence in system decisions and ensuring their effective use [5]. In addition, the formation of cause-and-effect relationships to explain the system's decision simplifies the perception of an AI system, making it more anthropomorphic. Explanations make the cause-and-effect relationships between input data and the model's solution explicit by presenting these relationships in a form that is obvious to end users [6-8].

Research in Explanatory Artificial Intelligence has been conducted in recent years under the DARPA program [9]. This program addressed the challenges of understanding the psychology of explanation, developing methods for constructing explanations, and developing methods for evaluating explanations.

Existing approaches to the evaluation of explanations do not pay enough attention to the different requirements of users of intelligent systems. The approaches developed focus mainly on determining the impact of input data on decisions when the intelligent system is represented as a black box, and on assessing the perception of explanations by users [10-12]. However, user confidence in the decision-making process and the results of an artificial intelligence system depends on the correctness of the decisions made, which should be represented as a binary score. The approaches developed to evaluate the correctness of decisions (numerical evaluation) determine the deviation of the result in case of significant deviations in the input data [12]. However, these approaches are mainly focused on image processing and do not take into account differences in input data, e.g., for recommender systems. However, such an evaluation can be generalized using the Theory of Possibilities [13], which allows the use of input data to be described in a probabilistic form, regardless of the type of data.

This indicates the relevance of the task of developing a possible evaluation of the correctness of an explanation for the user of an artificial intelligence system.

The purpose of the article is to develop a method for forming a possible assessment of the correctness of explanations for the end user in an artificial intelligence system.

The evaluation of the correctness of explanations makes it possible to increase the user's confidence in the AI solution and, as a result, to create conditions for the effective use of this solution.

To achieve this goal, the following tasks are solved:

- Structuring explanations according to user needs;

-Developing an explanation correctness indicator using the Theory of Possibilities;

- Developing a method for evaluating the correctness of explanations using the possibility approach.

User-needed structuring of explanations

Explanatory artificial intelligence is currently one of the key concepts in the development of intelligent systems, because the following factors are essential for the evaluation of explanations.

- The need to justify the decision of the artificial intelligence system;

- The importance of presenting the decision model in a "transparent" form;

- Improving the accuracy of the decisions of artificial intelligence systems;

- Identifying new knowledge about decision making.

Justifying a decision with an explanation makes the model underlying an AI system understandable and transparent. This enables internal users of the system (e.g., data scientists or developers) to identify potential shortcomings. The system can then be debugged and optimized based on the problems and opportunities identified, improving the accuracy of its decisions.

These factors reflect the impact of explanations on the internal mechanism of an artificial intelligence system. Such explanations are essential for users involved in the development and improvement of an intelligent system.

External users who use the system to solve practical problems should receive explanations in order to gain new knowledge about the reasons for the formation of a solution and the specifics of its use.

Taking into account the different needs of internal and external users, explanations should reveal the reasons for actions in the decision-making process and the reasons for the result obtained (or traceability and reconstructive explanations [14]).

Explanations of the decision-making process are intended for specialists in the development of an artificial intelligence system. Such explanations reflect the results implemented in the model on which the AI system is based.

Explanations of the result are intended for users who are directly using the system. Such explanations usually show the impact of input features on the resulting prediction.

For example, an explanation that highlights the characteristics of a runner in an image that justify classifying this image as a "person running". In this example, the difference between the developer's explanation and the user's explanation is that the model was able to analyze other features that were not essential to the classification and did not affect the result. Information about the stages of analysis of additional features is included in the explanation of the decision-making process in an intelligent system.

In other words, the difference between explanations of the type "What happened in the AI system?" and "Why did the AI system get this result?" is the use of different models. In the first case, the explanation should use a decision model (or one that is close to it in terms of accuracy).

In the second case, a simplified model of the AI system is used, reflecting only the key factors that influence the system's decisions. For example, decision trees, inference rules, etc.

Explanations for developers and users therefore have different requirements for accuracy and justification of the solution. Explanations for developers should be more detailed and consider the impact of both important and unimportant factors.

In other words, these explanations should provide greater accuracy and present the AI system as a "white" or "grey" box.

Explanations for users should reflect the main reasons for the decision and ensure the identification of new knowledge for the practical application of these decisions [15].

The general scheme for using explanations, taking into account the differences between external and internal users, is shown in the Fig. 1.

Method for the evaluation of the correctness of explanations for the end user

The users of explanations in AI systems are the external and internal specialists who develop and use these systems. These users can be divided into two groups:

- Specialists in artificial intelligence systems;

- Specialists in the domain in which such a system is used.

The first group includes: owners of artificial intelligence systems; system developers; decision model developers.

The second group includes: Experts in the field; End users of an intelligent system; representatives of regulatory bodies.

The users of the first group should receive an explanation of how the system works at different levels of detail.

Owners determine the capabilities and modes of operation of the AI system.

The explanation for such users should be at a high level and provide information on the general principles of the AI system. For example, an explanation of the solutions used in practice and the dynamics of using the system to support the owners' decisions on managing the system's development, financial costs, etc.

Developers work at the level of the system as a whole, including the interface, the database or knowledge base, and the intelligent core. Developers are usually not experts in the domain in which the system is used. Developers use explanations to resolve bugs in the system. In this case, building explanations takes into account the entire process of obtaining and processing data, including filtering out erroneous input data.

Decision model developers build and debug the intellectual core of the system. The core contains a debugged model that performs classification or prediction.

Users in this category select the type of model, build and train the model. These users do not need to have detailed knowledge of the domain. Users in the second group require explanation of the domain.

Users - subject matter experts - use explanations to certify that the system meets the requirements of practical applications. Experts usually do not have the experience and knowledge to build a machine learning model.

The user of an AI system uses it to solve practical problems in their business. Such a user is not necessarily an expert in the domain but has a basic knowledge system of the tasks he is solving. Such users usually have no knowledge of the construction and operation of artificial intelligence systems. The end user uses explanations to trust the resulting solution.

A regulator uses explanations to ensure that the AI solution meets regulatory requirements. In particular, if the data for the machine learning model in the intellectual core of the system is biased, the system's decisions will contain this bias and therefore may not meet regulatory requirements. For example, a recruitment system trained on biased data may reject qualified applicants for reasons that do not affect their performance (such as age or gender).

Thus, the explanation based on the user classification above is provided in two aspects: system and user. The main differences between these aspects of user-centred explanation are shown in the Tabl. 1 below.

A systemic explanation describes how an AI system works. Such an explanation should reflect causal relationships at different levels of the decision-making process. In other words, the developer's explanation reveals the situations that arise in the decision-making process. This explanation is used to influence the functioning of the intelligent system.

The user's explanation should describe the reasons for the decision, the relevance of the decision to the user's practical needs, and compliance with the standards or requirements of the domain. The main difference in this explanation is that users are affected by the AI system and do not participate directly in its operation. In this case, XAI explanations should be accurate, understandable, and meaningful to people who are not experts in the subject area. They should give reasons for the actions of the intelligent system without using technical terminology.

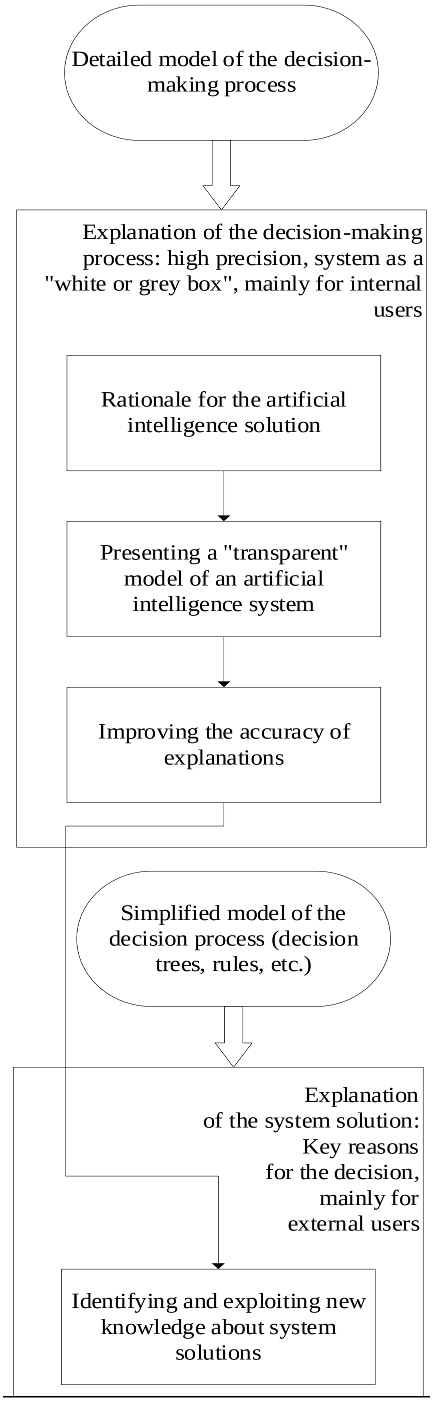
Detailed model of the decision-making process
Explanation of the decision-making process: high precision, system as a "white or grey box", mainly for internal users
Rationale for the artificial intelligence solution
Presenting a "transparent" model of an artificial intelligence system
Improving the accuracy of explanations
Simplified model of the decision process (decision trees, rules, etc.)
Explanation of the system solution: Key reasons for the decision, mainly for external users
Identifying and exploiting new knowledge about system solutions

Fig. 1. Differences in explanations for internal and external users

Table 1 – **Requirements for the evaluation of explanations in an artificial intelligence system**

Group	Users	Requirements
Internal users, specialists in AI systems, who modify the system to adjust the sensitivity and accuracy of the explanation.	Owners	A high-level explanation of the general principles of decision making in an AI system; an explanation of the dynamics of using the system.
	System developers	An explanation of the rationale for the overall decision-making process at a given level of detail, taking into account the system architecture.
	Decision model developers	An explanation of the decision model that makes it "transparent" to increase the accuracy of interpretation.
External users, subject matter experts who require/verify the accuracy of the explanation	Domain experts	An explanation that is consistent with domain knowledge
	End users	Explanation of the impact of inputs on the decision, use of counterfactuals
	Representatives from controlling organizations	Explanation based on counterfactuals to test regulatory constraints

Thus, when evaluating explanations, it is necessary to take into account the differences between the systemic and user aspects discussed. In the systemic aspect, the explanation reveals the internal mechanism of the intelligent system. Clarification of this mechanism may lead to changes in the sensitivity and correctness of the explanation. In the user aspect, the key evaluation should be the correctness of the explanation. The proposed method for assessing the correctness, in contrast to the possible assessment of the sensitivity of an explanation, takes into account the need to select alternative input data to obtain an explanation in an artificial intelligence system. Let's consider the basic idea of the method on the example of an explanation for a recommender system.

Suppose a recommender system offers a user a laptop with certain characteristics (processor, memory, hard drive). The explanation cites the processor model $x_{i,j} = i7 - 1185G7$ as the main reason for the choice, which should satisfy the user in terms of price and power.

This model is part of the same company's range of processors X_i:

$$X_i = \left\{ x_{i,j} \right\}, X_i \subset X. \qquad (1)$$

The set X_i is part of the set X of all possible values of processors. The possibility of choosing a processor $\Pi\left(X_i \right)$ is defined by the probability of choosing a given model $x_{i,j}$ as $\max\limits_j \pi\left(x_{i,j} \right)$.

Then, the need to choose is calculated taking into account the possibility of choosing the processors of all other firms represented $\Pi\left(X \setminus X_i \right)$ in the recommendation system:

$$N\left(X_i \right) = 1 - \Pi\left(X \setminus X_i \right). \qquad (2)$$

The essence of expression (2) is that the need to choose a particular processor $i7 - 1185G7$ depends on how often users of the recommendation system have chosen processors of other firms.

That is, if there is at least one very popular processor of another company, the need to choose the recommended processor decreases.

From the above considerations, it is clear that the criterion for the correctness of the explanation is

$$N\left(X_i \right) > 0.5. \qquad (3)$$

According to (2), if the probability of choosing an alternative processor is less than 0.5, then the presented explanation for the target laptop processor is correct. That is, the laptop model is recommended precisely because of the popularity of the processor. The developed method consists of the following steps:

Step 1. Calculation of the probability $\pi\left(x_{i,j} \right)$ of using input values from the set X.

Step 2: Calculate the probability of choosing an alternative $\Pi\left(X \setminus X_i \right)$.

Step 3. Calculation of the need $N\left(X_i \right)$ and checking the condition (3). If the condition is fulfilled, the explanation is correct.

Conclusions

The explanations are structured according to the needs of the user. It is shown that the explanation of the decision process is used by specialists in the development of intelligent systems.

Such an explanation represents a complete or partial sequence of steps to derive a decision in an artificial intelligence system.

End users mostly use explanations of the result presented by the system. Such explanations usually define the relationship between the values of input variables and the resulting prediction.

The article discusses the requirements for evaluating explanations, taking into account the needs of internal and external users of an artificial intelligence system. It is shown that it is advisable to use explanation fidelity evaluation for specialists in the development of such systems, and explanation correctness evaluation for external users.

An explanation correctness assessment is proposed that uses the necessity indicator in the theory of possibilities.

A method for evaluating the correctness of explanations in an artificial intelligence system using the necessity indicator has been developed. The method makes it possible to take into account the importance of the value of the input variable included in the explanation in comparison with the probability of choosing alternative values of the variables. Such a comparison makes it possible to ensure that the target value of the input variable is necessary for the explanation, i.e. the explanation is correct.

References

1. Adadi, A. and Berrada, M. (2018), “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)”, *IEEE Access (Vol. 6)*, pp. 52138–52160, doi: http://dx.doi.org/10.1109/ACCESS.2018.2870052
2. 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
3. Hoa Khanh, Dam, Truyen, Tran and Aditya, Ghose (2018), “Explainable software analytics”, ICSE-NIER '18: *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*, ACM, Gothenburg, Sweden, pp. 53–56, doi: http://dx.doi.org/10.1145/3183399.3183424.
4. Alonso, J.M., Castiello, C. and Mencar, C. (2018), “A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field”, In: Medina, J., et al. Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations, IPMU 2018, *Communications in Computer and Information Science*, vol 853. Springer, Cham, doi: https://doi.org/10.1007/978-3-319-91473-2_1
5. Tintarev, N. and Masthoff, J. (2007), “A survey of explanations in recommender systems”, *IEEE 23rd Int. Conference on Data Engineering Workshop*, IEEE, Istanbul, Turkey, 2007, pp. 801–810, doi: http://dx.doi.org/10.1109/icdew. 2007.4401070

6. Chalyi, S., Leshchynskyi, V. and Leshchynska I. (2021), “Counterfactual temporal model of causal relationships for constructing explanations in intelligent systems”, *Bulletin of the National Technical University "KhPI", Ser. : System analysis, control and information technology*, National Technical University "KhPI", Kharkiv, no. 2(6), pp. 41–46, doi: https://doi.org/10.20998/2079-0023.2021.02.07
7. Gunning D. and Aha, D. (2019) “DARPA’s Explainable Artificial Intelligence (XAI) Program”, *AI Magazine*, Vol. 40(2), pp. 44-58, doi: https://doi.org/10.1609/aimag.v40i2.2850.
8. Tintarev, N. and Masthoff, J. (2012), “Evaluating the effectiveness of explanations for recommender systems”, *User Model User-Adap Inter*., Vol. 22, pp. 399– 439, doi: https://doi.org/10.1007/s11257-011-9117-5.
9. Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D.I. and Ravikumar, P.K. (2019), “On the (in)fidelity and sensitivity of explanations”, *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, pp. 10965–10976, available at: https://dl.acm.org/doi/abs/10.5555/3454287.3455271
10. Dubois, D. and Prade, H. (2015), “Possibility Theory and Its Applications: Where Do We Stand?”, *Mathware and Soft Comp. Magazine*, *Springer Handbook of Comp. Intel.*, vol. 18, pp. 31–60, doi: https://doi.org/10.1007/978-3-662-43505-2_3.
11. Chalyi, S., Leshchynskyi, V.and Leshchynska, I. (2022), “Relational-temporal model of set of substances of subject area for the process of solution formation in intellectual information systems”, *Bulletin of National Technical University "KhPI". Series: System Analysis, Control and Inf. Technologies*, No. 1 (7), pp. 84–89, doi: https://doi.org/10.20998/2079-0023.2022.01.14
12. Wick, M.R. (1993), “Second generation expert system explanation”, *Second Generation Expert Systems*, Springer, Berlin, Germany, 1993, pp. 614–640, doi: http://dx.doi.org/10.1007/978-3-642-77927-5_26
13. Alvarez-Melis, D. and Jaakkola T.S. (2018), “Towards robust interpretability with self-explaining neural networks”, *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Inc., Montréal, Canada, pp. 7786–7795, available at: https://dl.acm.org/doi/10.5555/3327757.3327875#d31467704e1
14. Chalyi, S. and Leshchynskyi, V. (2023), “Probabilistic counterfactual causal model for a single input variable in explainability task”, *Advanced Information Systems*, Vol. 7, No. 3, pp. 54–59, doi: https://doi.org/10.20998/2522-9052.2023.3.08
15. Chalyi, S. and Leshchynskyi, V. (2023), “Evaluation of the sensitivity of explanations in the intelligent information system”, *Control, navigation and communication systems*, Vol. 2, pp. 165-169, doi: https://doi.org/10.26906/SUNZ.2023.2.165

Received (Надійшла) 29.08.2023
Accepted for publication (Прийнята до друку) 22.11.2023

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Чалий Сергій Федорович – доктор технічних наук, професор, професор кафедри інформаційних управляючих систем, Харківський національний університет радіоелектроніки, Харків, Україна;
Serhii Chalyi – Doctor of Technical Sciences, Professor, Professor of Professor of Information Control Systems Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: serhii.chalyi@nure.ua; ORCID ID: http://orcid.org/0000-0002-9982-9091.

Лещинський Володимир Олександрович – кандидат технічних наук, доцент, доцент кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна;
Volodymyr Leshchynskyi – Candidate of Technical Sciences, Associate Professor, Associate Professor of Software Engineering Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: volodymyr.leshchynskyi@nure.ua; ORCID ID: http://orcid.org/0000-0002-8690-5702.

Можливісна оцінка коректності пояснень для кінцевого користувача в системі штучного інтелекту

С. Ф. Чалий, В. О. Лещинський

Предметом вивчення в статті є процес оцінки пояснення в системі штучного інтелекту. **Метою** є розробка методу формування можливісної оцінки коректності пояснень для кінцевого користувача в системі штучного інтелекту. Оцінювання коректності пояснень дає можливість підвищити довіру користувача до рішення системи штучного інтелекту і, як наслідок, створити умови для ефективного використання даного рішення. **Завдання**: структуризація пояснень за потребами користувачів; розробка показника коректності пояснення з використанням теорії можливостей; розробка методу оцінки коректності пояснень з використанням можливісного підходу. Використовуваними **підходами** є: теоретико-множинний підхід, який застосовується для опису елементів пояснення в системі штучного інтелекту; можливісний підхід, який забезпечує представлення критерію оцінки пояснень в інтелектуальній системі; ймовірнісний підхід для опису ймовірнісної складової оцінки пояснення. Отримані наступні **результати**. Виконано структуризацію пояснень згідно потреб користувача. Показано, що для спеціалістів з розробки інтелектуальних систем використовується пояснення щодо процесу прийняття рішення. Таке пояснення представляє повну або часткову послідовність кроків з виводу рішення в системі штучного інтелекту. Кінцеві користувачі переважно використовують пояснення щодо результату, представленого інтелектуальною системою. Такі пояснення зазвичай задають зв'язок між значеннями вхідних змінних та отриманим прогнозом. Обґрунтовано вимоги до оцінки пояснень з урахуванням потреб внутрішніх та зовнішніх користувачів системи штучного інтелекту. Показано, що для спеціалістів з розробки таких систем доцільно використовувати оцінку вірності пояснення, а для зовнішніх користувачів – оцінку коректності пояснення. Запропоновано оцінку коректності пояснення, яка використовує показник необхідності в теорії можливостей. Розроблено метод оцінки коректності пояснення. **Висновки**. Наукова новизна отриманих результатів полягає в наступному. Запропоновано можливісний метод оцінки коректності пояснення в системі штучного інтелекту, який використовує показники можливості та необхідності. Метод розраховує необхідність використання цільового значення вхідної змінної у складі пояснення з урахуванням можливості вибору альтернативних значень змінних, що дає можливість впевнитись, що саме цільове значення вхідної змінної є необхідним для пояснення, а пояснення є коректним.

Ключові слова: інтелектуальна система; пояснення; процес прийняття рішення; каузальність; причинність.