UDC 004.891.3 doi: https://doi.org/10.20998/2522-9052.2023.3.08

Serhii Chalyi, Volodymyr Leshchynskyi

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

PROBABILISTIC COUNTERFACTUAL CAUSAL MODEL FOR A SINGLE INPUT VARIABLE IN EXPLAINABILITY TASK

Abstract. The subject of research in this article is the process of constructing explanations in intelligent systems represented as black boxes. **The aim** is to develop a counterfactual causal model between the values of an input variable and the output of an artificial intelligence system, considering possible alternatives for different input variable values, as well as the probabilities of these alternatives. The goal is to explain the actual outcome of the system's operation to the user, along with potential changes in this outcome according to the user's requirements based on changes in the input variable value. The intelligent system is considered as a "black box." Therefore, this causal relationship is formed using possibility theory, which allows accounting for the uncertainty arising due to the incompleteness of information about changes in the states of the intelligent system in the decision-making process. The **tasks** involve: structuring the properties of a counterfactual explanation in the form of a causal dependency; formulating the task of building a potential counterfactual causal model for explanation; developing a possible counterfactual causal model. The employed approaches include: the set-theoretic approach, used to describe the components of the explanation construction process in intelligent systems; the logical approach, providing the representation of causal dependencies between input data and the system's decision. The following **results** were obtained. The structuring of counterfactual causal dependency was executed. A comprehensive task of constructing a counterfactual causal dependency was formulated as a set of subtasks aimed at establishing connections between causes and consequences based on minimizing discrepancies in input data values and deviations in the decisions of the intelligent system under conditions of incomplete information regarding the functioning process of the system. A potential counterfactual causal model for a single input variable was developed**.** **Conclusions**. The scientific novelty of the obtained results lies in the proposal of a potential counterfactual causal model for a single input variable. This model defines a set of alternative connections between the values of the input variable and the obtained result based on estimates of the possibility and necessity of using these variables to obtain a decision from the intelligent system. The model enables the formation of a set of dependencies that explain to the user the importance of input data values for achieving an acceptable decision for the user.

Keywords: artificial intelligence system; explanation; possibility; causality; cause-and-effect relationship.

Introduction

Modern artificial intelligence (AI) systems commonly employ non-transparent methods for task resolution. These systems utilize models trained on data samples describing the subject domain. Typically, these data reflect practical solutions to current tasks within the domain [1]. However, due to the nature of the learning process, resulting models often remain unclear to users. Users are unable to directly access information about the system's working algorithm or discern the reasons behind the AI's decisions.

To address this issue, explanations are implemented [2-5]. Explanations elucidate the causal relationships that led to specific decisions for the user. These explanations consider the interplay between input object properties in the subject domain, events depicting property changes, and the sequence of actions leading to a solution. Through explanations, users can evaluate the actions culminating in a particular outcome and accept or reject AI recommendations [6].

An effective explanation within an AI system should focus on crucial cause-and-effect connections relevant to a specific decision, omitting extraneous details. This approach reduces the multitude of possible dependencies presented to the user. Therefore, explanations can incorporate both primary, factual connections among subject domain events and alternative dependencies.

Counterfactual explanations aim to interpret an AI system's decision by contrasting current outcomes with potential alternatives [7]. In essence, this method reveals decisions by describing necessary input data modifications for obtaining different outcomes. For example, if a banking AI system denies a user's loan request, a counterfactual explanation identifies which application data (such as current income, credit score, borrower's assets) requires alteration to achieve loan approval.

Alternative scenarios encompass data that is conceptually plausible but deviates from the current state of the subject domain [8]. For instance, in the loan approval scenario, an alternative scenario might entail the counterfactual assertion that "if the borrower had chosen a different type of insurance, they would have saved 10% on insurance payments."

Since machine learning algorithms render AI systems as "black boxes," information regarding causal relationships during decision-making is often incomplete. Consequently, considering alternatives for counterfactual explanations takes place under conditions of uncertainty, encompassing intermediate states and subject domain events, as well as the decision-making process.

This underscores the relevance of constructing sets of alternatives: counterfactual cause-and-effect dependencies concerning decision-making processes within AI systems.

Creating such alternatives under uncertainty demands the application of a possibility approach, particularly considering the potential impact of alternative causes on AI decisions. This approach enables users to compare multiple AI system outcomes and adjust input data to attain desired alternative solutions, thereby enhancing the effectiveness of AI system.

The overarching approach to providing explanations is founded on interpreting causal relationships. These relationships can be represented in rule-based form [9].

Counterfactual causal dependencies accounting for temporal event sequences [10], as well as probabilistic aspects of causality for event chains, are considered in works [11, 12]. The approach presented in these works has certain limitations related to comparing factual and alternative pairs: input data and resulting decisions in the presence of data utilization for specific decisions. The proposed approach for addressing this limitation is detailed in works [9, 10]. This approach focuses on leveraging event properties for determining causal dependencies. Graph-based modeling is used for representing causal relationships [13]. In this approach, causes and effects are represented as graph nodes, while causal connections are graph edges. These dependencies incorporate probabilistic evaluations. However, when forming counterfactual causal dependencies, only boundary probability values are significant, indicating the potential for achieving alternative outcomes. Additionally, these dependencies possess a fuzzy nature as explanations rely on knowledge about differences between input event properties or similarity with user background knowledge.

The mentioned aspect signifies the significance of possible causality description while tackling the task of constructing explanations. The importance of such depiction lies in the ability to more precisely unveil the interconnections between cause and effect within the context of decision-making. This unveils opportunities for users to comprehend the influence of input data values on outcomes. Additionally, this facet holds substantial importance in crafting more reliable and accurate models, as incorporated causal dependencies foster better alignment of the decision-making model within the intelligent system to user needs. Such an approach marks a pivotal stride towards enhancing the quality of artificial intelligence systems' operations, thereby facilitating a more informed approach to task resolution.

The aim of the article is to develop a counterfactual causal relationship model between the values of the input variable and the output of an artificial intelligence system, considering possible alternatives for different values of the input variable, as well as the probability of these alternatives. It needs to explain the user the actual result of the system's operation, as well as possible changes in this result according to the user's requirements based on changes in the value of the input variable.

The intelligent system is considered as a "black box." Therefore, this causal relationship is formed using possibility theory, which allows accounting for the uncertainty arising due to the incompleteness of information about changes in the states of the intelligent system in the decision-making process.

To achieve this goal, the following tasks are addressed:

Structuring the properties of the counterfactual explanation in the form of a causal relationship.

Formulating the task of constructing a possible counterfactual causal relationship model for explanation.

Developing a possible counterfactual causal relationship model.

In solving the first task, the properties of the cause and effect of the causal relationship are determined, which form the basis of the explanation.

In addressing the second task, conditions are established that the cause and effect of the counterfactual explanation must meet, so that the user can ascertain which input data ensure the achievement of the intelligent system's target decision.

Solving the third task provides the opportunity to obtain a set of alternative possible relationships that reflect user-interesting results of the artificial intelligence system.

Structuring the properties of the counterfactual explanation in the form of a causal relationships

Counterfactual explanation is a way of explaining the output of an artificial intelligence system by showing how the input attributes could be changed to get a different desired result.

This method helps to understand the causal relationships between the input and the output of the artificial intelligence system. Counterfactual explanation focuses on a few attributes that have the most impact on the output, making it easier for the human user of the system to comprehend.

However, this method also has some limitations. One of them is that it may not provide a complete and accurate explanation of the output, because it considers alternative, non-existing values of the input variables at the current moment. It may ignore some important factors that affect the output, or it may not explain why those factors matter. This may lead to a lack of justification or confidence in the output of the system in some cases.

Therefore, to construct a counterfactual explanation, it is necessary to define constraints on the properties of its structural elements.

To evaluate a counterfactual explanation, it is prudent to consider the properties inherent to explanations of this nature, which characterize the cause and the outcome realized within the intelligent system.

The distinctiveness of counterfactual explanations regarding input data is entwined with accounting for the uncertainty regarding the state of the subject domain and the decision-making process in artificial intelligence systems, as well as the significance of employing alternative values of variables closely related to actual input data.

When selecting input data for explanation, it's crucial to utilize minimal deviation between the values of alternative and factual input data.

Explanations should incorporate the plausible nature of input data, rooted in the probability of their utilization in decision-making within artificial intelligence systems. The peculiarities of counterfactual explanations concerning the results obtained within a system are connected to the fact that, firstly, it must uncover the decisions across several distinct yet crucial aspects. Secondly, explanations are intended to enable users to achieve a target (or near-target) result with slight alterations in the input data of the intelligent system. The resulting counterfactual explanation should be multifaceted, enabling users to comprehensively analyze the reasons for both the obtained and desired decisions.

Explanations should ensure minimal deviation between the resulting counterfactual decision and the projected (desired for the user of the intelligent system) outcome. The culmination of the discussed characteristics of counterfactual explanations is presented in the table.

Let's consider examples of counterfactual explanations with the properties listed in the table within various domains: banking, recommendation systems, medicine, and intelligent management systems.

For instance, an explanation concerning the decision to reject a loan application at a bank indicates the reason as a low credit score of the borrower. Counterfactual explanation: to achieve the desired outcome (loan approval), the credit score should be increased using credit cards. In this case, the requirement to minimize the deviation of an alternative input variable from the actual value lies in determining the minimum score the user needs to reach for loan approval.

In the medical field, an explanation for a proposed diagnosis involves an imprecise value of the patient's age. Counterfactual dependency: specifying the accurate age might lead to a cancellation of the diagnosis. This example considers a deviation in a single variable – the patient's age.

In a recommendation system, a high-priced smartphone is recommended based on a high camera resolution. Counterfactual explanation: to meet a budget constraint, the requirements for camera quality need to be lowered to a specific resolution value.

Here, the scenario sets a minimal deviation of the AI system's output (the cost of the recommended smartphone) from the actual device cost (the consequence) through a minimal change in the input variable – camera resolution (the cause). It's important to note that this example results in multiple alternative outcomes, as several smartphones may fit the price constraint with the specified camera quality.

Another case, related to traffic management systems, involves the reason for delays on a route being the alignment of travel time with the most probable peak traffic period.

An alternative approach: changing the travel time to the evening or morning could reduce travel time. Selecting the best time of day involves determining a time interval with minimal probability of heavy traffic while adhering to constraints, linked to the acceptable deviation from the target arrival time compared to the actual one (particularly constrained by working hours).

Thus, in this example, minimal deviations in both input and output of the AI system are considered, alongside the boundary probabilities of using specific input values.

Overall, the provided examples illustrate the significance of using structural elements of causal explanations as presented in the Table 1. It's important to note that while constructing counterfactual explanations, as seen in the examples, boundary probabilities of using particular input and output variable values are employed.

Table 1 – **Structural Elements of Counterfactual Causal Explanation**

Structural elements	Requirements	Comment
The reason	Minimal deviation of alternative values from the actual values	Minimization of deviations is aimed at simplifying the transition from the actual solution to the target that represents value for the user of the intelligent system
	Using the limit values of the probabilities of using input data	The maximum and minimum values of probabilities for the values of the input data provide a comparison of alternatives within the framework of the theory of possibilities, which creates conditions for the construction of alternative causal relationships.
The consequence	Multi-alternativeness as a condition of agreement with knowledge of the subject area	The user can use one of the alternatives, which is consistent with his knowledge of the subject area
	Minimal deviation of the counterfactual decision from the actual one	Since the counterfactual solution is targeted to the user, the explanation should reveal changes in the input data that provide a result that is closest to the expected one.

This aspect enables the formalization of causal dependencies in counterfactual explanations using the theory of possibilities.

This theory utilizes boundary probabilities to assess the possibility of using those values, further integrating trustworthiness evaluation of the possibility measure.

The combined estimation of possibility and trustworthiness for alternative input variable values and outcomes can be employed to establish causal dependencies that form counterfactual explanations within artificial intelligence systems.

Possibility Counterfactual Causal Model

Based on the analysis of the structure of a counterfactual explanation, the results of which are presented in the table, we will formulate the task of constructing causal dependencies for such explanations.

Considering the properties of the cause and effect, this task can be divided into two subtasks.

Subtask 1: Minimization of the deviation of alternative values of the input variable from the actual ones, while achieving constraints on the target decision of the artificial intelligence system.

Subtask 2: Minimization of the deviation of the counterfactual result from the actual result on a given set of input data values.

According to the given formulation, the counterfactual causal model includes a set of alternative cause-and-effect relationships between the values of the input variable and the system's decision, with the following characteristics:

Minimal deviation from the specified constraint regarding the difference between the actual and alternative outcomes of the intelligent system; the actual outcomes reflect previous decision implementations based on a set of known input variable values.

The constraints define a set of target outcomes for the user of the intelligent system.

The minimal possible deviation across a subset of input variable values, which ensures minimal deviation from the constraints on the outcome.

Deviation in input data is considered based on the probability of using input variable and outcome values using indicators of possibility and necessity.

The last characteristic is associated with uncertainty regarding the components and dependencies of the decision-making process in the intelligent system. The key idea is to find the most probable values of the variable, the potential influence of which on the system's outcome is maximal. These potential input data values should ensure the target outcome with the highest degree of confidence.

It should be noted that the possibility index [14] allows determining the probability deviation of the impact of input variable values on the outcome. Comparing the possibility indices for different variable values helps select the value with the minimum deviation.

The necessity index [14] in a generalized manner determines the degree of confidence in the obtained dependency.

This index demonstrates confidence through minimal probability of deviation from the system's outcome constraints (or deviation from the actual result).

Let's consider a formal possible counterfactual causal model by a single variable according to the provided description.

The input variable X has a set of possible values $\{x_i\}$.

The resulting impact (usually, probability of impact) of the input variable values on the system's decision is determined by normalized assessments $\pi(x_i)$, which map each value x_i to $[0,1]$.

The set of values X includes subsets X_j. Each of these subsets consists of values of the variable that were used during the decision-making process of the intelligent system at moments $t_{j,i}$ within a certain time period T_j:

$$X = \left\{ X_j : \forall x_{j,i} \in X_j \exists t_{j,i} \in T_j \right\}. \quad (1)$$

Since the intelligent system, when making similar decisions at different time intervals, can use the same input data, identical values $x_{j,i}$ can be part of different subsets $X_m \neq X_j$.

The distribution of assessments $\pi(x_{j,i}) : x_{j,i} \in X_j$ is defined by an ordered set P_j:

$$P_j = \left\langle \begin{matrix} \pi(x_{j,1}),...,\pi(x_{j,I}) : \\ \pi(x_{j,1}) = \max\limits_i \pi(x_{j,i}) \end{matrix} \right\rangle. \quad (2)$$

The possibility Π_j of impact for any value $x_{j,i} \in X_j$ corresponds to the upper bound of this subset, meaning it can be defined as the maximum element of the subset:

$$\Pi_j = \max_i \pi(x_{j,i}). \quad (3)$$

The possibility assessment for several subsets, obviously, will be equal to the maximum element of the union of these subsets.

Similarly, the possibility assessment is defined for the output data of the intelligent system.

According to (3), minimizing the input deviations $\Delta_{j,m}^{j,i}$ between the actual value of variable $x_{j,i}$ and the alternative value $x_{j,m} \in X_j$ of in the counterfactual causal relationship $c_{q,k}^{j,m}$ explaining the result $y_k \in Y$, takes the form:

$$\begin{aligned} \Delta_{j,m}^{j,i} &= \min_m \left| \pi(x_{j,i}) - \pi(x_{j,m}) \right| \\ &\Big| \exists c_k^m, c_k^j . \end{aligned} \quad (4)$$

According to equation (4), the minimization of deviations for input occurs for two dependencies - the actual and the counterfactual, if they explain the same result y_k, or the result with minimal deviation from the actual.

The index $\Delta_{j,m}^{j,i}$ contains a normalized deviation assessment. Therefore, in general, the set of such indices $\Delta^{j,i} = \left\{ 1 - \Delta_{j,m}^{j,i} \right\}$ can be considered as a set of possibility assessments for using input data to construct counterfactual explanations.

Accordingly, the maximum element of this set determines this possibility. In other words, the maximum element defines the most possible counterfactual explanation.

Then, the counterfactual explanation C_k should contain an ordered set of alternative causal dependencies

c_k^m explaining the same result, sorted by the deviation values of input variables, and differing within the threshold value ε:

$$\Delta_{j,m+l}^{j,i} = \min_m \left|\pi\left(x_{j,i}\right) - \pi\left(x_{j,m+l}\right)\right| \\ \left| x_{j,m+l} \in X_j \setminus \left\{x_{j,m+1},...,x_{j,m+l-1}\right\}. \right. \quad (5)$$

Each subsequent deviation is calculated for the current subset of values from which elements with previous, smaller deviations have been excluded:

$$\Delta_{j,m+l}^{j,i} = \min_m \left|\pi\left(x_{j,i}\right) - \pi\left(x_{j,m+l}\right)\right| \\ \left| x_{j,m+l} \in X_j \setminus \left\{x_{j,m+1},...,x_{j,m+l-1}\right\}. \right. \quad (6)$$

The set of alternative causal dependencies (5) explains counterfactual results if for similar input data, the intelligent system proposed the same or a result close to the actual decision y_k.

Otherwise, if the information about decision similarity is inaccurate or the decisions are slightly different, the necessity index N from possibility theory is used. This index defines the value of trust for possible (practically realized) subsets of the intelligent artificial system's decisions:

$$N(Y) = \inf_m N\left(\bigcap_q Y_q\right) \\ \left| \bigcap_q Y_q \neq \varnothing, Y_q \subseteq Y. \right. \quad (7)$$

Then, the user should trust the counterfactual in the form of a limit, or a threshold value, or an acceptable deviation from the actual result of the intelligent system, in the case of similar or higher trust in the counterfactual compared to trust in the actual result. Such a comparison makes sense because the level of trust is based on the minimum probability of using a specific result.

$$\Delta_{q,k}^{q,i} = \max_m \left(\pi\left(y_{q,k}\right) - N(Y)\right). \quad (8)$$

According to (8), the decision closest to the user's needs will be the one whose possibility of implementation in the intelligent system significantly exceeds the trust level in the system's decisions as a whole.

The counterfactual causal model based on the possibility theory contains cause-and-effect dependencies that satisfy the requirements (4) and (8):

$$C_k = \left\langle \begin{array}{l} c_{q,k}^{j,m} \left| \exists \Delta_{j,m}^{j,i},...,c_{q,k}^{j,m+l} \right| \exists \Delta_{q,k}^{q,i}, \exists \Delta_{j,m+l}^{j,i} : \\ : \forall i,m,l\, \Delta_{j,m+l}^{j,i} \leq \varepsilon \end{array} \right\rangle. \quad (9)$$

This approach allows building possible causal dependencies without delving into the specifics of the subject area.

Conclusions

The structuring of the counterfactual causal dependency has been performed. It has been demonstrated that such dependencies are multivariate, involving minimal changes in input data compared to actual values, as well as slight adjustments to the outcome in order to satisfy constraints that were not met in the actual decision.

A comprehensive task of constructing the counterfactual causal dependency as a set of subtasks for establishing the link between causes and effects based on the minimization of deviations in input data and deviations in the intelligent system's decisions has been formulated.

This is carried out in conditions of incomplete information about the functioning process of the intelligent system.

A possible counterfactual causal model has been developed for a single input variable, which defines a set of alternative connections between the values of the input variable and the obtained outcome based on the assessments of possibility and necessity for using these variables to derive the intelligent system's decision.

This model enables the formation of a set of dependencies that explain to the user which values of input data are crucial for achieving an acceptable decision for the user.

References

1. 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.
2. Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D. and Rinzivillo, S. (2021), "Benchmarking and survey of explanation methods for black box models", *CoRR arXiv:2102.13076*, available at: https://arxiv.org/abs/2102.13076.
3. Chalyi, S. and Leshchynskyi, V. (2020), "Temporal representation of causality in the construction of explanations in intelligent systems," *Advanced Information Systems*, vol. 4, no. 3, pp. 113–117, doi: https://doi.org/10.20998/2522-9052.2020.3.16.
4. Chalyi, S. and Leshchynskyi, V. (2020), "Method of constructing explanations for recommender systems based on the temporal dynamics of user preferences", *EUREKA: Physics and Engineering*, vol. 3, pp. 43–50, doi: https://doi.org/10.21303/2461-4262.2020.001228.
5. 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.
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. Beck, S.R., Riggs, K.J. and Gorniak, S.L. (2009), "Relating developments in children's counterfactual thinking and executive functions", *Thinking & Reasoning*, vol. 15, is. 4, pp. 337–354, doi: https://doi.org/10.1080/13546780903135904.

8. Byrne, R.M.J. (2019), “Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning”, Kraus S. (ed), *Proceedings of the twenty-eighth international joint conference on artificial intelligence*, IJCAI 2019, Macao, China, August 10–16, 2019, pp 6276–6282, doi: https://doi.org/10.24963/ijcai.2019/876.
9. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D. and Lee, S. (2019), “Counterfactual visual explanations”, *Proceedings of the 36th international conference on machine learning*, ICML 2019, 9–15 June 2019, Long Beach, California, USA, PMLR, Proceedings of machine learning research, vol .97, pp. 2376–2384, https://doi.org/10.48550/arXiv.1904.07451.
10. Chalyi, S. and Leshchynskyi, V. (2020), “Temporal representation of causality in the construction of explanations in intelligent systems”, *Advanced Information Systems*, vol. 4, no. 3, pp. 113–117, doi: https://doi.org/10.20998/2522-9052.2020.3.16.
11. Pearl, J. (2009), “Causality: Models, Reasoning and Inference”, *Econometric Theory*, vol. 19, pp. 675–685, Cambridge University Press, USA, doi: https://doi.org/10+10170S0266466603004109.
12. Halpern, J. Y. and Pearl, J. (2005), “Causes and explanations: A structural-model approach. Part II: Explanations”, *The British Journal for the Philosophy of Science*, Vol. 56 (4), pp. 889–911, doi: https://doi.org/10.48550/arXiv.cs/0208034.
13. Lewis, D. (2000), “Causation as influence”, *Journal of Philosophy*, vol. 97, no. 4 (Special Issue: Causation), pp. 182–197, available at: https://www.jstor.org/stable/2678389.
14. Levykin, V. and Chala, O. (2018), “Development of a method of probabilistic inference of sequences of business process activities to support business process management”, *Eastern-European Journal of Enterprise Technologies*, No. 5/3(95), pp. 16-24, doi: https://doi.org/10.15587/1729-4061.2018.142664.
15. Dubois, Didier and Prade, Henri. (2015), “Possibility Theory and Its Applications: Where Do We Stand?”, *Mathware and Soft Computing Magazine, Springer Handbook of Computational Intelligence*, vol. 18, p. 31–60, doi: https://doi.org/10.1007/978-3-662-43505-2_3.

Received (Надійшла) 29.05.2023
Accepted for publication (Прийнята до друку) 30.08.2023

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 контрфактуальна модель каузальної залежності по одній вхідній змінній в задачі побудови пояснень

С. Ф. Чалий, В. О. Лещинський

Анотація. Предметом вивчення в статті є процес побудови пояснень в інтелектуальних системах, представлених як чорна скринька. **Метою** є розробка контрафактної моделі причинно-наслідкової залежності між значеннями вхідної змінної та виходом системи штучного інтелекту з урахуванням можливих альтернатив для різних значень вхідної змінної, а також ймовірності цих альтернатив з тим, щоб пояснити користувачеві фактичний результат роботи системи, а також можливі зміни цього результату згідно вимог користувача на основі зміни значення вхідної змінної. Інтелектуальна система розглядається як «чорний ящик». Тому дана каузальна залежність формується з використанням теорії можливості, що дозволяє врахувати невизначеність, що виникає внаслідок неповноти інформації щодо зміни станів інтелектуальної системи у процесі прийняття рішення. **Завдання**: структуризація властивостей контрфактичного пояснення у формі каузальної залежності; формулювання постановки задачі побудови можливісної контрфактичної моделі каузальної залежності для побудови пояснення; розробка можливісної контрфактичної моделі причинно-наслідкової залежності. Використовуваними **підходами** є: теоретико-множинний, який застосовується для опису складових процесу побудови пояснень в інтелектуальних системах; логічний підхід, який забезпечує представлення каузальних залежностей між вхідними даними та рішенням системи. Отримані наступні **результати**. Виконано структуризацію контрфактної каузальної залежності. Сформульовано комплексну задачу побудови контрфактичної каузальної залежності як сукупності підзадач побудови зв'язку між причинами та наслідками на основі мінімізації відхилень значень вхідних даних та відхилень рішення інтелектуальної системи в умовах неповноти інформації щодо процесу функціонування цієї системи. Розроблено можливісну контрфактичну модель каузальної залежності по одній вхідній змінній **Висновки**. Наукова новизна отриманих результатів полягає в наступному. Запропоновано можливісну контрфактичну модель каузальної залежності по одній вхідній змінній, яка задає множину альтернативних зав'язків між значеннями вхідної змінної та отриманим результатом на основі оцінок можливості та необхідності використання цих змінних для отримання рішення інтелектуальної системи. Модель дає можливість сформувати множину залежностей, що пояснюють користувачеві, які значення вхідних даних є важливими для досягнення прийнятного для користувача рішення.

Ключові слова: система штучного інтелекту; пояснення; можливість; каузальність; причинно-наслідковий зв'язок.