DOI: https://doi.org/10.30837/csitic52021232899

The Construction of Explanations in Intelligent Systems on the Basis of Interactive Clarification of the Decision's Reasons

Volodymyr Leshchynskyi[1]

Irina Leshchynska[2]

[1]Kharkiv National University of Radio Electronics, 14 Nauky Ave,
Kharkiv UA-61166, Ukraine, volodymyr.leshchynskyi@nure.ua

[2]Kharkiv National University of Radio Electronics, 14 Nauky Ave,
Kharkiv UA-61166, Ukraine, iryna.leshchynska@nure.ua

Abstract. *The problem of interactive formation of explanations in intelligent information systems, as a problem of interactive interpretation and substantiation of stages of decision-making is considered. A combined approach to constructing an explanation is proposed, which combines the interpretation of the process of the intelligent system and the explanation of the obtained result. The approach implements the process of interactive clarification of the reasons for the decision as a basis for explanation.*

Keywords: *explanation, interpretation, explainable intelligent information system.*

I. Introduction and Problem Statement

The concept of explainable artificial intelligence, which provides for the automated construction of explanations together with the formation of the solution of the intelligent system, has become widespread in recent years. The importance of the explained artificial intelligence is due to the influence of ethical problems on the practical use of solutions of intelligent information systems. Such problems arise in case of distrust of the user to the received decision. Distrust is a consequence of the opacity of algorithms and the mechanism of intelligent systems. For example, when using machine learning methods, the user may not always be able to justify the decision and, as a consequence, use this solution to solve practical problems.

This contradiction leads to the problem of interactive formation of explanations in intelligent information systems, as a problem of interactive interpretation and justification of the stages of decision-making. The explanation allows the user to understand the sequence of decision formation in the intelligent system, as it reflects the causal links between the input data, the actions of forming the result, and the resulting solution. The purpose of the explanation is the user's understanding of the process of functioning of the intelligent system, which makes this system "transparent" and, as a consequence, increases the user's confidence in the solution [1].

II. Problem Solution and Results

Solving the problem of interactive construction of explanations requires the formalization of the process of user interaction with the intelligent information system. When describing the sequence of human interaction with the intellectual system, it is advisable to distinguish two aspects. In the first case, the main attention is paid to the decision-making process as a sequence of actions that have causal links. In the second case, the final result is considered first of all in comparison with existing practices and achievements. The first aspect considers the possibility of constructing explanations using existing intelligent technologies. In this case, the main attention is paid to the interpretation of the sequence of actions of the intelligent system. The use of "process" explanation in full requires from the human user theoretical knowledge in the subject area and decision-making algorithms in intelligent systems. The second aspect considers the need to use explanations in terms of human-intellectual system interaction in a way that is similar to the ways of social interaction between people. In this case, general knowledge about the intermediate results or the obtained solution as a whole, as well as further comparison of the result with known to the user or widely used solutions are used to construct explanations. That is, the explanation in the first aspect is focused on the interpretation of decision-making technology, and in the second - to determine the key benefits of final and intermediate decisions from the standpoint of its practical application.

When constructing an explanation in the first aspect, the processes of forming an explanation and making decisions in an intelligent system are integrated. The same data is used for interpretation as for decision making. The result is a step-by-step interpretation of the decision-making process. The user of the intelligent system in this case must be a qualified specialist who understands the principles of operation of the artificial intelligence system he uses. For example, when interpreting the results of logical inference in the expert system, the user must understand the ideas of direct and inverse inference. He will then be able to determine the causal relationships between the input data and the result obtained by interpreting the sequence of rules in the output chain.

The explanation in the second aspect can be formed separately from the decision-making process. Additional data that differ from the data for decision-making in the intelligent system can be used for explanation. Structurally, the intelligent system should contain additional modules to create explanations [6]. The consumer in this case receives an explanation similar to the form of explanations of human behavior [7], considering the typical social attitudes and expectations. An important difference between this explanation is the use of only key causal relationships that characterize the decision-making process.

The proposed combined approach to the construction of explanations combines the interpretation of the process of the intelligent system and the direct explanation of the result. Interpretation and explanation can be presented at two levels: generalized and applied. At the generalized level, the explanation gives the user a detailed description of the current behavior of the intelligent system, which makes it possible to increase the latter's confidence in the decisions obtained. At this level, the explanation is based on formal, general theories that describe the decision-making process in an intelligent system. This explanation makes it possible to describe not only the current, context-sensitive patterns, but also the general principles on which the decision was made. At the applied level, the explanation explains the reasons for a particular decision, as well as the reasons for the events, individual facts, etc. that underlie this decision.

Building an explanation at the generalized level consists of the following steps.

Stage 1. Determining the set of possible reasons for decision-making as a whole or for individual stages of decision-making.

The explanation should contain a set of reasons for the decision in the intelligent information system.

Stage 2. Determining the reasons for the decision at the current stage, indicating their probability.

A generalized explanation of the result reveals the theoretical methods that were used in obtaining the solution. Therefore, for example, in the case of application of probabilistic methods, the deterministic explanation is supplemented by probabilistic indicators.

Stage 3. Arranging the reasons for the decision according to a certain evaluation indicator.

The generalized explanation of the result contains the whole set of reasons for the obtained decision, as well as a numerical assessment of the comparison of these reasons with the facts of the alternative solution.

Stage 4. Formation of differences between the proposed and alternative solutions.

The comparison of the proposed and alternative solutions in a generalized explanation is performed using formal criteria.

Building an explanation at the application level consists of the following steps.

Stage 1. Determining a subset of possible reasons for decision-making as a whole or for individual stages of decision-making.

The generalized explanation of the result uses the model of the subject area. This model identifies the key properties of the subject area that affect the decision-making process. Thus, the model of the subject area determines the context of perception of the result and may not fully meet the prejudices and expectations of the person receiving the explanation. Therefore, the explanation should contain only those facts that can be identified as reasons for the decision in the context of the perception of the recipient.

Stage 2. Selection of the most probable reasons of the received decision and representation of pair "causes-result" in the determined form.

The explanation identifies the reasons for the decision of the intelligent system, or the individual stages of the decision-making process, even in the case of using probabilistic methods to determine these reasons. Within the context of user perception, the explanation must specify a deterministic causal relationship between the input data and the solution of the intelligent system. The use of a probabilistic description at this stage increases the qualification requirements of the user, which generally does not increase the confidence of people who receive explanations.

Stage 4. Identification of key differences in the result obtained at each stage of decision-making in the intelligent system.

The explanations, as well as the reasons underlying it, should determine the differences between the obtained solution and a similar solution that is expected for the human user. Intelligent systems form solutions to partially structured and unstructured problems. That is, such solutions are not typical, such as those used in solving structured problems. Therefore, the comparison of a new solution with a typical one makes it possible to "embed" the reasons for the explanation in the context of the recipient's perception of such an explanation.

The generalized explanation of the result uses the model of the subject area. This model identifies the key properties of the subject area that affect the decision-making process. Thus, the model of the subject area determines the context of perception of the result and may not fully meet the prejudices and expectations of the person receiving the explanation. Therefore, the simplified interpretation should contain a limited number of reasons for the decision or the reasons for the events and facts underlying this decision.

Stage 2. Selection of the most probable reasons of the received decision and representation of pair "causes-result" in the determined form.

The explanation identifies the reasons for the decision of the intelligent system, or the individual stages of the decision-making process, even in the case of using probabilistic methods to determine these reasons. Within the context of user perception, the explanation must specify a deterministic causal relationship between the input data and the solution of the intelligent system. The use of a probabilistic description at this stage increases the qualification requirements of the user, which generally does not increase the confidence of people who receive explanations.

Stage 4. Identification of key differences in the result obtained at each stage of decision-making in the intelligent system.

The explanations, as well as the reasons underlying it, should determine the differences between the obtained solution and a similar solution that is expected for the human user. Intelligent systems form solutions to partially structured and unstructured problems. That is, such solutions are not typical, such as those used in solving structured problems. Therefore, the comparison of a new solution with a typical one makes it possible to "embed" the reasons for the explanation in the context of the recipient's perception of such an explanation.

III. Conclusions

The proposed approach to the construction of explanations implements the process of interactive clarification of the reasons for the decision, as well as the facts on which this decision is based. Each iteration of the explanation process identifies the key reasons for the current decision. If the current explanation detail satisfies the user, the explanation process ends. Alternatively, the facts – reasons for the decision are considered as the current decision. In the next iteration of the explanation process, the user receives the reasons for these facts.

REFERENCES

[1] J.E. Mercado, M.A. Rupp, J.Y. Chen, M.J. Barnes, D. Barber, K. Procci, “Intelligent agent transparency in human–agent teaming for multi-UxV manage-men”t, Hum. Factors, №58(3), pp. 401–415, 2016.

[2] P.Miller, L.`Howe, Sonenberg, “Explainable AI: beware of inmates running the asylum”, IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), pp.36–42, 2017/