Conversational XAI: Formalizing its Basic Design Principles

Marco Garofalo[1,3][0009−0005−9108−0038], Alessia Fantini[1,2][0009−0007−0337−2423], Roberto Pellugrini[4][0000−0003−3268−9271], Giovanni Pilato[2][0000−0002−6254−2249], Massimo Villari[3][0000−0001−9457−0677], and Fosca Giannotti[4][0000−0003−3099−3835]

[1] University of Pisa, Pisa, Italy
{marco.garofalo,alessia.fantini}@phd.unipi.it
[2] Institute for High Performance Computing and Networking National Research Council of Italy, Palermo, Italy
{alessia.fantini,giovanni.pilato}@icar.cnr.it
[3] University of Messina, Messina, Italy
{marco.garofalo,massimo.villari}@unime.it
[4] Scuola Normale Superiore, Pisa, Italy
{roberto.pellungrini,fosca.giannotti}@sns.it

Abstract. eXplainable Artificial Intelligence (XAI) aims to explain the predictions and operations performed by an AI model. Its goal is to make AI models more understandable to humans. However, XAI methods sometimes produce explanations in implementation-dependent formats and these artifacts may stimulate different perceptions in users with different backgrounds. *Conversational XAI systems* have been proposed to provide explanations in the form of conversation based on natural language. This new trend for XAI systems focused on a human-centered approach provides more powerful forms of explanation representation. In this study, we analyze the current state of the art of Conversational XAI systems and propose a general formalization based on currently available literature. Moreover, we devise a general Conversational XAI architecture that includes two new components designed to improve the user experience both functionally taking into account the recurrent questions and in terms of trustworthiness by explicitly providing metrics for the explanation.

Keywords: Explainable Artificial Intelligence · Conversational Interface · Human-Computer Interaction.

1 Introduction

Explainability is a key factor when it comes to trustworthy Artificial Intelligence (AI). eXplainable Artificial Intelligence (XAI) is an active field of research that addresses the problem of explaining decisions made by an AI model according to reliable criteria. As denoted in [5], there are mainly two phases in which one can intervene in requiring an explanation from the model: *(i)* use algorithms that produce models that are inherently interpretable [24]; *(ii)* query the model once trained, observing what output it provides in relation to any test sample, whether synthetically generated instances or from the original dataset. This second approach is also known as *post-hoc* explanation. It is easy to see that the most flexible approach concerns post-hoc methods since they often do not depend on the black-box model thus being model agnostic. In [10], the methods proposed for solving the problem of generating an explanation against a black-box model are widely described and depending on the type of problem, and the type of input data (tabular, images, or text), the visualization of the resulting explanation can take different forms [3,19,30]. However, as pointed out in [21], the visualization of the explanation is the result of the intuition of the researchers who developed the explainer without actually considering the perception of the end user. In this direction, several recent proposals in the literature place the end user at the center of the design process of the resulting explanation.

As emerged from interviews with experts in the field, such as physicians or researchers who regularly use XAI techniques in their workflow, practitioners prefer to interact more with the explanations provided by an XAI method rather than interpreting its visualization [2,15]. This indicates how a conversational experience fosters human-AI interactivity, abstracting the formalisms required to interrogate a trained model and thus laying the foundation for an experience based on transparency and trust. These insights motivate the adoption of Conversational XAI, an explainability technique that aims to simplify the exchange of prediction-related explanations between AI and humans by means of natural language. However, there is no clear and commonly accepted definition of what a Conversational XAI system is in the literature, so in this paper, we propose a general definition by formalizing the key components suitable for achieving a satisfactory conversational interaction from the end user's perspective. We analyzed the state of the art of Conversational XAI systems, starting with the rationale behind the concept. In order to identify commonly used components, we selected the currently available literature in which reproducible implementations are discussed. Finally, with the formalization we propose, we highlight the importance of two new components that we introduce to improve the user experience both functionally and in terms of trustworthiness.

The remainder of this study is structured as follows: we analyze the state of the art of Conversational XAI systems by following its evolution in Section 2. In Section 3 we highlight some problems related to the non-formal definition of conversational systems and then propose a possible general formulation. In Section 4 we discuss the contribution of our formalization and its possible impact on real conversational systems. Finally, in Section 5, we conclude the paper and give reasoning ideas for future scenarios.

2 Related Work

Conversational AI agents are the core element of Conversational XAI systems. These agents are designed to facilitate human-like interactions through natural language. According to [35], conversational agents can be categorized based on their use cases. First, there are rule-based agents that use a predefined set of rules to guide the conversation. They are particularly effective in scenarios where conversations focus on providing information or answering frequently asked questions. However, a major limitation of this approach is the finite size of the rule set. Second, generative agents use Natural Language Processing (NLP) techniques and generative models to understand user intent and generate natural language responses. They are commonly used in applications where conversational agents need to provide more engaging interactions, such as virtual assistants, social chatbots, and conversational games. One problem with this type of conversational agents is that of *out-of-scope intent*, where the user's input or query falls outside the defined scope or domain of the system's capabilities. An attempt is made to respond to this problem with out-of-scope intent classification, trying to identify queries that do not belong to any of the intents supported by the system. The classification of out-of-scope intents, however, is scarcely studied due to the lack of publicly available information, as pointed out by [17]. Finally, retrieval-based agents retrieve preexisting answers from a knowledge base or a collection of predefined answers based on user input. They excel at information retrieval tasks, similar to rule-based agents. However, retrieval-based agents are relatively simplistic in nature. They lack flexibility and the ability to generate new responses, as they solely rely on the information stored in the knowledge base. The basic architecture of a Conversational AI agent is based on three elements [12]: *(i)* the Natural Language Understanding (NLU) unit that must be able to classify user intentions from textual input and extract entities, i.e., discrete pieces of information within the sentence; *(ii)* the Dialogue Management System, which is the component responsible for managing the flow of the conversation. Typical functionalities of a Dialogue Management System include conversation state tracking and context management; finally, *(iii)* the Natural Language Generation (NLG), which is responsible for making the conversation more engaging through text generation based on user input and conversation history managed by the Dialog Management System.

While Conversational AI aims to create natural and interactive conversations between humans and AI agents, Conversational XAI goes a step further by addressing the need for explainability and interpretability. In [23], the explanation process is defined as a socio-cognitive process composed of two phases: *(a)* the cognitive phase in which an explanation is selected as reliable from a set of a priori identified causes; *(b)* the social phase in which knowledge of the explanation is actually transferred from the explainer to the human. It is the latter part that generally receives the least attention when building an XAI method, thus failing to establish trust between the model and the end user. Several proposals show the advantages of designing a conversational interface through the use of Wizard of Oz (WOz) prototypes [38], in which the real AI model has not been developed and is replaced by a real agent that emulates its behavior, "magically" executing requests sent to the system, hence the reference to the Wizard of Oz. In [2], more generally, WOz is proposed as a technique for collecting feedback from users during the use of the emulated AI model by subjecting users to different scenarios such as true positive or false positive classifications. In [11], WOz prototyping is used to observe human-robot interaction tailored to an explainability use case, where the "wizard" emulates the responses of an AI-powered robot, attempting to explain its functionality. As a result of this experiment, the need to have functional requirements, such as a NLU component that understands user intent related to explainability, is demonstrated, since the same intent can be formulated in multiple forms. Interesting work has been carried out in [15] by investigating the perception of current real-world explanation methods by domain experts. Key results show how domain experts are not fully satisfied with current XAI techniques and would prefer a more interactive approach with the explainer accompanied by metrics that evaluate the explanation. In addition, the authors identify best practices that may be considered when designing an interactive explanation system and present requirements for interactivity, explainability, and context management. A recent proposal [28] explores the impact of an extended model-agnostic XAI technique [29], called `Doctor XAI`, on a clinical decision support system (DSS). The purpose of `Doctor XAI` is to provide explanations for `Doctor AI`, a deep learning model for next visit diagnosis prediction. The authors conducted tests on the developed prototype involving healthcare experts, asking them to estimate the probability of a patient suffering from acute myocardial infarction based on his or her medical history, supported by the `Doctor AI` explanation system. The results show that practitioners generally tend to trust algorithmic suggestions if they are accompanied by an intelligible explanation, in this case through textual natural language. In [4] an explainable recommendation system addresses the problem of including incremental user feedback in the learning process. The proposed system calculates a score that identifies how much the user likes the object of the recommendation, generating a text-language explanation that motivates the prediction. The end user can provide feedback that is used to iteratively fine-tune the model. In [36] a tool named `TalkToModel`[5] has been proposed that tackles different aspects of user-model interaction through a conversational interface. Different figures such as data scientists, physicians, or generic end users are considered as target of the explanation, and the tool aims at generalizing the type of model to which explainers can be associated through a system of translation by sequence between sentences entered by the user and grammatical commands given to the `TalkToModel` system. The generalization comes at the cost of providing an initial dataset containing examples of translation between user inputs and commands used to train a Large Language Model (LLM) in charge of predicting the command. Similarly, in [27], a tool called `XAgent`[6] was developed under the assumption that the end user of the conversational system has no prior Machine Learning knowledge. In this work, the intents defined by the XAI question are mapped explicitly to XAI methods, e.g., questions related to the importance of features are answered using SHAP [19] as an XAI method, while requests for counterfactuals activate DICE [25]. The author extends a

[5] https://github.com/dylan-slack/TalkToModel

[6] https://github.com/bach1292/XAGENT

dataset of XAI queries collected from [16] by adding paraphrased versions of the same queries using GPT-3 [1]. In this approach, the text does not totally replace the explanation, contrary to other proposals, but it is used to complement a visualization produced by a method of XAI, e.g., a waterfall plot showing the importance of features accompanied by explanatory text. Several approaches are based on the derivation of a Conversational XAI system from a dialogue model often represented as a state machine in which the dialogue actors, i.e., the explainer and the human, communicate in a sequence of shifts that often have as their final state an evaluation of the explanation provided by the explainer [21,20,27]. In this configuration, the concept of conversation context proves to be necessary because the user might take up examples discussed earlier in the conversation. An explainer-agnostic system, called `ConvXAI`[7], has been proposed in [22] where natural language acts as a bridge between the explanation received from an XAI method and the end user. In `ConvXAI` the dialogue model contains three components: explanation, argumentation, and clarification extending the work previously done by [20]. The newly introduced clarification component handles any queries about details regarding explanation or argumentation with respect to the selected XAI method. The results of this approach show that the perception of an explanation depends on the end user and that support for the explanation, whether in the form of natural language or metrics, is interpreted as more reliable, enabling trust between humans and models. The state machine is not the only way to model a dialogue, recently the author of [42] proposes Behaviour Trees (BT) as a representation of a dialogue, justified by the granular level of definition of a single dialogue state, in order to implement conversational explanation experiences (EE).

In [13], the authors created an open explanation system using dialogue, specifically, they implemented a chatbot called `dr_ant`[8] that allows the recipient of the explanation to interact with a Machine Learning model and its explanations. In this experiment, the authors trained a random forest model that predicted the probability of survival on the Titanic dataset[9]. Users could dialogue with the model to understand the logic behind its predictions. The purpose of their study was twofold: first, to propose a prototype of a Conversational XAI system, and second, to find out what types of queries the end user asks to understand the model's decision-making process. In particular, the latter goal provided a better comprehension of how to meet the explanatory needs of a human operator. The authors conducted an analysis of the collected dialogues for each category, they calculated the number of conversations with at least one such question. For example, the `what-if` category includes questions such as "*How was it calculated/derived?*" while the `feature-importance` category covers questions such as "*What makes me most likely to survive?*". Regarding the system architecture, it consists of several components. Among these, there are a web-based Interface, Explainers, and a Dialog Agent with NLU and NLG components.

[7] https://github.com/Naviden/ConvXAI

[8] https://github.com/ModelOriented/xaibot

[9] https://www.kaggle.com/c/titanic/data

The authors note that users involved in the conversation seem to be more focused on the model's decisions rather than its metrics. This aspect indicates a preference for qualitative data and underscores the importance of an interactive experience: users seem more interested in understanding the answers provided and formulating hypotheses by actively interacting with the tool than in interpreting passive responses.

3 Methodology

Examination of the state of the art of Conversational XAI systems shows that there are several strategies for implementing this type of conversational interface. Much of the effort has been made to define dialogue patterns inspired by social interaction between human beings. However, because Conversational XAI is a fairly recent area of research, Conversational XAI systems have no commonly accepted definition. On the one hand, this allows some flexibility in designing the architecture of these systems; on the other hand, it is important to identify the key components that can lead to an effective user experience. Moreover, textual representation only satisfies some use cases of explainability, resulting in poor generalization by those implementations that do not take into account the need for different forms of explanation. For example, when the training dataset consists of images, many of the current state-of-the-art explainability techniques [6,39,37,3,34] are based on gradient extraction, which provide an almost immediate visualization at a glance. In this case, an explanatory text should *contextualize* the result of the XAI method and not replace it.
Observing the current state of the art for Conversation XAI, our proposal is a general formal definition that can accommodate all the fundamental components of these systems:

Definition 1. *Conversational XAI is a set of technologies, which use natural language as a means of communication between an explainer (XAI method) and an explainee (human). It is formally represented by Equation 1:*

$$C_{XAI} = \langle DS, b, E, R, M, D \rangle \tag{1}$$

where DS is a Dialog System, namely the conversational agent in charge of generating an answer based on the user's inputs; b is the black-box model we want to explain, $E = \{e_1, e_2, \ldots, e_n\}$*, with* $|E| \geq 1$*, is the set of explainers,* $R = \{r_1, r_2, \ldots, r_m\}$ *with* $|R| \geq 0$ *is the set of the user-defined routines,* $M = \{m_1, m_2, \ldots, m_z\}$ *with* $|M| \geq 1$ *is a set of explanation metrics,* $D = \{X, Y\}$ *is the original dataset with which b was trained, with X denoting the data points and Y denoting the labels. Note that the dataset can also be inaccessible once the model has been deployed; in this case,* $D = \emptyset$*.*

We do not place any restrictions on the type of dialog system because there are several ways to implement it [20,22,42]. However, we note that to be defined as conversational, a system must understand the user's intent and generate an appropriate response in textual natural language. In this sense, the former functionality is generally performed by an NLU component, while the latter is provided by an NLG or a predefined template-filling mechanism. Note that we have defined a set of explainers E because many of the current XAI techniques depend on the type of data on which the AI models are trained, so good generalization is possible by supporting different explainers. In addition, having different explainers applicable to the same type of data allows one to evaluate the explanations from different perspectives: for example, to check the importance of features, one could compare the explanation generated by SHAP [19] with that generated by LIME [32]. Our formalization (Equation 1) introduces two new elements with respect to the implementations proposed in the currently available literature:

i) *the routines set R*, aimed at improving the user experience by automating the execution of several recurring questions in the form of routines. Such component will provide a unique input to the resulting XAI Conversational system in the execution of several tasks. To illustrate the concept of *routine*, consider a scenario in healthcare. A physician using an AI model at the diagnosis stage can make use of Conversational XAI systems to interact with the AI model in a transparent way, and obtain explanations about the prediction of the diagnosis. During the interaction with the Conversational XAI system, the physician can ask preliminary questions common to each diagnosis. The ability to aggregate them into a routine means that for each diagnosis there is no need to annotate the different common preliminary questions but only the routine identifier as the input prompt. As a result, the input preliminary questions are less prone to human errors and the interaction is not unnecessarily long and repetitive.

ii) *the metric set M*, which represents the measures to evaluate the explanation. The end user will be provided with all the available elements to evaluate the reliability of an explanation generated by an XAI method, including state-of-the-art metrics, such as *fidelity* or methods of *insertion* and *deletion* [9,26]. Validating explanations with a metric means having an indicator parameter for the reliability of the explanation. In this way, the end user visualizes, as the response of the Conversational XAI system, the prediction of the AI model, the explanation versus the prediction, and the reliability of the explanation.

Finally, we have made explicit the original training dataset D as an optional component of the Conversational XAI system. This is because some of the proposed implementations assume that they have the dataset available [36,27], but in many applications, explaining a black-box model does not necessarily imply having access to the source dataset and thus in the formalization it is represented as optional. Figure 1 summarizes the proposed formalization in diagram form.

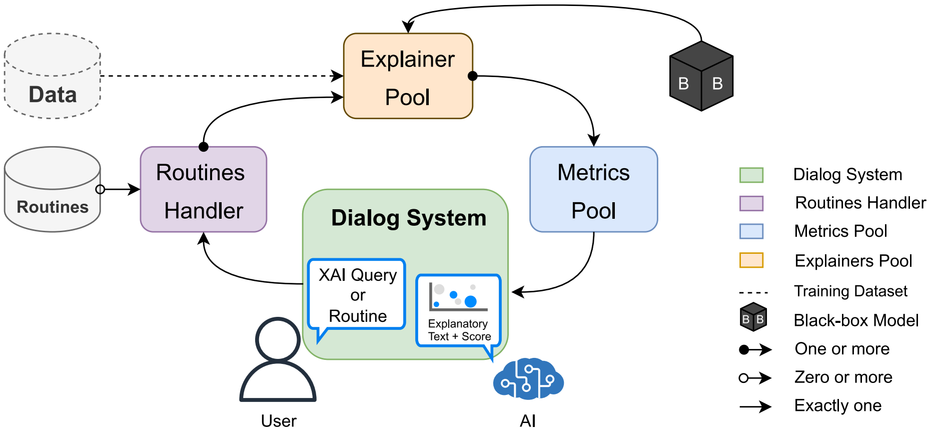

Fig. 1. Full representation of a general Conversational XAI system according to our proposed formalization.

4 Discussion

As pointed out in Section 2, natural language is generally preferred as a means of communication between humans and AI. Therefore, we propose a general formalization that a Conversational XAI architecture can implement. In our proposal, we combine the components necessary to translate an explanation generated by XAI methods into a form understandable by any type of end user, regardless of their background.

Our formalization remains flexible with respect to different implementations since it does not refer to specific dialogue patterns studied in the literature. Furthermore, we call it realistic since we openly consider the possibility of not having the training dataset available, as might happen in real scenarios. We introduced two components aimed at improving the user experience both functionally, by adding the concept of XAI routines to automate recurring tasks, and in terms of trustworthiness, as we believe that every explanation produced by XAI methods should indicate its degree of reliability, thus giving the user all the elements to make his or her own assessments. As an example, consider a data scientist who needs to debug a black-box model and resorts to XAI methods. He might use the method in the traditional way or define XAI routines that can be reused in different contexts and thus with different black-box models. To verify that our formulation is general, we apply it to the Conversational XAI models considered in our literature review, whose differences are summarized in Table 1, extending the respective architectural diagrams of each tool.

Table 1. Comparison of selected proposals. From left to right: Conversational XAI System; technology used for NLU; conversational context support; technology used for NLG; metric calculation on explanation; supported explainers.

Method	NLU	Context	NLG	Metrics	Explainers
Kuźba et al. [13]	Dialogflow[10]	✓	Dialogflow		CeterisParibus [14] iBreakDown [7]
Slack et al. [36]	T5[31]	✓	Template Filling	✓	LIME [32] SHAP[19] DiCE[25] PDP[8]
Nguyen et al. [27]	RoBERTa[18] + NN	✓	Template Filling		LIME [32] SHAP[19] DiCE[25] Proto[40] CFProto[40] Anchor[33]
Malandri et al. [22]	RASA[11]	✓	Template Filling		LIME [32] SHAP[19] FoilTree[41]

ConvXAI We have represented the original architecture of `ConvXAI` [22] according to our formalization in Figure 2. Although `ConvXAI` introduces the concept of clarification, it refers to a possible state the conversation may be in, and its function is based on an element of disambiguation of the conversation itself. Differently, in our formalization (Equation 1), we identify the set of metrics as the set of tools that allow us to calculate a reliability score of the explanation to be shown to the end user. `ConvXAI` also provides a good user experience through context management, allowing the user to make references to elements that emerged backward during the conversation. We propose a further step, external to the dialog system, adding the set of routines. Finally, in the original `ConvXAI` architecture, explicit reference is made to the dataset used to train the black-box model. In our conception of the Conversational XAI system, however, the training dataset is treated as an uncertain element, since it is not always present in all real-world scenarios.

TalkToModel Figure 3 shows how we can model a typical flow described in [36] by adding the contribution of our formalization. Even though `TalkToModel` introduces the *fudge score* used to select the most faithful explanation among those generated by different methods, this metric is not shown explicitly to the

[10] https://cloud.google.com/dialogflow

[11] https://rasa.com/

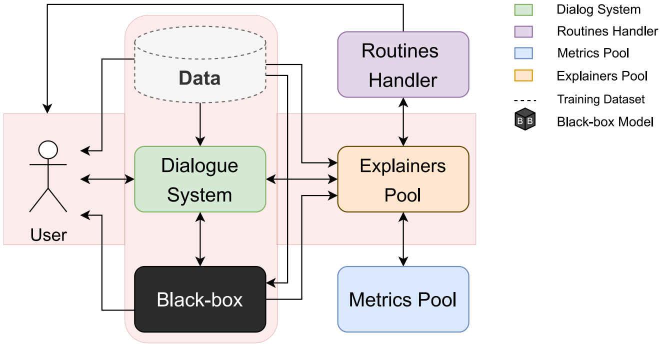

Fig. 2. Extended architecture of `ConvXAI`. The architecture proposed in [22] is highlighted within the pink area, while our extension is represented by the components `Routines Handler` and `Metrics Pool`. In addition, the `Data` component is represented with dashed lines to symbolize its optionality.

end user, whereas in our proposal we encourage the visualization of an explanation reliability score alongside the explanation itself. `TalkToModel` offers not only XAI-related operations but also those related to the dataset and the AI model itself. However, there remains the risk of running into limitations when trying to explain a black-box model without having access to the training dataset. The typical `TalkToModel` execution flow involves the user entering textual inputs that are processed one at a time to execute the corresponding commands. With our formalization, we add the concept of *routine*, which in the context of `TalkToModel` results in the invocation of functionality from different domains within the same user request.

XAgent `XAgent` [27] explicitly maps user intent to the authors' systematically chosen methods of XAI. In this configuration, from the point of view of user experience, the set of routines becomes particularly advantageous: in order to launch the execution of multiple methods, the user must be able to formalize his request so that all the desired intents are well exposed. This request may be arbitrarily ambiguous due to the amount of content and may result in repeating the query multiple times. A direct approach, on the other hand, such as defining a routine, avoids the composition of long sentences. In Figure 4 we show how the `XAgent` [27] architecture can be extended by introducing the component that can execute user-defined routines. We have also added a step prior to the generation of the response identified by the calculation of metrics on the explanation.

dr_ant The original architecture of `dr_ant` [13] is based on a multi-shift chatbot whose task is to answer users' questions about the underlying black-box model trained on the Titanic dataset. We show how our proposed formalization can be applied to this type of system in Figure 5. We emphasize the fundamental aspects necessary to achieve a conversational experience that can be evaluated

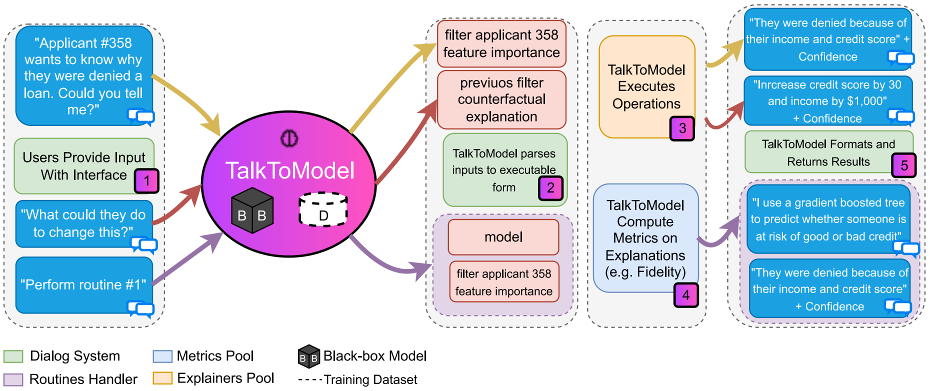
"Applicant #358 wants to know why they were denied a loan. Could you tell me?"
Users Provide Input With Interface
1
"What could they do to change this?"
"Perform routine #1"
TalkToModel
B B
D
filter applicant 358 feature importance
previuos filter counterfactual explanation
TalkToModel parses inputs to executable form
2
model
filter applicant 358 feature importance
TalkToModel Executes Operations
3
TalkToModel Compute Metrics on Explanations (e.g. Fidelity)
4
"They were denied because of their income and credit score" + Confidence
"Inrcrease credit score by 30 and income by $1,000" + Confidence
TalkToModel Formats and Returns Results
5
"I use a gradient boosted tree to predict whether someone is at risk of good or bad credit"
"They were denied because of their income and credit score" + Confidence
Dialog System
Routines Handler
Metrics Pool
Explainers Pool
Black-box Model
Training Dataset

Fig. 3. Extended overview of `TalkToModel`. The original architecture proposed in [36] takes place in four stages. We introduce a new type of input, such as a user-defined routine, to automate recurring tasks. We also add an additional step, such as calculating metrics that must be formatted and displayed in the form of a degree of confidence in the final result.

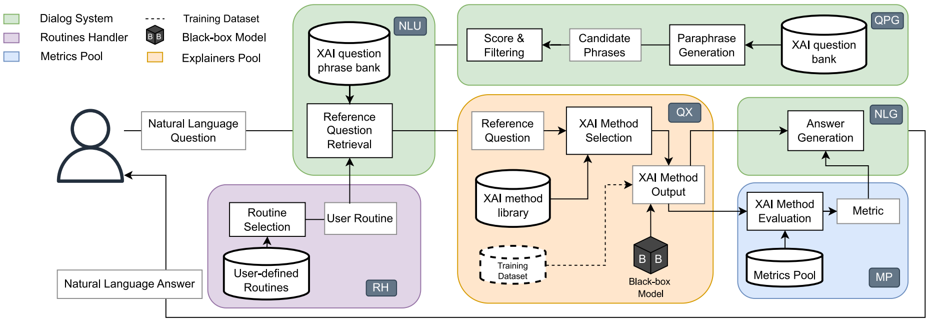

Fig. 4. Extended architecture of `XAgent` We added the visualization of the training dataset, represented by a dotted line as optional, and the black-box model to be explained. In addition, we added the user-defined routines component and the pool of metrics used to evaluate the explanation.

by the end user himself. Furthermore, since users' backgrounds may vary, the ability to automate a set of questions in the form of routines is a functional aspect that can enhance both the experience of the experienced user, such as a data scientist, and the user with no prior knowledge, who can find his own aggregation of common questions that meets his needs.

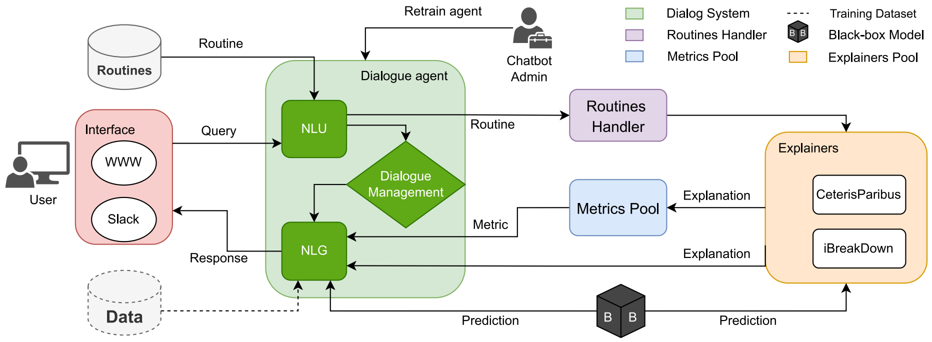

Fig. 5. Extended architecture of `dr_ant`. The user can request one or more explanation tasks from the agent in the form of a routine. The explanation produced is evaluated with appropriate metrics contributing to the final answer. Requests concerning the initial dataset, such as feature distribution information, can be fulfilled if the dataset is available.

5 Conclusion and Future Work

In this study, we reviewed the current literature on conversational XAI systems and provided a possible general formalization. Our proposal outlines what are the key components to build a Conversational XAI system necessary for a satisfactory user experience. To verify that our formalization is general, we applied it to Conversational XAI models available in the literature, showing how different implementations can conform to our formalization. Our goal is to provide other researchers with a starting point for implementing conversational explanation systems by defining the components needed to relate transparency and trust requirements to real-world scenarios. We examined the trend toward which current conversational systems are moving, noting that complex representations of explanations are becoming increasingly common in order to provide end users with all the elements to evaluate the decisions made by the AI model.

Future work involves the implementation of the formalized components into conversational systems in order to estimate their impact on current systems in terms of performance and user experience. Just as the advancement of technology has generated new jobs needed for current automated systems, we believe that the deployment and rapid growth of XAI will lead to the need for a specialized black-box systems explainer figure, such as an AI explainer engineer or an

AI explainer architect, for entities adopting AI within their business in the coming years. In addition, collaboration among professionals with multidisciplinary backgrounds during the design phase can enable the creation of increasingly human-friendly XAI systems.

ACKNOWLEDGMENTS This work has been partially supported by the Italian Ministry of Health, Piano Operativo Salute (POS) trajectory 2 “eHealth, diagnostica avanzata, medical device e mini invasività” through the project “Rete eHealth: AI e strumenti ICT Innovativi orientati alla Diagnostica Digitale (RAIDD)”(CUP J43C22000380001) and the Italian Ministry of Health, Piano Operativo Salute (POS) trajectory 4 “Biotechnology, bioinformatics and pharmaceutical development” through the Pharma-HUB Project “Hub for the repositioning of drugs in rare diseases of the nervous system in children” (CUP J43C22000500006).

This work was also partially supported by the European Union under ERC-2018-ADG G.A. 834756 (XAI), G.A. 952026

References

1. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020). https://doi.org/10.48550/ARXIV.2005.14165
2. Browne, J.T.: Wizard of oz prototyping for machine learning experiences. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. p. 1–6. CHI EA '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3290607.3312877
3. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (mar 2018). https://doi.org/10.1109/wacv.2018.00097
4. Chen, Z., Wang, X., Xie, X., Parsana, M., Soni, A., Ao, X., Chen, E.: Towards explainable conversational recommendation. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. pp. 2994–3000. International Joint Conferences on Artificial Intelligence Organization (7 2020). https://doi.org/10.24963/ijcai.2020/414, main track
5. Dosilovic, F.K., Brcic, M., Hlupic, N.: Explainable artificial intelligence: A survey. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) pp. 0210–0215 (5 2018). https://doi.org/10.23919/MIPRO.2018.8400040
6. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2017). https://doi.org/10.1109/iccv.2017.371
7. Gosiewska, A., Biecek, P.: Do not trust additive explanations (2020)
8. Greenwell, B.M., Boehmke, B.C., McCarthy, A.J.: A simple and effective model-based variable importance measure (2018)
9. Guidotti, R.: Evaluating local explanation methods on ground truth. Artificial Intelligence **291**, 103428 (2021). https://doi.org/10.1016/j.artint.2020.103428
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5) (aug 2018). https://doi.org/10.1145/3236009
11. Jentzsch, S.F., Höhn, S., Hochgeschwender, N.: Conversational interfaces for explainable ai: A human-centred approach. In: Calvaresi, D., Najjar, A., Schumacher, M., Främling, K. (eds.) Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 77–92. Springer International Publishing, Cham (2019)
12. Kulkarni, P., Mahabaleshwarkar, A., Kulkarni, M., Sirsikar, N., Gadgil, K.: Conversational ai: An overview of methodologies, applications & future scope. In: 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA). pp. 1–7 (2019). https://doi.org/10.1109/ICCUBEA47591.2019.9129347
13. Kuź ba, M., Biecek, P.: What would you ask the machine learning model? identification of user needs for model explanations based on human-model conversations. In: ECML PKDD 2020 Workshops, pp. 447–459. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-65965-3_30

14. Kuźba, M., Baranowska, E., Biecek, P.: pyceterisparibus: explaining machine learning models with ceteris paribus profiles in python. Journal of Open Source Software **4**(37), 1389 (2019). https://doi.org/10.21105/joss.01389
15. Lakkaraju, H., Slack, D., Chen, Y., Tan, C., Singh, S.: Rethinking explainability as a dialogue: A practitioner's perspective (2022). https://doi.org/10.48550/ARXIV.2202.01875
16. Liao, Q.V., Gruen, D., Miller, S.: Questioning the AI: Informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM (apr 2020). https://doi.org/10.1145/3313831.3376590
17. Liu, P., Li, K., Meng, H.: Out-of-scope domain and intent classification through hierarchical joint modeling. In: Conversational AI for Natural Human-Centric Interaction: 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore. pp. 3–16. Springer (2022)
18. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
19. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions (2017). https://doi.org/10.48550/ARXIV.1705.07874
20. Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: A grounded interaction protocol for explainable artificial intelligence (2019). https://doi.org/10.48550/ARXIV.1903.02409
21. Madumal, P., Miller, T., Vetere, F., Sonenberg, L.: Towards a grounded dialog model for explainable artificial intelligence (6 2018), http://arxiv.org/abs/1806.08055
22. Malandri, L., Mercorio, F., Mezzanzanica, M., Nobani, N.: Convxai: a system for multimodal interaction with any black-box explainer. Cognitive Computation (2022). https://doi.org/10.1007/s12559-022-10067-7
23. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. CoRR **abs/1706.07269** (2017)
24. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
25. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM (jan 2020). https://doi.org/10.1145/3351095.3372850
26. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Computing Surveys (feb 2023). https://doi.org/10.1145/3583558
27. Nguyen, V.B., Schlötterer, J., Seifert, C.: Explaining machine learning models in natural conversations: Towards a conversational xai agent (2022). https://doi.org/10.48550/ARXIV.2209.02552
28. Panigutti, C., Beretta, A., Fadda, D., Giannotti, F., Pedreschi, D., Perotti, A., Rinzivillo, S.: Co-design of human-centered, explainable ai for clinical decision support. ACM Trans. Interact. Intell. Syst. (mar 2023). https://doi.org/10.1145/3587271, just Accepted
29. Panigutti, C., Perotti, A., Pedreschi, D.: Doctor xai: An ontology-based approach to black-box sequential data classification explanations. In: Proceedings

of the 2020 Conference on Fairness, Accountability, and Transparency. p. 629–639. FAT* '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3372855

30. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models (2018). https://doi.org/10.48550/ARXIV.1806.07421
31. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html
32. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. CoRR **abs/1602.04938** (2016), http://arxiv.org/abs/1602.04938
33. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018). https://doi.org/10.1609/aaai.v32i1.11491
34. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (oct 2019). https://doi.org/10.1007/s11263-019-01228-7
35. Singh, S., Beniwal, H.: A survey on near-human conversational agents. Journal of King Saud University - Computer and Information Sciences **34**(10, Part A), 8852–8866 (2022). https://doi.org/10.1016/j.jksuci.2021.10.013
36. Slack, D., Krishna, S., Lakkaraju, H., Singh, S.: Talktomodel: Explaining machine learning models with interactive natural language conversations (2022). https://doi.org/10.48550/ARXIV.2207.04154
37. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise (2017). https://doi.org/10.48550/ARXIV.1706.03825
38. Steinfeld, A., Jenkins, O.C., Scassellati, B.: The oz of wizard: Simulating the human for interaction research. In: 2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 101–107 (2009). https://doi.org/10.1145/1514095.1514115
39. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks (2017). https://doi.org/10.48550/ARXIV.1703.01365
40. Van Looveren, A., Klaise, J.: Interpretable counterfactual explanations guided by prototypes. In: Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J., Lozano, J.A. (eds.) Machine Learning and Knowledge Discovery in Databases. Research Track. pp. 650–665. Springer International Publishing, Cham (2021)
41. van der Waa, J., Robeer, M., van Diggelen, J., Brinkhuis, M., Neerincx, M.: Contrastive explanations with local foil trees (2018)
42. Wijekoon, A., Corsar, D., Wiratunga, N.: Behaviour trees for creating conversational explanation experiences (2022). https://doi.org/10.48550/ARXIV.2211.06402