A Multi-layered Approach for Tailored Black-box Explanations

Clément Henin, Daniel Le Métayer

A Multi-layered Approach for Tailored Black-box Explanations

Clément Henin[1,2] *[0000−0002−7615−9988] and Daniel Le Métayer[1]

[1] Univ Lyon, Inria, INSA Lyon, CITI, Villeurbanne, France {clement.henin, daniel.le-metayer}@inria.fr
[2] École des Ponts ParisTech, Champs-sur-Marne, France

Abstract. Explanations for algorithmic decision systems can take different forms, they can target different types of users with different goals. One of the main challenges in this area is therefore to devise explanation methods that can accommodate this variety of situations. A first step to address this challenge is to allow explainees to express their needs in the most convenient way, depending on their level of expertise and motivation. In this paper, we present a solution to this problem based on a multi-layered approach allowing users to express their requests for explanations at different levels of abstraction. We illustrate the approach with the application of a proof-of-concept system called IBEX to two case studies.

Keywords: Algorithmic decision system · explainability · transparency · black-box model · machine-learning · artificial intelligence · interactive

1 Introduction

Explainability has generated increased interest during the last decade because accurate ML techniques often lead to opaque Algorithmic Decision Systems (hereafter "ADS") and opacity is a major source of mistrust. Indeed, even if they should not be seen as a silver bullet, well designed explanations can play a key role, not only to enhance trust in a system, but also to allow its users to better understand its outputs and therefore to make a better use of them. In addition, they are necessary to make it possible to challenge decisions based on the results of an ADS. On the legal side, Recital 71 of the European General Data Protection Regulation, which concerns decisions "based solely on automated processing", states that a data subject has the right "to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision."

Explainability methods produce different types of explanations in different ways, based on different assumptions on the system [1]. In this paper, we focus on a category of methods, called "black-box explanation methods", which do not assume the availability of the code of the ADS or its underlying model. The

* Corresponding author (clement.henin@inria.fr)

only assumption is that input data can be provided to the ADS and its outputs can be observed.

In practice, explanations can take different forms, they can target different types of users (hereafter "explainees") with different interests. One of the main challenges in this area is therefore to devise explanation methods that can accommodate this variety of situations. This is especially crucial to avoid the "inmates running the asylum" phenomenon [2] and be able to design a system that can be used by lay persons. A first step to address this challenge is to allow explainees to express their needs in the most convenient way, which is not an easy task especially for users lacking technical expertise. In this paper, we present a solution to this problem based on a multi-layered approach allowing users to formulate their requests for explanations at different levels of abstraction. The three levels of abstraction considered here are called respectively the *context*, the *requirements* and the *technical options*:

1. The *context* provides high-level information about the profile of the explainee and his/her objectives.
2. The *requirements* characterize the desired explanations, including, for example, their format, degree of simplicity and generality.
3. The *technical options* are lower-level choices related to the available explanation techniques.

We provide a mapping between the different levels of abstraction to generate explanations tailored to the needs of each explainee. In addition, we make it possible for explainees to react to an explanation. They can, for example, request more detailed, or simpler explanations, or explanations in a different form.

The idea is that lay users should be able to express their needs at the highest level of abstraction, without any knowledge of the requirements and technical options. On the other hand, expert users, for example the designers of the ADS, may prefer to express their requests directly as requirements or technical options. Regardless of the level of abstraction adopted by the user, ultimately his/her needs have to be translated into technical options. In this paper, we describe a heuristic method to derive requirements from contexts and suggest the derivation of technical options from requirements for different explanation methods.

We first present the two higher levels of abstraction (context and requirements) in Section 2. In Section 3, we show the derivation of requirements from contexts and suggest how technical options can be derived from requirements. In Section 4 we illustrate the approach with the application of our proof-of-concept system IBEX (for "Interactive Black-box Explanations") to two case studies. Section 5 discusses related work and Section 6 concludes with prospects for future work.

2 Context and requirements

In this section, we present successively the higher levels of abstraction of our framework: the context (Section 2.1) and the requirements (Section 2.2). The

mapping between these levels is described in Section 3. The methodology followed to devise the framework relies on a detailed analysis of existing explanation methods [3] as well as existing literature to identify the needs and the expectations of the users. The most relevant references are included in the text and further discussed in Section 5.

2.1 Context

The context is the highest level of abstraction, which should be accessible to any explainee, including lay users, to express their needs in a simple, non technical, way. Contexts are made of the *ADS* to be explained[3] and four elements related to the explainee's query: *Profile*, *Objective*, *Focus* and *Point of interest*.

- *Profile* takes a value in the set $\{TE, AU, DE, LU\}$. TE represents technical experts, AU auditors, DE domain experts and LU lay users. Technical experts include designers, developers, testers, i.e. people having some knowledge about the design or the techniques used to implement the ADS. Auditors are also assumed to have a high level of expertise but they are involved in a specific task of auditing or evaluating the ADS. Domain experts are not assumed to have any expertise about the ADS itself or the technology used but they are knowledgeable about the application domain. Examples of domain experts include medical doctors, judges or police officers. The last category, lay users, includes users who are not assumed to possess any specific knowledge. They may be persons affected by decisions relying on the ADS or simple citizens.[4]
- *Objective* takes a value in the set $\{I, T, C, A\}$. I represents the improvement of the ADS, T trust enhancement, C challenging a decision and A taking actions based on a decision. The improvement of the ADS includes its testing, assessment of its accuracy and any action to detect potential weaknesses. Trust enhancement includes a variety of objectives related to the use of the ADS (avoiding wrong decisions [1], enhancing the acceptance of the results [1], increasing the predictability of the output [6] and being comfortable with the strengths and limitations of the ADS [7]) or its purpose (causality, transferability [8, 5]). Challenging a decision and taking an action based on a decision are two alternative reactions for the person affected by a decision [9]. Actions that can be taken based on a decision include actions that can have an impact on the person's record and therefore on future decisions. An example of action for the customer of a bank could be to reduce his/her outstanding loan balance to increase his/her chances to have his/her new credit application accepted.
- *Focus* characterizes the scope of the explanation. It takes a value in the set $\{G, L\}$. G stands for global explanation and L for local explanation. An

[3] With the associated learning data set, if available.

[4] Other taxonomies of explainees' profiles have already been proposed, in particular in [4] and [5]. Our contribution is consistent with them, but involves some simplifications, justified by pragmatic needs.

explanation is global if the explainee is interested in the behaviour of the ADS for the whole input dataset. Otherwise, it is local, which means that the explainee is interested in the behaviour of the ADS for (or around) a specific input value.
– *Point of interest* defines the input value x which is the point of interest of the explainee when the focus of the explanation is local (otherwise, the context does not involve any point of interest).

We should emphasize that some of these elements can be omitted by explainees if they are not sure about them. The only mandatory element is the *ADS*. Explanations can be generated from partially defined contexts. The drawback is that such explanations may not correspond to the expectations of the explainee who may then have to refine his/her needs through further interaction steps.

2.2 Requirements

Requirements provide an intermediate level of abstraction. They characterize the desired explanations more precisely than the context but still in an abstract way. They can be useful to certain lay users, depending on their level of proficiency, and to expert users. The requirements are made of seven elements[5]: *Format*, *Simplicity*, *Generality*, *Point of interest*, *Realism*, *Actionability* and *Nature*. Apart from *Realism*, which is, to the best of our knowledge, an original contribution, these elements are motivated by previous work and experimental studies, as mentioned below.

– *Format* includes the different forms of explanations that can be generated [1, 8]. The impact of the format on the acceptance of explanations is analyzed in [10, 11]. Examples of formats include rule based explanations (RB), feature importance (FI), counterfactual explanations (CF), decision trees (DT) and partial dependence plots (PD).
– *Simplicity* is a key requirement as it generally relates to understandability [1, 6, 12]. It is usually expressed through a fixed scale of values. The current version of IBEX considers three increasing levels of simplicity: Simplicity = $\{1, 2, 3\}$.
– *Generality* characterizes the size of the class of input values that should be covered by the explanation ([6] p.44). Some authors use the word "cover" to denote the same concept [12, 13]. It is also expressed through a fixed scale of values. The current version of IBEX considers three increasing levels of generality: Generality = $\{1, 2, 3\}$. Level 1 covers a single input (the point of interest), level 3 a wide class of inputs and level 2 is intermediate. Note that generality is defined only for local explanations since global explanations cover, by definition, the whole input dataset.
– *Point of interest* has the same definition as above (for contexts). It also belongs to the requirements for the sake of comprehensiveness (each level is

[5] In addition to the ADS, as defined in the context.

assumed to be self-contained). Like generality, the point of interest is defined only for local explanations.

- *Realism* characterizes the level of realism required for an explanation. By "realism", we mean the fact that the explanation process takes into account the actual distribution of the input data. Realistic explanations are preferable for explainees interested in the actual usage of the ADS. On the other hand, explainees interested in the internal logic of the ADS, independently of its actual usage, may proceed without the constraint of realism. Let us consider this notion with the example of a credit scoring system. The ADS systematically outputs the maximum risk when the application file mentions a previous credit fraud. Although this feature has a tremendous impact on the score, it is rarely used in practice, as few credit applicants are in this situation. The realistic approach takes into account the low probability of this feature while the non-realistic approach only considers the model itself, and thus assigns great importance to this feature. The current version of IBEX considers three increasing levels of realism: Realism $= \{1, 2, 3\}$.
- *Actionability* expresses the fact that actionable explanations should be preferred. An actionable explanation is an explanation involving only actionable features of the input dataset ([9] p.42). For example, in the input file of a loan applicant, the age variable is not actionable whereas the number of outstanding loans is actionable. The current version of IBEX considers two options: Actionability $= \{T, F\}$. Value T means that actionability is a requirement. In this case, the explainee has to provide the list of actionable features.
- *Nature* corresponds to the presence or absence of probability in the explanations ([6] p.44). The current version of IBEX considers two options: Nature $= \{T, F\}$. Value F means that probabilistic explanations are not desired and value T that they are acceptable.

Like contexts, requirements can be partially defined. In addition, they may be expressed in terms of preferences rather than fixed choices. For example, a technical expert may characterize simplicity by $3 > 2 > 1$ to express a preference for simple explanations but can also cope with intermediate or complex explanations. On the other hand, lay users may prefer to characterize simplicity by selecting only value 3. In the following, the former are called soft requirements and the latter hard requirements. In addition, soft requirements may also be prioritized (ranked by order of importance). For example, a technical expert who wants to debug or improve the ADS may consider that generality is more important than simplicity ($general > simple$).

3 From contexts to explanations

In order to produce explanations, the needs described in the previous section have to be translated into technical options of the generic explainer. In this section, we present the two phases of this process, the translation of the context into

requirements in Section 3.1 and the translation of requirements into technical options in Section 3.2. The whole process is sketched in Figure 1.

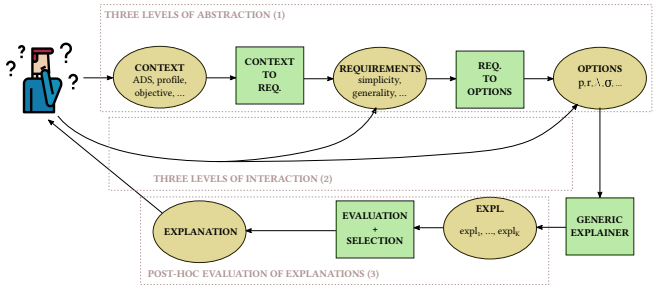

Fig. 1: Overview of the approach.

3.1 From context to requirements

The first step of the translation procedure consists in using the *Focus* element of the context to select the subset of formats that can be used. For example, if $Focus = G$ (meaning that the explainee is interested in a global explanation), then counterfactual explanations (CF) are not appropriate. If $Focus = L$ (local explanation), then the *Point of interest* element of the requirements is obtained directly from the same element in the context. The other elements of the requirements are derived from the *Profile* and *Objective* elements of the context as presented in Table 1.

In the following, we provide some intuition about the choices made in Table 1. Usually, simple explanations are preferred over complex explanations ([6] p.44). Simplicity is expressed as a soft requirement with a low priority unless the profile is *Lay User*. Lay users generally expect explanations that are as simple as possible, thus a hard requirement is used ($simple = 3$).

The generality of an explanation (which is relevant only for local explanations) enhances the explainee's capabilities to understand the outcomes of the ADS for input values that have similarities with the point of interest. Therefore the values of the generality element should be maximum ($general = 3$) when the objective is to increase the trust in the model ([6] p.44). On the other hand, when the objective for a lay user is to challenge a specific decision or to take actions to obtain better decisions from the ADS, a lower level of generality is more appropriate.

High levels of realism favour the generation of explanations that are supported by training data [14]. Depending on the context, this choice can be an advantage or a drawback. Explanations that are not supported by training data make it possible to analyze decision boundaries that are part of the model, but are not necessarily reflected in actual field data, as mentioned in the credit

Technical Expert		**Domain Expert**		
*Improve**	*Trust*	*Trust**	*Challenge*	*Action*
format: RB >DT >FI >PD >CF	format: RB >DT >FI >PD >CF	format: RB >DT >FI >PD >CF	format: RB >DT >FI >PD >CF	format = CF
simplicity: 3 >2 >1	simplicity: 3 >2 >1	simplicity: 3 >2 >1	simplicity: 3 >2 >1	simplicity: 3 >2 >1
generality: 3 >2 >1	generality = 3	generality = 3	generality: 1 >2 >3	generality: 1 >2 >3
realism = 1	realism: 3 >2 >1	realism = 3	realism = 1	realism = 2
actionability = F	actionabililty = F	actionability = F	actionability = F	actionability = T
nature = T	nature = T	nature: T >F	nature : F >T	nature = F
general >form >simple	simple >real >form	simple >nat >form	form >nat >simple	simple >gen
Auditor		**Lay User**		
Trust	*Challenge**	*Trust**	*Challenge*	*Action*
format: RB >DT >FI >PD >CF	format: RB >DT >FI >PD >CF	format: RB >DT >FI >PD >CF	format: RB >DT >FI >PD >CF	format = CF
simplicity: 3 >2 >1	simplicity: 3 >2 >1	simplicity = 3	simplicity = 3	simplicity = 3
generality = 3	generality: 1 >2 >3	generality = 3	generality: 1 >2 >3	generality: 1 >2 >3
realism = 3	realism = 1	realism = 3	realism = 1	realism = 2
actionability = F	actionability = F	actionability = F	actionability = F	actionability = T
nature: T >F	nature: T >F	nature : F >T	nature : F >T	nature = F
simple >nat. >form	form >nat. >simple	nat. >form	form >nat.	

Table 1: Translation of the context into requirements. Hard requirements appear in black type and soft requirements in green type. Objectives marked with a star are used as default settings.

scoring example of Section 2.2. When the objective of the explainee is trust enhancement, decision boundaries that are actually used must be the primary concern, which justifies the choice of realistic sampling. On the other hand, technical experts may want to investigate these "theoretical" decision boundaries in order to assess the robustness of the model in all conditions.

As shown by previous studies ([6] p.44), the use of probabilities in explanations is usually not illuminating for explainees ($nature = F$), especially when they are interested in a single point of interest. However some profiles, such as auditors and technical experts, may be interested in an overall view of the situation, which is provided by the use of probabilities ($nature: F > T$).

To conclude this section, we would like to emphasize that Table 1 corresponds to the choices made in IBEX but they are not hard-wired in the implementation. The architecture of the system can accommodate different choices of translation and this flexibility will be used to improve it based on the feedback of the users and field experience.

3.2 From requirements to technical options

The translation of the requirements into technical options depend on the available explanation methods. For instance, simplicity can be translated into an acceptable number of non-zero coefficients for explanations expressed as feature importance or number of nodes for decision trees. Generality has an impact on the range of the sampling of the explanation method, *i.e.* the average distance between the point of interest and the samples. Interested readers can find in [15] the details of the translation for IBEX, which makes it possible to generate different forms of explanations based on a variety of parameters.

In general, the translation procedure may yield several possible solutions (sets of technical options), in particular when soft requirements are involved. In such cases, it is necessary to choose among them the set of technical options that is the most likely to address the needs of the explainee. To address this issue and to ensure that the explanation generated by the explainer will meet the requirements, the translation process of IBEX includes a last *post-hoc evaluation* step: the generation of the explanations corresponding to the different technical options derived in the previous step, followed by an evaluation of their properties.

Generally speaking, the assessment of the qualities of explanations is still an open research question. We consider here their compliance with respect to requirements as defined in Section 2.2. More precisely, we focus on the *Simplicity* and *Generality* elements, which are often expressed as soft requirements. The assessment of simplicity is based on the number of items involved in the explanation (*e.g* the number of rules in a rule-based model, the number of modifications in a counterfactual example, etc.). This use of the size of an explanation as a proxy for simplicity is common [16]. It has some limitations (size does not always reflect simplicity) but it is operational and it can be instantiated to any explanation format. In IBEX, the assessment of generality relies on a test of the explanation on inputs from the population that are close to the point of interest. If the explanation is not valid for a minimum number of inputs (threshold T_1) then the generality is 1; if it is valid for the T_1 closest inputs but not for T_2 inputs ($T_2 > T_1$), then the generality is 2; if it is valid for the T_2 closest inputs then the generality is 3[6].

To conclude this section, it is important to stress that the definition of the needs of the explainee (at one of the three levels of abstractions) is only the first interaction step of the explainee with IBEX. When an explanation has been generated by IBEX based on the set of technical options resulting from the initial step, the explainee can reply to IBEX with a new request. This request can refer to the initial explanation (e.g. asking for a "richer", or "less simple", explanation, or an explanation in a different format) or can be entirely new and expressed again at any level of abstraction. By allowing explainees to interact at a different abstraction levels, IBEX gives them the opportunity to express their needs in a very precise and interactive way.

[6] In the current version of IBEX , threshold T_1 is set to 10 and T_2 is set to 50.

4 IBEX at work: application to case studies

In this section, we illustrate our approach with the application of our proof-of-concept system IBEX to two case studies. The implementation of the interaction protocol of IBEX follows directly the approach presented in the previous sections and interested readers can find in [15, 17] complementary information about the explanation techniques available in IBEX. The code of IBEX is publicly available[7].

Interactions at any level of abstraction are feasible with IBEX. By default, the interaction is done at the context level which is the most appropriate for lay users. These interactions take place as follows (questions asked by IBEX):

1. Choose a data set.
2. Are you interested in global (G) or local (L) explanations?
3. What is your point of interest? (optional question: for local explanations only)
4. How do you want to be considered by IBEX: as a technical expert (TE), a lay user (LU), a domain expert (DE) or an auditor (AU) ?
5. What is the objective of the explanation: is it to improve the ADS (I), to enhance your trust in the ADS (T), to challenge the ADS (C), or to take future actions based on results of the ADS (A)?
6. What are your actionable features? (optional question: for objective A only)

The user may skip any of these questions (except the first one) if he/she is not sure about the answer. In any case, IBEX then generates a first explanation based on this (potentially partial) context and asks whether the user has further questions. If so, the user has two options: either ask an entirely new question (simple iteration of the protocol) or ask a question based on the previous explanation. In this case, he/she can express his/her wishes as a tuning of the requirements, for example "simpler explanation", "more general explanation", or "actionable explanation". Alternatively, the user can ask to see the requirements derived from the previous question and modify any of its components by himself/herself. In both cases, IBEX will then generate new technical options and a new explanation based on the new requirements. The user will then have again the options to stop, ask an entirely new question or a question involving the previous answer.

In order to illustrate the benefits of the approach in terms of versatility and interactivity, we consider two possible ways of using IBEX to get explanations about hypothetical ADS based on publicly available datasets.

1. The first situation corresponds to a lay user requesting explanations about the ADS with the objective of enhancing trust (Section 4.1).
2. The second situation is a lay user requesting explanations with the objective of taking future actions to improve his/her record. (Section 4.2).

[7] https://gitlab.inria.fr/chenin/ibex

4.1 Explanations to enhance trust

The first case study involves the adult census data set[8]. This data set, which has been extracted from the 1994 US census, contains personal information about American citizens such as their age, education level or marital status. The goal of the ADS is to predict, from these features, if the individual earns more or less than $50,000$ per year. A lay user who wants to enhance his/her trust in the ADS would choose the following answers: *data set=adult census, focus=G, profile=LU and objective=T*. From this context, IBEX has generated the explanation presented in Figure 2. We can see that the explanation is simple, it is composed of a decision tree with only two nodes and three leaves, which is consistent with the choices presented in Table 1.

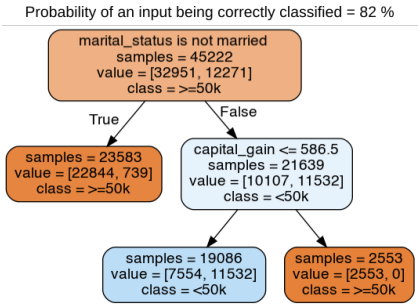

Fig. 2: Explanation generated by IBEX for the adult census data set from the initial context. The explanation is a decision tree applied to the training data set with labels replaced by model's outputs. The following information is associated with each node: number of samples meeting the conditions leading to the subtree (*samples*), numbers of samples belonging respectively to the class $>= 50k$ and to the class $< 50k$ (*value*), and the majority class for this subset of samples (either $< 50k$ or $>= 50k$). To meet the simplicity constraint (only 3 leaves), the tree must be approximate and is therefore only valid for a part of the inputs (*82 %*). IBEX has used the following requirements to generate this decision tree: *format=DT, simplicity=3, actionability=F, nature=T, realism=3*.

The requirements generated by IBEX for this context are presented in the left part of Figure 3. We can see that $nature = F > T$, meaning that an explanation that does not involve any probability would have been preferred by the user. Nevertheless, the explanation generated by IBEX involves a probability. The reason is that the first explanation formats that were considered by IBEX (*FI* and *PD*, which are non probabilistic, in accordance with the soft requirement $nature : F > T$) led to explanations that were considered too complex to satisfy the hard requirement $simplicity = 3$. For this reason, the post-hoc evaluation step of IBEX made the choice of a decision tree format.

[8] https://archive.ics.uci.edu/ml/datasets/Adult

```
HARD REQUIREMENTS:
        actionability : F
        simplicity : 3
        generality :
        realism : 3
SOFT REQUIREMENTS:
        nature : F > T
        format : DT > PD > PC
```

(a)

```
HARD REQUIREMENTS:
        actionability : F
        simplicity : 2
        generality :
        realism : 3
        format : DT
SOFT REQUIREMENTS:
        nature : F > T
```

(b)

Fig. 3: (a) Requirements derived by IBEX from the initial context (G, LU, T) corresponding to the columns Lay User / Trust of Table 1. Because the focus is "Global", formats corresponding to local explanations (RB, FI, CF) are not considered and the generality requirement is empty as it is only applicable to local explanations; (b) Revised requirements based on the user's request "less simple".

Let us assume now that the user is almost satisfied with this first explanation but he/she suspects that the logic of the ADS is much more complex and this explanation is a bit simplistic. Through the IBEX interface, he/she can either request a "less simple" explanation or ask IBEX to show the requirements derived from the previous question and modify by himself/herself the simplicity element. In the first case, IBEX would generate the requirements shown in the right part of Figure 3 leading to a richer decision tree, as shown in [15].

4.2 Explanations to take actions

The second case study concerns the German credit data set[9] which contains information about the credits (amount, duration, purpose, etc.) and the applicants (type of job, number of ongoing credits, etc.). The ADS classifies applications as risky ("bad") or safe ("good"). Let us consider an individual whose credit application has been rejected and who would like to know how to improve it to have it accepted in the future. The profile for this query is lay user (LU) and the objective is to prepare future actions (A) for a specific input (L). From Table 1, we can see that IBEX associates this context with the *CF* format and the average level *2* of realism. Indeed, level *1* would lead to unlikely modifications of the application that might not be of any practical use. At the other extreme, level a generation method based on a high realism (*3*) would involve only real examples. Restricting the search for counterfactuals to real examples is not necessary in our case and would probably yield counterfactuals that are too far away from the optimal value. Also, possible modifications need to be limited to actionable features (e.g. duration of the credit or number of ongoing credits), which are provided by the explainee. The counterexample generated from this context by IBEX, shown in Table 2, suggests two modifications of the current application: the duration of the credit and telephone ownership.

[9] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Actionable features	Credit amount	Duration	Ongoing credits	Job	Telephone ownership	Output
Current	10722	47	1	unskilled resident	yes	**Bad**
CF	10722	36	1	unskilled resident	none	**Good**

Table 2: Realistic counterfactual explanations based on actionable features. The first line shows the current attribute values of the Point of Interest while the second line shows the attribute values of a counterfactual (CF) input being classified as "Good" (low credit risk) by the system.

Comparing these two case studies gives insight of the diversity of explanations that can be generated with IBEX. We see that the framework proposed to define the context and the associated requirements offer an understandable way to interact with explanations, event for a lay user. Further examples of the use of IBEX are presented in [17] .

5 Related works

To the best of our knowledge, no existing explanation system provides the diversity of explanations and the interaction capabilities offered by IBEX. Some authors have already proposed taxonomies of explainee's profiles [4, 18], explanations' objectives [6, 8, 19] or combinations of profiles and objectives [7, 20, 5]. The impact of the type of question on the explanation has been analyzed through a user study in [11]. In the same vein, different forms of explanations are studied in [21]. Some works also aim at identifying appropriate sets of features of explanations [22, 1, 6]. These contributions are related to this paper in the sense that the categorization of explainees' needs is a key element of our interactive approach. However, the goal of these contributions is to identify and categorize these needs, rather than to design a generic interactive explainer. To the best of our understanding, none of them suggests an operational mapping to actual explanations as presented here.

Some contributions involve a form of interaction with explainees. AIX360 [23] contains eight explainability algorithms and allows users to choose among them based on a taxonomy including criteria such as "understand the data or the model" or "self-explaining model or post-hoc explanations". As it takes into consideration the user's needs, AIX360 provides a first level of interaction with explainees. However, the three levels of abstraction available in IBEX allow for richer interactivity, for instance, by allowing to choose the levels of simplicity, generality and realism of the explanation. Moreover, the generic explainer can be customized to fulfill the requirements of the explainee, which is not possible with the portfolio approach of AIX360. Finally, IBEX offers the possibility to react to an explanation, which is also a distinguishing feature.

Glass-Box [24] allows explainees to interact with an adaptive explainer through a voice-based (or chat-based) interface. The system provides local explanations, under the form of counterfactuals, and allows explainees to react in order to obtain a new explanation. Although Glass-Box has similarities with IBEX, its

interactive capabilities are limited to the choice of actionable features for counterfactuals (which is also included in IBEX).

The bLIMEy system (for "build LIME yourself"), is a generic explainer relying on the framework proposed in [3]. However, bLIMEy does not include an analysis of the context of the explainee's query, neither does it include a mapping from this context to technical options, as done in IBEX.

Some authors consider interactive explanation frameworks from a more theoretical point of view. For instance, [25] defines the specifications of a dialogue system for explanations and [26] proposes an interaction protocol for XAI. These works are related to IBEX, and could be useful sources of inspiration to enhance its interaction facilities. However, their goal is not to propose an operational explanation system.

Finally, on the implementation side, many projects have recently emerged to provide implementations of existing methods [27–29]. The goal of these projects is to integrate a variety of existing methods, but they do not include a comprehensive interaction module and a fine-grain decomposition of components as done in IBEX.

6 Conclusion

The main goal of the work described in this paper is to address the variety of needs in terms of explanations of ADS and to design an explanation system that can be used by a wide range of explainees, including lay users. We have shown, through the IBEX prototype, the feasibility of an interactive explanation system based on our multi-layered approach. IBEX is a generic explanation generation system based on a variety of parameters and fine-grained components that can be combined in different ways. The architecture of IBEX and its components are described in [17]. As stated above, IBEX is a proof of concept implementation and it can be improved and extended in several directions. A first improvement concerns the user interface, which is very basic in the current version. In particular, it would be interesting to provide a richer and higher-level language to interact with explainees, for instance a restricted version of natural language that could be used by explainees to express questions such as "Why is it the case that my application has been rejected ?" or "Why has this file been accepted and not this one ?" or to express explanations. In some cases, requirements or technical options for the generation of explanations could be derived directly from such questions. In other cases, the explanation system would in turn ask a question to the explainee in order to allow him/her to refine his/her initial request. Dialogue specifications could rely on models such as [26]. Another extension of the tool would be to include an additional component to deal with input data that are not meaningful for humans, as the pixels of an image for example. An initial task is necessary to extract an interpretable representation from such data, as done in LIME [30], for example.

In order to prove its usability as an explanation system in real life, IBEX should be tested through a randomized user study involving different types of

explainees, which we plan to do in the near future with applications in the health care and the judicial sectors. In this perspective, a key aspect of explanations that has not been developed in this paper is their assessment. Different criteria have been proposed to assess the quality of an explanation [31]. Our framework makes it possible to specify quality objectives, either as constraints or as criteria, but it does not provide any help to evaluate the relevance of these objectives (for example through an assessment of the understanding of the explainee). This is a major avenue for further research.

References

1. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys (CSUR) 51 (5) (2018) 93.
2. T. Miller, P. Howe, L. Sonenberg, Explainable AI: Beware of inmates running the asylum, in: IJCAI-17 Workshop on Explainable AI (XAI), Vol. 36, 2017.
3. C. Henin, D. Le Métayer, Towards a generic framework for black-box explanations of algorithmic decision systems (Extended Version), Inria Research Report 9276, https://hal.inria.fr/hal-02131174.
4. R. Tomsett, D. Braines, D. Harborne, A. D. Preece, S. Chakraborty, Interpretable to whom? a role-based model for analyzing interpretable machine learning systems, CoRR abs/1806.07552 (2018).
5. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai, arXiv:1910.10045 [cs]ArXiv: 1910.10045 (Dec 2019).
6. T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2017). doi:10.1016/j.artint.2018.07.007.
7. A. Weller, Challenges for transparency, arXiv:1708.01870 [cs]ArXiv: 1708.01870 (Jul 2017).
8. Z. C. Lipton, The mythos of model interpretability, arXiv:1606.03490 [cs, stat]ArXiv: 1606.03490 (Jun 2016).
9. S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard journal of law & technology 31 (2018) 841–887.
10. S. Stumpf, V. Rajaram, L. Li, M. Burnett, T. Dietterich, E. Sullivan, R. Drummond, J. Herlocker, Toward harnessing user feedback for machine learning 10.
11. B. Y. Lim, A. K. Dey, D. Avrahami, Why and why not explanations improve the intelligibility of context-aware intelligent systems, in: Proceedings of the 27th international conference on Human factors in computing systems - CHI 09, ACM Press, 2009, p. 2119. doi:10.1145/1518701.1519023.
12. H. Lakkaraju, E. Kamar, R. Caruana, J. Leskovec, Interpretable & Explorable Approximations of Black Box Models, arXiv preprint arXiv:1707.01154 (2017).
13. M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: AAAI Conference on Artificial Intelligence, 2018.
14. T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, M. Detyniecki, The dangers of post-hoc interpretability: Unjustified counterfactual explanations, arXiv:1907.09294 [cs, stat]ArXiv: 1907.09294 (Jul 2019).

15. C. Henin, D. Le Métayer, A multi-layered approach for interactive black-box explanations, Inria Research Report 9331, https://hal.inria.fr/hal-02498418.
16. F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, arXiv e-prints (2017) arXiv:1702.08608arXiv:1702.08608.
17. C. Henin, D. Le Métayer, A generic framework for black-box explanations, in: Proceedings of the International Workshop on Fair and Interpretable Learning Algorithms (FILA 2020), IEEE, 2020.
18. G. Ras, M. van Gerven, P. Haselager, Explanation methods in deep learning: Users, values, concerns and challenges, CoRR abs/1803.07517 (2018). arXiv:1803.07517.
19. A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.
20. C. T. Wolf, Explainability scenarios: towards scenario-based XAI design, in: Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19, ACM Press, 2019, p. 252–257. doi:10.1145/3301275.3302317.
21. F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, H. Wallach, Manipulating and measuring model interpretability, arXiv:1802.07810 [cs]ArXiv: 1802.07810 (Feb 2018).
22. M. Hall, D. Harborne, R. Tomsett, V. Galetic, S. Quintana-Amate, A. Nottle, A. Preece, A systematic method to understand requirements for explainable AI (XAI) systems 7.
23. V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al., One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, arXiv:1909.03012 [cs, stat]ArXiv: 1909.03012 (Sep 2019).
24. K. Sokol, P. Flach, One explanation does not fit all: The promise of interactive explanations for machine learning transparency, KI - Künstliche Intelligenz (Feb 2020). doi:10.1007/s13218-020-00637-y.
25. D. Walton, A dialogue system specification for explanation, Synthese 182 (3) (2011) 349–374. doi:10.1007/s11229-010-9745-z.
26. P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence, arXiv:1903.02409 [cs]ArXiv: 1903.02409 (Mar 2019).
27. H. Nori, S. Jenkins, P. Koch, R. Caruana, Interpretml: A unified framework for machine learning interpretability, arXiv preprint arXiv:1909.09223 (2019).
28. J. Klaise, A. Van Looveren, G. Vacanti, A. Coca, Alibi: Algorithms for monitoring and explaining machine learning models.
29. P. Biecek, Dalex: Explainers for complex predictive models in r, Journal of Machine Learning Research 19 (84) (2018) 1–5.
30. M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
31. A. Dhurandhar, V. Iyengar, R. Luss, K. Shanmugam, A formal framework to characterize interpretability of procedures, arXiv:1707.03886 [cs]ArXiv: 1707.03886 (Jul 2017).