Data Mining and Knowledge Discovery
https://doi.org/10.1007/s10618-022-00872-x

Explanatory artificial intelligence (YAI): human-centered explanations of explainable AI and complex data

Francesco Sovrano[1] · **Fabio Vitali**[1]

Received: 15 June 2021 / Accepted: 15 September 2022

Abstract

In this paper we introduce a new class of software tools engaged in delivering successful explanations of complex processes on top of basic Explainable AI (XAI) software systems. These tools, that we call cumulatively Explanatory AI (YAI) systems, enhance the quality of the basic output of a XAI by adopting a user-centred approach to explanation that can cater to the individual needs of the explainees with measurable improvements in usability. Our approach is based on Achinstein's theory of explanations, where explaining is an illocutionary (i.e., broad yet pertinent and deliberate) act of pragmatically answering a question. Accordingly, user-centrality enters in the equation by considering that the overall amount of information generated by answering all questions can rapidly become overwhelming and that individual users may perceive the need to explore just a few of them. In this paper, we give the theoretical foundations of YAI, formally defining a user-centred explanatory tool and the space of all possible explanations, or *explanatory space*, generated by it. To this end, we frame the *explanatory space* as an hypergraph of knowledge and we identify a set of heuristics and properties that can help approximating a decomposition of it into a tree-like representation for efficient and user-centred explanation retrieval. Finally, we provide some old and new empirical results to support our theory, showing that explanations are more than textual or visual presentations of the sole information provided by a XAI.

Keywords Explanatory AI · Explainable AI · Theory of explanations · Human–computer interaction · Human-centred explanations

Responsible editor: Martin Atzmüller, Johannes Fürnkranz, Tomas Kliegr, Ute Schmid

✉ Francesco Sovrano
francesco.sovrano2@unibo.it

Fabio Vitali
fabio.vitali@unibo.it

[1] Department of Computer Science and Engineering (DISI), University of Bologna, Bologna, Italy

Published online: 10 October 2022

Abbreviations

AI	Artificial intelligence
XAI	Explainable AI
YAI	Explanatory AI
SUS	System usability scale
NCS	Need for cognition score
2EC	2nd-level exhaustive explanatory closure
RL	Reinforcement learning

1 Introduction

Is there anything more in explaining a complex process than letting humans have direct access to the information necessary to understand it? That is to say, is there anything more in explaining than the explanatory information itself as produced by a XAI? Well, we certainly believe so. Slowly but steadily we are understanding that XAI approaches are just pathways, necessary but highly incomplete routes to understanding complexity.

The purpose of this paper is to build on the growing awareness that good explanations start from, but are not, the output of however improved forms of XAI, but constitute a complementary and vastly different endeavour. We identify a new class of tools, that we named Explanatory AI (YAI), and that we deem to be the missing connection between XAI and human understanding of complex behaviours of digital systems, and that we assume can be considered somehow independent and separate from the XAI tool they explain. In particular, regardless of the adopted model of explaining, or the direction taken to produce explanations, we aim to prove that user-centred explanations are necessary for any sufficiently complex system, because generic explanations are either inadequate or too burdensome for most needs.

By way of a metaphor, we consider the goal-oriented aspects of providing explanations to a complex project akin to searching for information about a bank robbery using the recording of a closed circuit camera (CCC) in front of the bank entrance. The fact that the CCC system is able to store hours of good quality video is instrumental, but not sufficient, to determine the usefulness of the CCC service. For instance, investigators may know the time of the robbery but not the face of the robbers, or may know their faces but not how long they waited outside of the entrance, or the number of people that entered, or the direction they fled to, or the licence plate of the car they drove, or whether they had been there before for recognisance of the place, or even the same questions could be made not for the bank robbery but for a night burglary at the liquor store two doors down the bank, etc. It is the specific goal of the investigator, and not the quality and technicalities of the recording machine, that determines the questions that the CCC system must provide an answer to, and therefore simply providing 48 h of good quality video with no tool for navigating it other than watching it in $1\times$ speed, is not enough.

Similarly, our objective is to design and prototype YAI to extend and improve the reach of explanations, just like advanced search and playback functions in modern video players improve the reach of CCC systems. In fact, one can imagine lots of

examples (other than the CCC system) where searching for an explanation is equivalent to looking for a needle in a haystack of explainable information also coming from one or more XAI. For example, suppose that the user of a complicated AI-based credit approval system deployed by a bank needs to know why his/her loan application was rejected and what to do in order to have it accepted instead. In this case, the bare output of a XAI might not be enough for fully understanding the details of how to change the outcome of a loan application, while the whole documentation about how a credit score is computed and used for approval might be too burdensome, complicated and technical for a lay person. In fact, the XAI might be able to tell the applicant that she/he was rejected because of an excessive amount of "credit inquiries", but it cannot tell how to reduce the number of such inquiries, or that only the "hard inquiries" should be avoided, or what is a "hard inquiry", etc.

In particular, it appears from preliminary studies, as the one by Liao et al. (2020), that users are interested in asking a variety of different questions about an AI-based system, pointing to complex and heterogeneous needs for explainability that go beyond the output of a single XAI. More specifically, on the one hand we frame XAI as that component of an explanatory tool that generates explainable information to supplement the content of a possibly large amount of documents (i.e., manually created) that explain the details of an AI-based system. On the other hand we frame YAI as the component responsible for selecting, from a large collection of explainable information, the most relevant and useful explanations for the user. So that, in other words, YAI is needed on top of XAI whenever a large-enough amount of (explainable) information about a system has to be conveyed in a user-centred, *pragmatic*, manner to a person of interest. Importantly, this type of scenario is not unrealistic if we consider that the European Artificial Intelligence (AI) Act[1] (which will come into force by 2024 in the EU) is likely to require developers of a high-risk AI-based system to provide sufficiently detailed technical documentation about the system and the underlying AI (Sovrano et al. 2022b).

Therefore, the research question we answer with this paper is how to generate user-centred and goal-driven explanations, out of a sufficiently large collection of explainable information, with a software, when no assumption about the users can be made. To this end, we rely on a recent extension of Achinstein's theory of explanations (Sovrano and Vitali 2022a, 2021a; Achinstein 2010), where explaining is an *illocutionary* act of *pragmatically* answering a question. In particular, as discussed also in Sect. 2, we mean that there is a subtle and important difference between simply "answering questions" and "explaining", and this difference is in both *illocution* (i.e., informed and pertinent answering not just the main question, but also other questions such as `why`, `how`, `when`, `what`, `who`, etc.) and *pragmatism* (i.e., tailoring explanations to the specific background knowledge, needs and goals of the person receiving them).

Therefore, building on the concepts of *pragmatism* and *illocution*, in Sect. 3 we provide the theoretical foundations of YAI. In particular, we formally define what is

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

an explanatory process and how it generates an *hypergraph*[2] of questions and answers called *explanatory sub-space*. Furthermore, we define the main properties of a user-centred *explanatory sub-space* and some heuristics to decompose it into a tree-like representation for efficient and user-centred explanation retrieval.

All these contributions are meant to show that user-centred explanations are better understood as individual goal-driven paths within a huge, and possibly unbounded, *explanatory space*, and the direction, length and components of each of these paths directly and substantially depend on the type of need, the goal and the background of the human it is meant for. Hence, in Sect. 5 we show some old and new empirical results in support of the proposed theory, also publishing the related source code on GitHub[3].

2 Background

In this section we provide enough background to understand and support the rest of the paper.

Hereby we discuss how Achinstein's theory of explaining as a question answering process is compatible with existing XAI literature, highlighting how deep is in this field the connection between answering questions and explaining.

2.1 Achinstein's theory of explanations and archetypal questions

Achinstein (1983), was one of the first scholars to analyse the process of generating explanations as a whole, introducing his philosophical model of a pragmatic explanatory process. According to the model, explaining is an illocutionary act of question answering (Achinstein 2010). Illocution here means that explaining comes from a clear intention of producing new understandings in a specific explainee (the person receiving the explanation) by providing a correct content-giving answer to an open question. Therefore, according to this view, answering by "filling the blank" of a predefined template answer (as most of one-size-fits-all explanations do) prevents the act of answering from being explanatory, by lacking illocution. In fact, in some contexts, highlighting logical relationships may be the key to making the person understand. In other contexts, pointing at causal connections may do the job, and in still further contexts, still other things may be called for.

Building on these ideas, recent efforts towards the automated generation of explanations (Sovrano and Vitali 2021a, 2022a), have shown that it may be possible to define *illocution* in a more "computer-friendly" way and consequently measure the degree of explainability (Sovrano and Vitali 2022b, 2021b). As stated by Sovrano and Vitali (2021a), illocution in explaining involves informed and *pertinent* answers not just to the main question, but also to other (archetypal) questions of various kinds, even unrelated to causality, that are relevant to the explanations.

[2] A hypergraph is a generalization of a graph in which an edge can join any number of vertices (Bretto 2013).

[3] https://github.com/Francesco-Sovrano/YAI4Hu.

Definition 1 (*Archetypal question*) An archetypal question is an archetype applied on a specific aspect of the explanandum[4]. Examples of archetypes are the interrogative particles (`why`, `how`, `what`, `who`, `when`, `where`, etc.), or their derivatives (`why-not`, `what-for`, `what-if`, `how-much`, etc.), or also more complex interrogative formulas (`what-reason`, `what-cause`, `what-effect`, etc.). Accordingly, the same archetypal question may be rewritten in several different ways, as "why" can be rewritten in "what is the reason" or "what is the cause".

2.2 XAI and question answering

The idea of answering questions as explaining is not new to the field of XAI and it is also quite compatible with everyone's intuition of what constitutes an explanation. Despite the different types of explainability one can choose (i.e., rule-based, case-based), it appears to be always possible to frame the information provided by explainability with one or (sometimes) more questions. In fact, it is common to many works in the field (Ribera and Lapedriza 2019; Lim et al. 2009; Miller 2019; Gilpin et al. 2018; Dhurandhar et al. 2018; Wachter et al. 2018; Rebanal et al. 2021; Jansen et al. 2016; Madumal et al. 2019) the use of archetypal or more punctual questions to clearly define and describe the characteristics of explainability, regardless its type.

For example, Lundberg et al. (2020) assert that the local explanations produced by their TreeSHAP (an *additive feature attribution* method for feature importance) may enable "agents to predict why the customer they are calling is likely to leave" or "help human experts understand why the model made a specific recommendation for high-risk decisions". While Dhurandhar et al. (2018) clearly state that they designed CEM to answer the question "why is input x classified in class y?". Also Rebanal et al. (2021) propose and study an interactive approach where explaining is defined in terms of answering `why`, `what` and `how` questions. For further concrete examples of how archetypal questions are related to XAI algorithms, we point the reader to a recent survey by IBM Research (Liao et al. 2020).

Anyway, these are just some examples, among many, of how Achinstein's theory of explanations is already implicit in existing XAI literature, highlighting how deep is in this field the connection between answering questions and explaining. A connection that has been implicitly identified also by authors like Lim et al. (2009), Miller (2019), Gilpin et al. (2018) that analysing XAI literature were able to hypothesise that a good explanation, about an automated decision-maker, answers at least `what`, `why` and `how` questions.

Nonetheless, despite its compatibility, practically none of the works in XAI ever explicitly mentioned Ordinary Language Philosophy, preferring to refer Cognitive Science instead (Sovrano et al. 2022b, 2021). This is probably because Achinstein's illocutionary theory of explanations is seemingly difficult to be implemented into a software, by being utterly pragmatic and by failing to give a precise definition of *illocution* as intended for a computer program. In fact, *user-centrality* is challenging and sometimes not clearly connected to XAI's main goal of "opening the black-box" (e.g., understanding how and why an opaque AI model works).

[4] It means "what has to be explained" in Latin.

Therefore it appears that XAI is more focused on producing explainable software and explanations that generally follow a one-size-fits-all approach, by answering well to just one (or sometimes few) punctual pre-defined questions, as suggested by the exploratory study carried out by Liao et al. (2020). In particular, Liao et al. (2020) show that user needs for explainability are manifold, i.e., they could be about the terminology, the performance of the system, its output and input, etc. Furthermore, Liao and Varshney (2021) also show that no single XAI appears to be able to cover all of the identified user needs, suggesting that in the most generic scenario a plethora of different XAI might be required for better explaining an AI-based system.

To this end, there is some work in the intersection of XAI and Human–computer Interaction, called Human-centred XAI, trying to overcome the one-size-fits-all nature of XAI for more "human-centrality". What appears to be common to Human-centred XAI is that proper explaining involves some kind of conversation between the explainer and the explainee. For example, Dazeley et al. (2021) propose to define levels of explanation and describe how they can be integrated to create a human-aligned conversational explanatory system that provides more insights into an agent's: beliefs and motivations; hypotheses of other (human, animal or AI) agents' intentions; interpretation of external cultural expectations; or, processes used to generate its own explanation. Also, Vilone and Longo (2022) propose some kind of conversational, argument-based, explanation system for a machine-learned model, in order to enhance its degree of explainability by employing principles and techniques from computational argumentation that frame the act of explaining as akin to non-monotonic logic. Finally, we can find also applications of Human-centred XAI for education, as in Khosravi et al. (2022). In particular, Khosravi et al. (2022) present a framework that considers six key aspects in relation to explainability for studying, designing and developing educational AI tools. These key aspects focus on the stakeholders, benefits, approaches for presenting explanations, widely used classes of AI models, human-centred designs of the AI interfaces and potential pitfalls of providing explanations.

3 Explanatory artificial intelligence: theoretical foundations

The purpose of this paper is to build on the growing awareness that good explanations start from, but are not, the output of however improved form of XAI, but constitute a complementary and vastly different endeavour. So, in this section we will present our main contribution: the theoretical foundations of YAI. To do so, we start from discussing what is a one-size-fits-all explanation and why it is not sufficient for user-centrality, then we draw the difference between XAI and YAI providing a formal definition of user-centred explanatory tool and *explanatory space*. Soon after, we discuss the main properties of an *explanatory space* and some heuristics to explore an *explanatory space* in a user-centred and efficient way.

3.1 User-centrality and the problem with one-size-fits-all explanations

Computational irreducibility is typical of emerging phenomena such as physical, biological and social ones (Beckage et al. 2013). For these systems it is possible to simulate every step of the evolution of the system's behaviour, but it is not possible to predict a result of this simulation without letting the system take each evolutionary step. Thus, standing on the definition given by Zwirn and Delahaye (2013), user-centrality in explaining is *computationally irreducible* because generally speaking nothing besides the user itself (while unfolding an explanation) can predict whether an explanation is really useful, usable, satisfactory, etc.

Therefore, we take a strong stand against the idea that static, one-size-fits-all approaches to explanation have a chance of being pragmatic (i.e. user-centred). This is to say that XAI-based tools answering just a few specific `why`, `how` and `what` questions are not enough for properly explaining in a user-centred way. In fact, by definition one-size-fits-all explanations are based on the idea that the same piece of information can fit all, therefore assuming that it would be usable and useful *a priori* for anybody. The main types of one-size-fits-all explanations are the following:

- *Normal XAI-based explanations* answering one only question or just few.
- *Selected narratives* answering only one type of questions, e.g., How-Why Narratives answering only `how` or `why` questions.
- *Exhaustive explanatory closures* answering by giving as answer a a whole library of content (as analogy) surely containing the sought answer. More specifically, a 1st-Level Explanatory Closure is about giving as explanation all the available information immediately. A 2nd-Level Explanatory Closure is explaining the 1st Level Explanatory Closure itself with additional available information about aspects of the explanandum becoming apparent after one interaction with it. A 3rd-Level Explanatory Closure is like a 2nd-Level one but all the information is given to the user after two levels of interaction, etc.

Indeed, a one-size-fits-all explanation, to really fit all, should contain all the possible answers to all the possible questions of all the possible users (e.g., an Exhaustive 2nd-Level Explanatory Closure). But this kind of explanation would be useless for a human, being overwhelming in size and content as soon as the complexity of the explanandum increases beyond a fairly trivial threshold. In other terms, an explainable dataset or system by itself it is not a user-centred explanation, and a generic Nth-Level Explanatory Closure is necessary for user-centrality. In fact, the interest of a user in the output of an explanation system often may lie in a few short statements out of the hundreds of thousands that the explanation system may be able to generate, and these few lines depend on the function that the user gives to the explanation. This is why we must assume that, in general, the purpose of the explanation is known to the user but not to the explanation system, and it cannot be decided in advance but it becomes knowable only during the evolution of the task for which the explanation is required.

For example, a complex big-enough *explainable* software can be super hard to *explain*, even to an expert, and the optimal (or even sufficient) explanation might change from expert to expert. In this specific example, an explainable software is necessary but not sufficient for explaining.

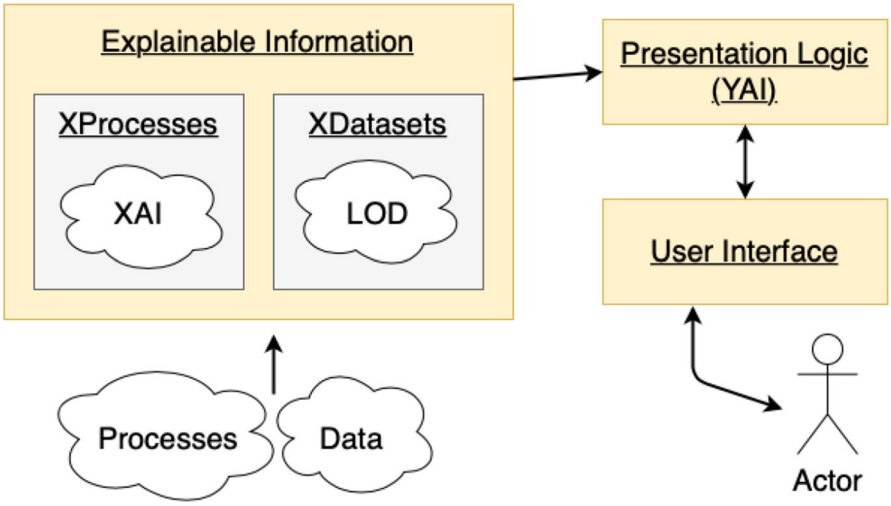

Fig. 1 XAI versus YAI: an abstract model of explanatory tool. This model shows how to decompose the flow of explanatory information that moves from raw representations of processes/data to the explainee (or actor). Raw data are refined into explainable datasets (e.g., Linked Open Data, LOD in short). Raw processes are refined into explainable processes. Explainable information can be used by YAI to generate pragmatic explanations

3.2 XAI versus YAI

A user-centred explanatory tool requires to provide goal-oriented explanations. Goal-oriented explanations imply explaining facts that are relevant to the user, according to her/his background knowledge, interests and other peculiarities that make him/her a unique entity with unique needs that may change over time. Therefore, to model a user-centred explanatory process we need to:

- Disentangle *making things explainable* (i.e., XAI) from *explaining* (i.e., YAI): in a way, this is tantamount to separating the presentation logic from the application logic. In fact only explaining has to be user-centred.
  In this sense, we like to say that we need both the Xs and the Ys of AI[5].
- Design a presentation logic that allows personalised explanations out of some explainable information.

In Fig. 1 we show a simple model of an explanatory tool, obtained by our own need to clearly separate between explainability and explanations. In particular, to increase the overall cohesion of the explanatory system, in this model we require an explicit logical separation between the functionalities related to *producing explainable information*, and those related to *producing pragmatic explanations*. In addition, we envision another logical separation in the production of actual explanations between *building explanations* (i.e., the presentation logic) and *interfacing with users*. Independently, *producing explainable information* should be separated in *generating explainable processes* and *producing explainable datasets*.

[5] XX and XY are the human chromosomes responsible for biological gender.

One of the most interesting benefits coming from this distinction of YAI from XAI is that it would meet the *Single Responsibility Principle* (Martin 2002), making easier to integrate an explanatory layer in an existing application layer (without changing the latter). What we can see is that nowadays in many XAI applications, intended as explanatory tools, the presentation logic is not explicitly separated from the application logic.

3.3 Definition of user-centred explanatory process and space: the SAGE properties

We believe that an explanatory tool is *an instrument for articulating explainable information* into an explanatory discourse. This definition of explanatory tool is drawn from the essential best-practices of scientific inquiry (Berland and Reiser 2009), involving:

- *Sense-making of phenomena* classical question answering to collect enough information for understanding, thus building an explainable explanandum (perhaps through XAI).
- *Articulating understandings into discourses* re-ordering and aggregation of explainable information to form an explanatory narrative or more generally a discourse to answer research questions.
- *Evaluating* pose and answer questions about the quality of the presented information (e.g., argument them in a public debate).

More formally, we propose the following definition of explanatory process, taking under consideration that for user-centrality an explainee must be able to specify as input of the process her/his goals, otherwise not inferable due to the computational irreducibility of the phenomenon.

Definition 2 (*Explanatory process*) Let an *explanans* (plural is *explanantia*) be a text in natural language (i.e., English) answering one or more questions. A user-centred *explanatory process* or *explanatory discourse articulation* (stylised in Fig. 2) is a function p for which $p\,(D, E_t, i_t) = E_{t+1}$, where:

- D is the *explanandum* a set of explainable pieces of information; a set of *answers* organised to build archetypal explanations that are useful to the explainee.
- E_t is the *explanans*, at time step $t \geq 0$. E_t can be any meaningful rephrasing of the information contained in D.
- i_t is the *interaction of the explainee* at step t.

We can iteratively apply p, starting from an initial explanans E_0, until satisfaction. The user interaction i is a tuple made of an action a taken from the set A_p of possible actions for p, and a set of auxiliary inputs required by the action a. Whenever A_p allows any explainee to specify its needs and goals to maximise the usability of E_{t+1}, p is said to be user-centred.

In order to understand how to implement such a user-centred explanatory process p we need first to define the characteristics of the space of all the possible explanations that can be generated by p. We call this space of explanations the *explanatory subspace* of p.

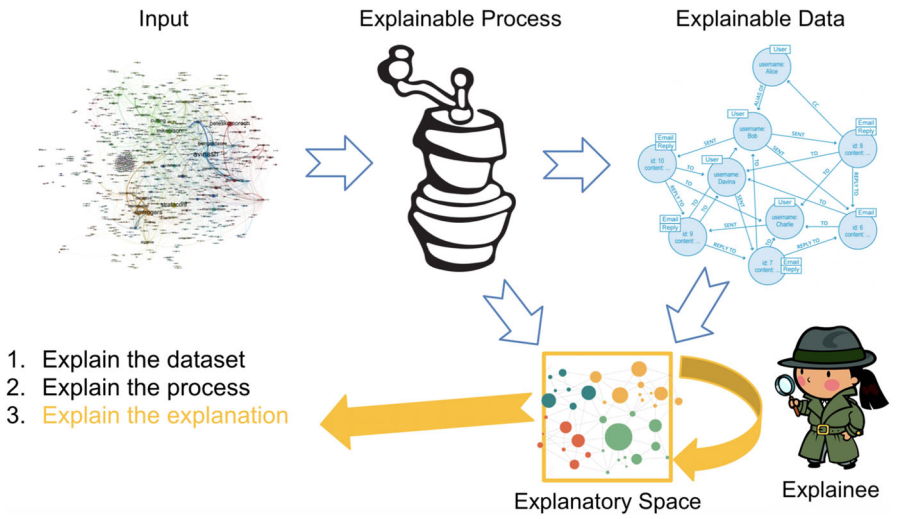

Fig. 2 Stylized interactive explanatory process: a user-centred explanatory process explains an explanandum to an explainee, thus producing as output an explanans that is meaningful for the specific explainee

Definition 3 (*Explanatory sub-space*) An *explanatory sub-space* is a *hypergraph* $H_p = (\xi_p, \epsilon_p)$ of interconnected explanantia reachable by an explainee interacting with a process p, given an explanandum D, a set of actions A_p and an initial explanans E_0. Thus, the set of *hyperedges* ϵ_p is the set of all possible explanantia that can be generated by p about D:

$$\epsilon_p = \{E_0\} \cup \{\forall u > 0, \forall i_u \in A_p \mid p\,(D, i_u, E_{u-1})\} \tag{1}$$

While, the set of nodes ξ_p is the set of questions q and answers a covered by the explanantia[6]:

$$\xi_p = \{\forall E \in \epsilon_p, \forall < q, a > \in E \mid q\} \cup \{\forall E \in \epsilon, \forall < q, a > \in E \mid a\} \tag{2}$$

This leads us to the definition of an *explanatory space*.

Definition 4 (*Explanatory space*) An *explanatory space* is the *hypergraph* $H = (\xi, \epsilon)$ resulting from the union of the *explanatory sub-spaces* of each p in the set of all the possible explanatory processes P. In particular we have that:

$$\epsilon = \{\forall p \in P, \forall E \in \epsilon_p \mid E\} \tag{3}$$

$$\xi = \{\forall p \in P, \forall i \in \xi_p \mid i\} \tag{4}$$

[6] It is always possible to represent natural language sentences as networks of questions and answers. Indeed, casting the semantic annotations of individual propositions as narrating a question-answer pair recently gained increasing attention in computational linguistics (He et al. 2015; FitzGerald et al. 2018; Michael et al. 2018; Pyatkin et al. 2020).

Therefore, according to Definition 2, we have that an *explanatory sub-space* H_p, in order to be user-centred, should be *adaptable*. More specifically, it should also be:

- *Sourced* bound by the explanandum D. The space should be a description of D.
- *Adaptable* bound by the narrative purposes of the explainee and his/her queries i. The space should be structured in a way that would minimise the number of queries for the explainee to achieve its objective.
- *Grounded* bound by the explanatory process p as *illocutionary question answering*. The space should be structured in order to effectively and efficiently answer questions.
- *Expandable* bound by the characteristics of the web of explanantia E. The space should form a coherent information network that can be explored and described throughout linguistic structures such as narration or, more generally, discourse.

We will refer to these properties of an *explanatory sub-space* as the SAGE properties and we will use them to define a set of actions A_p to embed user-centrality in an explanatory process.

3.4 Efficient exploration of explanatory spaces: the ARS heuristics

In graph theory, tree decompositions are used to speed up solving certain computational problems on graphs, and more generally hypergraphs (Gottlob et al. 2016). Indeed, many instances of NP-difficult problems on graphs can be efficiently solved via tree decomposition (Bachoore and Bodlaender 2007). So, if an *explanatory space* is a hypergraph, then any efficient explanatory process p should be able to approximate a decomposition of such hypergraph into some kind of hypertree, allowing the explainee to efficiently navigate through the vast underlying space and find the answers he/she is seeking.

More specifically, decomposing an *explanatory space* H into a hypertree is equivalent to ordering and prioritising all the explanantia and the pieces of information within the explanantia so that the explainee can efficiently navigate and read the *explanatory space* from the root (i.e., any initial explanans) to the leaves of its decomposition. Though, several different hypertree decompositions might exist for the same *explanatory space* with no assurance that all of them are effective as they should at explaining to a human. That is because the output of an explanatory process should be pragmatic, user-centred. In particular, a good explanatory process should be able to adapt to the needs of a human explainee with a specific background knowledge and specific goals.

To this end, Sovrano et al. (2020b) proposed a few heuristics for user-centred exploration of an *explanatory space*, designed to maximise the *adaptability* of the explanatory process. These heuristics are namely:

- ***A**bstraction* for identifying the nodes (also called *explanandum aspects*) of the tree decomposition of the *explanatory space*. This is done by aggregating explanations according to some kind of taxonomy defining a hierarchy of abstractions.
- ***R**elevance* for ordering the information internal to explanandum aspects according to its relevance to the goal of the explainee.

– ***S**implicity* for selecting the viable edges of the tree decomposition and the information internal to an *explanandum aspect*. This can be done by filtering the content of the explanandum aspects or also by prioritising certain abstractions over others.

We will refer to them as the ARS[7] heuristics.

By definition, both the SAGE properties and the ARS heuristics (the SAGE-ARS model) pose some constraints on the ways of interaction that are allowed for exploring the *explanatory space*. In other words, these constraints help to define a set A_p of actions that would allow the user to explore in a user-centred way a decomposition of the whole *explanatory space* starting from an initial explanans E_0 (i.e., the output of a XAI), according to Definition 2.

Following the definition of explanatory tool drawn from the best-practices of scientific inquiry described in Sect. 3.3, some primitive actions that can be implemented are:

– *Open Question Answering* for *sense-making of phenomena* the user writes a question and then it gets one or more relevant punctual answers.
– *Aspect Overviewing* for *articulating understandings* the user selects an aspect of the explanandum (i.e., contained in a punctual answer) receiving as explanation a set of *relevant* archetypal answers involving other different aspects that can be explored as well. Archetypal answers can also be expanded, increasing the level-of-detail according to the *simplicity* heuristic.
– *Argumentation* for *evaluating* the user evaluates the explanations, identifying counter-arguments or weak points that can be used for further (automated) reasoning.

The first two primitive actions are said to be the main primitives for explaining, because aligned to *sense-making* and *articulation of understandings*. In fact, we previously defined an explanatory tool as "an instrument for articulating explainable information".

Specifically, we can see an overview as an appropriate summary of an explanandum aspect, while a specific answer can be seen as a sequence of information (a path) that can span more explanandum aspects.

So, we have that for each SAGE property, we can identify a set of SAGE commands for these primitive actions that may be used by the explainee during the explanatory process, as suggested also by Sovrano et al. (2020b):

– "*Sourcing* commands" used to access the source of an explanation fragment (e.g., a law, a scholarly paper, a rule, etc.), i.e., as shown in the right sub-figure of Fig. 3.
– "*Adapting* commands" used to provide the explanatory process with sufficient information to model the background knowledge and the goals of the explainee in order to personalise the content of the explanations.
– "*Grounding* commands" used to ask questions.
– "*Expanding* commands" used to navigate the tree decomposition of the *explanatory space* and get a partial view of it. Examples of expanding commands might be: (1) *Get Overview* it opens an explanatory overview about a concept. (2) *More* it

[7] "Ars" means art in Latin.

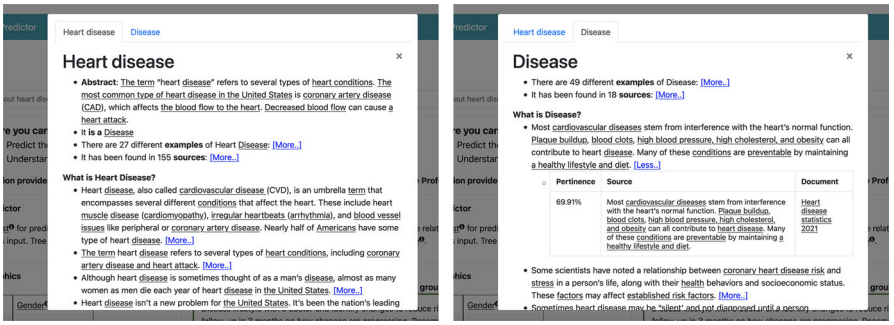

Fig. 3 Heart disease predictor & YAI4Hu: a screenshot showing the overview modal of YAI4Hu containing two cards about as many aspects of the heart disease predictor. The first card is about "Heart Diseases", the second card is about "Diseases"

shows additional details available in the explanation but currently hidden from the interface, because of the *simplicity* policy. A concrete example is shown in Fig. 3. (3) *Less* it removes the information added with the "More" and "Get Overview" commands.

Details on how to build an explanatory tool based on our theoretical framework of YAI (i.e., YAI4Hu) can be found in Sovrano and Vitali (2022a).

4 YAI4Hu: a possible implementation of the SAGE-ARS Model

An example of YAI implementing the SAGE-ARS model is YAI4Hu (Sovrano and Vitali 2022a), an explanatory tool approximating a Nth-Level Explanatory Closure, as discussed in Sect. 3.1. More specifically, YAI4Hu covers both *Aspect Overviewing* and *Open Question Answering*, but not *Argumentation*. In particular, YAI4Hu is a fully automatic explanatory tool, relying on pre-existing documentation about an AI system (i.e., generated by a XAI or manually created) to extract a special knowledge graph out of it for efficient answer retrieval. So that an explainee can ask questions about the content of the documentation or explore it by means of *Aspect Overviewing*. More specifically, *Open Question Answering* is implemented with an answer retrieval system, i.e., the system described in Sovrano et al. (2020a). Furthermore, also *Aspect Overviewing* is implemented with an answer retrieval system whose questions though are not asked by the explainee but are indeed instances of archetypal questions about the aspect to overview. So that the explainee can specify which aspect to overview and then get an explanation about it in the form of answers to a set of pre-defined archetypal questions (e.g., why is this aspect/concept important, what is this aspect/concept, etc.).

In YAI4Hu, through *Aspect Overviewing*, a user can navigate the whole *explanatory space* reaching explanations for every identified aspect of the explanandum.

In fact, every sentence presented to the user is annotated (as in Sovrano and Vitali 2021a) so that users can select which aspect to overview by clicking on the annotated syntagms. Annotated syntagms are clearly visible because they have a unique style that makes them easy to recognize, as shown in Fig. 3.

After clicking on an annotation, a modal opens showing a card with the most relevant information about the aspect (see Fig. 3). This is in accordance with the *relevance* heuristic.

The most relevant information shown in a card is:

- A short description of the aspect (if available): abstract and type.
- The list of aspects taxonomically connected.
- A list o archetypal questions and their respective answers ordered by estimated pertinence. Each piece of answer consists of an *information unit* and its summary.

All the information shown inside the modal is annotated as well. This means (for example) that clicking on the taxonomical type of the aspect, the user can open a new card (in a new tab) displaying relevant information about the type, thus being able to explore the *explanatory space* according to the *abstraction* policy, as shown in Fig. 3. On the other side, the *simplicity* policy is ensured by the "More" and "Less" buttons (that allow to increase/decrease the level of detail of information) and by the fact that not all the words in the explanantia are linked to an overview despite being nodes of the *explanatory space*.

5 Empirical results

Overall, our proposed model of YAI is defined around the idea that explaining is somehow akin to exploring a possibly unbounded hypergraph of questions and answers called *explanatory space*. This *explanatory space*, to be efficiently explored through an explanatory process, is then decomposed into some form of tree, allowing the explainee to navigate through the vast underlying space and find the answers she/he is seeking.

In particular, our central hypothesis is that an explanatory process that implements the ARS heuristics and the SAGE commands is user-centred, therefore producing better explanations through an easy-to-navigate tree decomposition of the *explanatory space*. In other words, our hypothesis is equivalent to say that not all the decompositions of an *explanatory space* are equally useful to a human (if no assumption is made about the background knowledge of the explainee), and that the SAGE-ARS model can produce a decomposition that is user-centred and useful. In this sense, one can say that the focus of our work is on explanations for lay users (e.g., data subjects).

To verify our hypothesis, we collected some old and new empirical results presented throughout this section. In particular, our hypothesis is verified through ad hoc user studies comparing the user-centrality of baseline, one-size-fits-all, explanatory tools to that of tools implementing the SAGE-ARS model (i.e., YAI4Hu), measured in terms of *usability*.

In short, we adopt the definition of *usability* as the combination of *effectiveness*, *efficiency*, and *satisfaction*, as per ISO 9241-210. ISO 9241-210 defines *usability* as the "extent to which a system, product or service can be used by specified users to achieve specified goals with *effectiveness*, *efficiency* and *satisfaction* in a specified context of use" (International Organization for Standardization 2010). *Effectiveness* ("accuracy and completeness with which users achieve specified goals") and *efficiency* ("resources

used in relation to the results achieved. […] Typical resources include time, human effort, costs and materials.") can be assessed through objective measures (in our case, pass vs. fail at domain-specific questions and time to complete tasks, respectively).

Satisfaction, defined as "the extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations", is a subjective component and it needs a direct confrontation with the user (in our case performed with a System Usability Scale questionnaire, or SUS in short; Brooke 2013). Importantly, in all the user studies mentioned in this section, it was made clear to all participants what was their expected objective (i.e., to get an explanation; to complete a quiz with the best score possible), so that it was possible to properly measure *satisfaction* as the ability of the system to meet the goals of the user. For this reason, users were explicitly and immediately informed when failing or succeeding to meet their expected goals, so that a user could know whether he really acquired the explanation he was supposed to seek, thus being satisfied. For this reason, we made sure to not pay or reward the participants. In fact, if participants only participated in the study because they would get paid/rewarded, their goal would be to get money as fast as possible and not to get an actual explanation.

5.1 Old empirical results

Overall, when no major assumption is made about the background knowledge of an explainee, the empirical results produced by Sovrano and Vitali (2022a) to evaluate YAI4Hu have shown that the SAGE-ARS model qualitatively and quantitatively outperforms in terms of usability the following one-size-fits-all explanatory tools (also described in Sect. 3.1):

- *A 2nd-level exhaustive explanatory closure (2EC)* a rather static tool showing the output of a XAI and the whole documentation about the explanandum (hundreds of web-pages of explanatory contents) exhaustively.
- *A how-why narrator* a simplified version of YAI4Hu that does not allow *Open Question Answering*, showing only `how` and `why` explanations through *Aspect Overviewing*.

This is proved with a user study involving more than 60 participants and 2 different and complex explananda (better described in Sect. 5.2), showing that YAI4Hu produces statistically relevant improvements on *effectiveness* (hence a P value lower than .05) over the baseline one-size-fits-all tools. Furthermore, the observed improvements in *effectiveness* were also visibly aligned with an increase in *satisfaction*.

In particular, the hypothesis defended by Sovrano and Vitali (2022a) was that, given an arbitrary explanatory process, increasing its *goal-orientedness* and *illocutionary power* results in the generation of more *usable* explanations. In this sense, *goal-orientedness* is defined as the "ability to answer the explicit questions of an explainee" while *illocutionary power* is defined as the "ability to anticipate and answer the implicit (archetypal) questions of an explainee". Importantly, in YAI4Hu, *goal-orientedness* was mostly implemented in terms of *Open Question Answering* while *illocutionary power* as *Aspect Overviewing*.

Though, from our point of view, one of the key results of the user study of Sovrano and Vitali (2022a) is that not every decomposition of the *explanatory space* is maximally useful for a generic human, as hypothesised. Indeed, we have that both the How-Why Narrator and YAI4Hu implement the ARS heuristics, nonetheless YAI4Hu outperforms the How-Why Narrator in terms of *effectiveness* and *satisfaction*. In fact, the main difference between the two explanatory tools is that the How-Why Narrator is less *grounded* (the G of SAGE), not fully implementing the act of explaining as an illocutionary act of question answering.

Seemingly, the How-Why Narrator outperforms a 2EC explanatory tool that is even less *grounded* and that does not fully implement neither *relevance* nor *simplicity*. This shows that, as expected, an overwhelming and shallow decomposition of the *explanatory space* can be not useful for a human. So, even if several different decompositions of the *explanatory space* can be found, not all of them are equally useful and explanatory, suggesting that for properly explaining we might need to fully implement both the ARS heuristics and the SAGE commands.

5.2 New empirical results

The experiments carried out by Sovrano and Vitali (2022a) do not specifically show whether we need a YAI to explain the output of a XAI, or whether a simple XAI-based explanatory tool (that does not implement any of the ARS heuristics or SAGE commands) is sufficient, not being overwhelming as 2EC or complex as the How-Why Narrator or YAI4Hu. Therefore, to understand this point, we ran the user study of Sovrano and Vitali (2022a) also on a simple XAI-based explainer.

In particular, the XAI-based explainer we considered is a one-size-fits-all explanatory tool providing the bare output of a XAI as fixed explanation for all users, together with the output of the wrapped AI, a few extra details to ensure the readability of the results, and a minimum of context. Importantly, assuming that there is no single XAI capable of explaining every detail of an AI, we are not particularly interested in the type of XAI chosen for our experiment. Instead, we are interested in showing that, regardless the type of XAI, any XAI-based explainer cannot explain to a lay person as a user-centred explanatory tool, when no assumption is made about the background knowledge of the explainee.

So, as Sovrano and Vitali (2022a), we considered the following 2 explananda:

– A credit approval system based on a simple Artificial Neural Network and on a XAI called CEM (Dhurandhar et al. 2018). Given the specific characteristics of this system, it is possible to assume that the main goal of its users is about understanding what are the causes behind a loan rejection and what to do to get the loan accepted. The mere output of the XAI can answer to the question: "What are the minimal actions to perform in order to change the outcome of the credit approval system?". Nonetheless many other relevant questions might be to answer before the user is satisfied, reaching its goals. Generally speaking, all these questions can be shaped by contextually implicit instructions (for more details see Sect. 2.1) set by specific legal or functional requirements (Bibal et al. 2021). These questions include: "How to perform those minimal actions?", "Why are these actions so important?", etc..

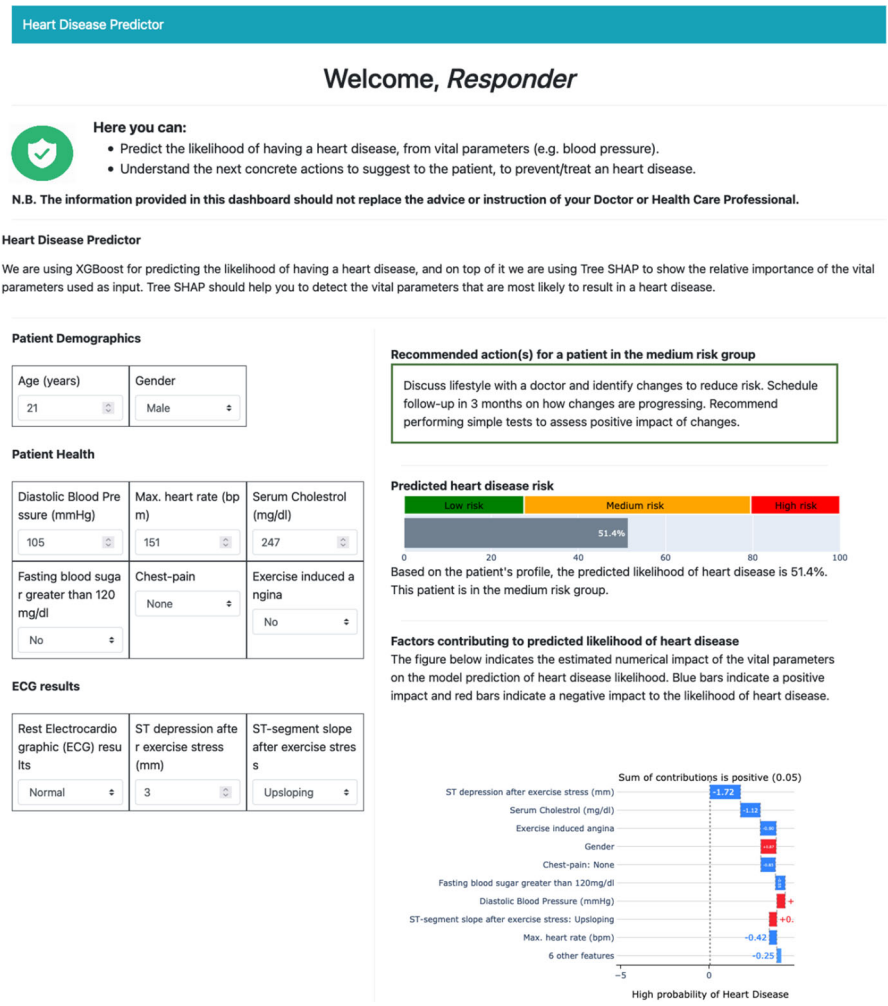

Fig. 4 Heart disease predictor & XAI-based explainer: a screenshot of the XAI-based Explainer explanatory tool for the heart disease predictor

– A heart disease predictor based on an AI called XGBoost (Chen and Guestrin 2016) and on a XAI called TreeSHAP (Lundberg et al. 2020). A screenshot of the heart disease predictor is shown in Fig. 4. This explanandum is about health and the system is used by a first level responder of a help-desk for heart disease prevention. TreeSHAP can be used to answer the following questions: "What are the most important factors leading that patient to this probability of heart disease?", "How important is a factor for that prediction?". The first level responder is responsible for handling the patient's requests for assistance, forwarding them to the right physician in the eventuality of a reasonable risk of heart disease. First level responders get basic questions from callers, they are not doctors but they

have to decide on the fly whether the caller should speak to a real doctor or not. So they quickly use the XAI system to figure out what to answer to the callers and what are the next actions to suggest. This system is used directly by the responder, and indirectly by the caller through the responder. These two types of users have different but overlapping goals and objectives. It is reasonable to assume that the goal of the responders is to answer in the most efficient and effective way the questions of the callers. To this end, the questions answered by TreeSHAP are quite useful, but many other important questions should also be answered, including: "What is the easiest thing that the patient could actually do to change his heart disease risk from medium to low?", "How could the patient avoid raising one of the factors, preventing his heart disease risk to raise?", etc.

In particular, both TreeSHAP and CEM are local explanation methods. More specifically, TreeSHAP (Lundberg et al. 2020) is an algorithm to compute exact SHAP values for Decision Tree based models (e.g. XGBoost). Hence, TreeSHAP belongs to the category of XAIs used for estimating local feature importance rankings, together with SHAP (Lundberg and Lee 2017) and LIME (Ribeiro et al. 2016), two of the most cited XAI algorithms. Interestingly, several XAI of this type (e.g., SHAP and LIME) are commonly model-agnostic and in practice they can be used on any kind of AI. In this sense, using SHAP instead of TreeSHAP in the heart disease predictor would have resulted in the same exact final interface and type of explainability covered. On the other hand, CEM (Dhurandhar et al. 2018) is a generic algorithm for generating contrastive explanations that approximatively show what to modify in order to change the outcome of an AI.

To test the XAI-based Explainer we found 23 new participants coming from the same pool of Sovrano and Vitali (2022a). As in Sovrano and Vitali (2022a), our test evaluated *effectiveness* and *satisfaction* only on people with a normal Need for Cognition Score (NCS)[8]. In fact, it is reasonable to assume that only the most dedicated and focussed users (those with a high NCS) can handle (also with satisfaction) the effort to search in a one-size-fits-all exhaustive explanatory closure as 2EC. On the other end, users with a too low NCS may be more prone to avoid any (also minimally) challenging cognitive task, especially if it involves understanding a complex-enough explanandum, preferring the most naive XAI-based Explainer. For these reasons we believe that it is important to test the usability of a user-centred explanatory tool on people with a normal NCS, as in Sovrano and Vitali (2022a).

More specifically, we measured the *effectiveness* of the XAI-based explanatory tools with the same quizzes and SUS questionnaires of Sovrano and Vitali (2022a). In particular, the questions selected for the quiz on the credit approval system are shown in Table 1 and those of the quiz on the heart disease predictor are shown in Table 2. Furthermore, in order to better understand the relevance to XAI of the questions considered for this user study, we have aligned each question to the types of explainability needs identified by Liao et al. (2020) in their XAI Question Bank. In particular, one can argue that these questions were arbitrarily chosen and might not be of interest for every explainee, and that the answers to these questions might not always

[8] The NCS (Cacioppo and Petty 1982; de Holanda Coelho et al. 2020) is a user characteristic that refers to the user's tendency to engage in and enjoy thinking.

Table 1 Quiz—credit approval system: *Question*, *Archetype*, *QB Type* and *Steps* are shown

Question	Archetype	QB type (Liao et al. 2020)	Steps			
			XAI	2EC	HWN	YAI4Hu
What did the credit approval system decide for Mary's application?	what, how	Output	0	0	0	0
What is an inquiry (in this context)?	what	Terminological	− 1	1	1	1
What type of inquiries can affect Mary's score, the hard or the soft ones?	what, how	How (global)	− 1	1	1	1
What is an example of hard inquiry?	what	Terminological	− 1	1	− 1	1
How can an account become delinquent?	how, why	How to be that	− 1	1	1	1
Which specific process was used by the Bank to automatically decide whether to assign the loan?	what, how	How (global)	0	0	0	0 (no OQA)
What are the known issues of the specific technology used by the Bank (to automatically predict Mary's risk performance and to suggest avenues for improvement)?	what, why	Performance	− 1	1	1	1 (no OQA)

QB Type is the type of question according the taxonomy of user needs for explainability proposed by Liao et al. (2020) in their XAI Question Bank (QB). *Archetype* indicates which interrogative particle is representative of the question. *Steps* is the minimum number of steps (in terms of links to click, overviews to open and/or questions to pose) required by each explanatory tool. Negative *steps* mean that the correct answer cannot be found, while 0 *steps* means that the answer is immediately available without clicking on any link. On the other hand, "no OQA" means that Open Question Answering does not answer correctly to the question. XAI stands for the XAI-based explainer, while HWN stands for the How-Why Narrator

Table 2 Quiz—heart disease predictor: for more details about how to read this table see the caption of Table 1

Question	Archetype	QB type (Liao et al. 2020)	Steps			
			XAI	2EC	HWN	YAI4Hu
What are the most important factors leading that patient to a medium risk of heart disease?	`what, why`	Why	0	0	0	0 (no OQA)
What is the easiest thing that the patient could actually do to change his heart disease risk from medium to low?	`what, how`	How to be that	0	0	0	0 (no OQA)
According to the predictor, what level of serum cholesterol is needed to shift the heart disease risk from medium to high?	`what, how`	How to be that	0	0	0	0 (no OQA)
How could the patient avoid raising bad cholesterol, preventing his heart disease risk to shift from medium to high?	`how`	How to be that	−1	1	2	2
What kind of tests can be done to measure bad cholesterol levels in the blood?	`what, how`	Input	−1	1	−1	1
What are the risks of high cholesterol?	`what, why-not`	Output, What if	−1	1	2	1
What is LDL?	`what`	Terminological	−1	1	2	1
What is Serum Cholestrol?	`what`	Terminological	−1	1	1	1
What types of chest pain are typical of heart disease?	`what, how`	How to still be this	−1	1	1	1

Table 2 continued

Question	Archetype	QB type (Liao et al. 2020)	Steps			
			XAI	2EC	HWN	YAI4Hu
What is the most common type of heart disease in the USA?	what	Social	−1	1	1	1
What are the causes of angina?	what, why	Why	−1	1	2	1
What kind of chest pain do you feel with angina?	what, how	Terminological	−1	1	1	1
What are the effects of high blood pressure?	what, why-not	Why not, Follow-up	−1	1	1	1
What are the symptoms of high blood pressure?	what, why, how	How (global), Input	−1	1	1	1
What are the effects of smoking to the cardiovascular system?	what, why-not	Why not, Follow-up	−1	1	3	1
How can the patient increase his heart rate?	how	How to be that	−1	1	3	1
How can the patient try to prevent a stroke?	how	How to be that	−1	1	3	2
What is a Thallium stress test?	what, why	Terminological	−1	1	3	1

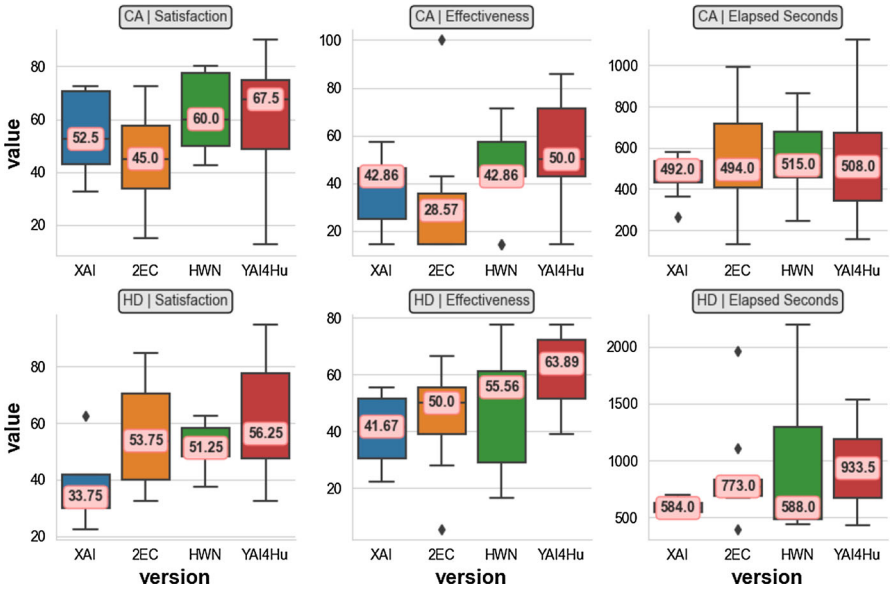

Fig. 5 All versus YAI4Hu—normal NCS: in this figure only participants with a normal NCS are considered. Results are shown in the form of box plots (25th, 50th, 75th percentile, and whiskers covering all data and outliers). The numerical value of medians is shown inside pink boxes. The 1st row is for the heart disease predictor (HD), while the 2nd for the credit approval system (CA). Satisfaction is shown in the 1st column, effectiveness in the 2nd, and elapsed seconds in the 3rd. In this picture we abbreviate XAI-based Explainer as *XAI* and How-Why Narrator as *HWN*

be correctly given by the explanatory tools (i.e., for the adopted AI and XAI providing approximate or wrong information). Nonetheless, regardless the correctness of the explainable information used for generating the explanations, with these quizzes we can analyse the quality of the considered explanatory tools and their presentation logic on a large variety of different explainability needs (i.e., almost all of those identified by Liao et al. (2020), as shown in Tables 1 and 2), without making assumptions about the background knowledge of the explainee[9].

As shown in Figs. 5 and 6, overall YAI4Hu is visibly the most *effective* and *satisfactory* explanatory tool in both the explananda, followed by the How-Why Narrator. The XAI-based Explainer seems to be overall the worst explainer together with 2EC. This is probably for the quizzes containing questions outside the scope of the XAI-based Explainer. Though, even if 2EC can technically answer to all the questions of the quizzes, we have 2EC performing significantly worse than the How-Why Narrator or YAI4Hu.

Yet, interestingly, as shown in Fig. 7, the performance of YAI4Hu is better than (on the heart disease predictor) or equal to (on the credit approval system) that of the plain XAI-based Explainer, in terms of median effectiveness score, even on those explanations specifically targeted by the XAI-based Explainer (questions 1, 2 and 3

[9] The only assumption that is made is that explainees can read and understand English.

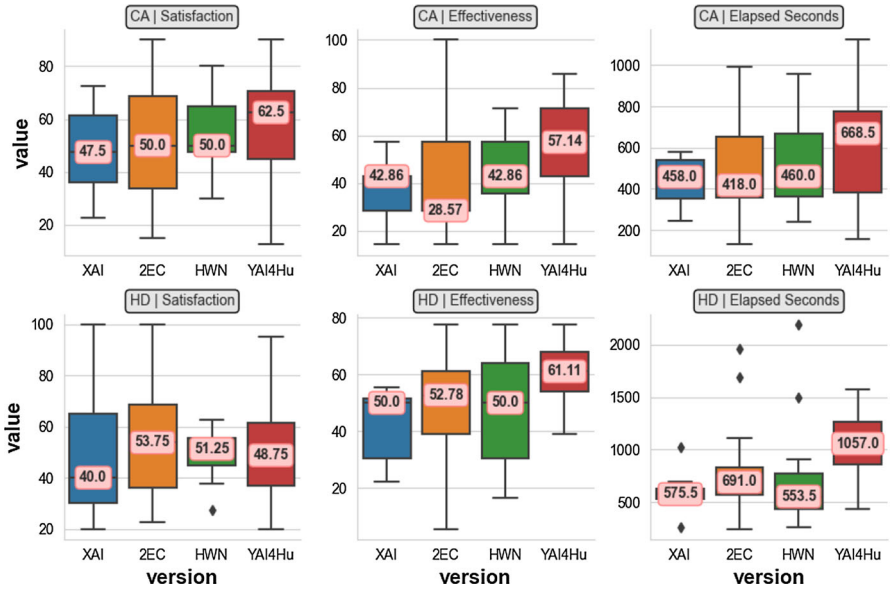

Fig. 6 All versus YAI4Hu—any NCS: in this figure participants with any NCS are considered, not just those with a NCS within the interquartile range. For more details about how to read this figure, see Fig. 5

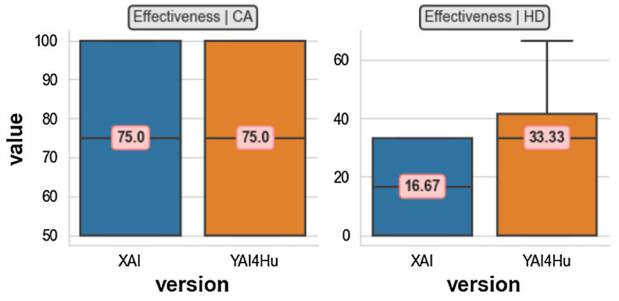

Fig. 7 XAI-based explainer versus YAI4Hu–normal NCS on questions answerable with the information provided by XAI-based Explainer: comparison between XAI-based explainer and YAI4Hu only on those questions answerable with the information provided by XAI-based explainer. For more details about how to read this figure, see Fig. 5. Note that rows and columns in this figure are switched with respect to Fig. 5

in the quiz of the heart disease predictor). This shows that explanations can be more than a textual or visual presentation of the information provided by a XAI.

The difference in usability between participants with normal NCS and non-normal NCS can be noticed by looking at the differences between Figs. 5 and 6. As expected we can see a drop in satisfaction for the more user-centred tools and an increase in effectiveness for 2EC and the XAI-based Explainer (at least on the heart disease predictor).

6 Conclusions, limitations and future work

In this paper, we defended the thesis that good explanations start from, but are not, the output of however improved form of XAI or other one-size-fits-all approaches. In the attempt to tackle the computational irreducibility of user-centrality in explanations, we designed a theoretical model of Explanatory AI (YAI) that clearly disentangles *explaining* from *making things explainable*. This model is based on the definition of the SAGE-ARS, a set of properties, commands and heuristics that an explanatory process can have to allow a pragmatic exploration of an *explanatory space* (the space of all possible explanations reachable by a generic user with a YAI), through a set of pre-defined primitive actions including *Open Question Answering* and *Aspect Overviewing*. Although, importantly, we do not claim that the SAGE-ARS model is the only possible model to achieve user-centrality with an explanatory process.

In particular, in line with Ordinary Language Philosophy, we framed an explanatory tool as a software for illocutionary question answering, formally defining an *explanatory space* as a hypergraph of questions and answers, and a user-centred explanatory process as a function for approximating meaningful tree decompositions of such space through the ARS heuristics and the SAGE commands.

To show that our theory is sound and that not every decomposition of the *explanatory space* is optimal at explaining to humans, we collected some old and new empirical results. These results showed that, when no assumptions are made about the background knowledge of explainees, every one-size-fits-all explanatory tool we considered performs worse than a user-centred explanatory tool by being overwhelming in the amount of information provided (as in 2EC), or by providing too little information for the needs of the explainees (as in the XAI-based Explainers or in the How-Why Narrator).

Therefore, we can conclude with great force that whatever approach is used to describe an *explanatory space*, it should make sure that such description is more expressive than bare-bones XAI outputs and at least as expressive as a Nth-Level Explanatory Closure, allowing users to identify and create their own goal-driven narratives as paths within the *explanatory space*.

One current limitation of the empirical results we gathered is that they rely on explanatory tools like YAI4Hu that do not fully implement the SAGE-ARS model, i.e. by not implementing sophisticated mechanisms for *adaptivity* (the A of SAGE) or by implementing in a naive way the *relevance* heuristic. In general, we consider it a future challenge to be able to implement user-centred explanatory tools that fully adhere to the SAGE-ARS model in a way that maximizes its usability. Besides that, the empirical results we collected are partial and they consider only a few and very specific XAI (i.e., local explanation methods) and explananda, pointing to the need for a more comprehensive and thorough evaluation of the SAGE-ARS model. This new evaluation should include not only examples of XAI but also all the other types of explanatory systems used in practice with humans. These include intelligent tutoring systems (VanLehn 2011) and, more generally, other explanatory tools for education.

Nonetheless, another possible limitation of our work is the fact that we evaluated it only on generic lay persons, without making any assumption on the background knowledge of the explainee. In fact, it is possible that YAI4Hu might be less effective

with different types of users (e.g., field experts, data scientists). That is because the *simplicity* heuristic is designed to give simpler information first, while an expert user might be more interested in having more specific and complex information first. So, as future work, smarter and more adaptive strategies to *Aspect Overviewing* might be designed to improve the user experience of an expert explainee on YAI4Hu, even if we believe that features like the *Open Question Answering* of YAI4Hu might easily overcome the issue.

Furthermore, if the SAGE-ARS model and its underlying theoretical framework is generic enough to fairly capture the complexity of pragmatically explaining in absolute terms, in principle we should be able to use it for explaining to any kind of intelligent agent (i.e., not just a human). Therefore, being able to use the SAGE-ARS model for improving machine learning as well as human learning might be the very next step to defend the importance of YAI. To this end, a recent work in the field of Reinforcement Learning (RL) applied to robotics showed that it is possible to apply our SAGE-ARS heuristics to improve the performance of several seminal off-policy RL algorithms through explanation-aware experience replay in rule-dense environments (Sovrano et al. 2022a). Hence, we leave as future work further applications of YAI in more domains where pragmatically explaining is of vital importance for intelligence.

Funding Open access funding provided by Alma Mater Studiorum - University di Bologna within the CRUI-CARE Agreement.

References

Achinstein P (1983) The nature of explanation. Oxford University Press, https://books.google.it/books?id=0XI8DwAAQBAJ

Achinstein P (2010) Evidence, explanation, and realism: essays in philosophy of science. Oxford University Press

Bachoore EH, Bodlaender HL (2007) Weighted treewidth algorithmic techniques and results. In: Tokuyama T (ed) Algorithms and computation, 18th international symposium, ISAAC 2007, Proceedings, Lecture notes in computer science, Springer, Sendai, vol 4835, pp 893–903, https://doi.org/10.1007/978-3-540-77120-3_77

Beckage B, Kauffman S, Gross LJ, Zia A, Koliba C (2013) More complex complexity: exploring the nature of computational irreducibility across physical, biological, and human social systems, Springer, Berlin Heidelberg, pp 79–88. https://doi.org/10.1007/978-3-642-35482-3_7

Berland LK, Reiser BJ (2009) Making sense of argumentation and explanation. Sci Educ 93(1):26–55

Bibal A, Lognoul M, de Streel A, Frénay B (2021) Legal requirements on explainability in machine learning. Artif Intell Law 29(2):149–169. https://doi.org/10.1007/s10506-020-09270-4

Bretto A (2013) Hypergraph theory: an introduction. Mathematical engineering, Springer International Publishing, https://books.google.co.uk/books?id=lb5DAAAAQBAJ

Brooke J (2013) Sus: a retrospective. J Usability Stud 8(2):29–40

Cacioppo JT, Petty RE (1982) The need for cognition. J Personal Soc Psychol 42(1):116–131. https://doi.org/10.1037/0022-3514.42.1.116

Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R (eds) Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, San Francisco, pp 785–794, https://doi.org/10.1145/2939672.2939785

Dazeley R, Vamplew P, Foale C, Young C, Aryal S, Cruz F (2021) Levels of explainable artificial intelligence for human-aligned conversational explanations. Artif Intell 299:103525. https://doi.org/10.1016/j.artint.2021.103525

de Holanda Coelho GL, Hanel PH, Wolf LJ (2020) The very efficient assessment of need for cognition: developing a six-item version. Assessment 27(8):1870–1885

Dhurandhar A, Chen P, Luss R, Tu C, Ting P, Shanmugam K, Das P (2018) Explanations based on the missing: towards contrastive explanations with pertinent negatives. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31: annual conference on neural information processing systems 2018, NeurIPS 2018, Montréal, Canada, pp 590–601, https://proceedings.neurips.cc/paper/2018/hash/c5ff2543b53f4cc0ad3819a36752467b-Abstract.html

FitzGerald N, Michael J, He L, Zettlemoyer L (2018) Large-scale QA-SRL parsing. In: Gurevych I, Miyao Y (eds) Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, Long Papers, Association for computational linguistics, vol 1, Melbourne, pp 2051–2060, https://doi.org/10.18653/v1/P18-1191

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter MA, Kagal L (2018) Explaining explanations: An overview of interpretability of machine learning. In: Bonchi F, Provost FJ, Eliassi-Rad T, Wang W, Cattuto C, Ghani R (eds) 5th IEEE international conference on data science and advanced analytics, DSAA 2018, IEEE, Turin, pp 80–89, https://doi.org/10.1109/DSAA.2018.00018

Gottlob G, Greco G, Leone N, Scarcello F (2016) Hypertree decompositions: questions and answers. In: Milo T, Tan W (eds) Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI symposium on principles of database systems, PODS 2016, San Francisco, ACM, pp 57–74, https://doi.org/10.1145/2902251.2902309

He L, Lewis M, Zettlemoyer L (2015) Question-answer driven semantic role labeling: using natural language to annotate natural language. In: Màrquez L, Callison-Burch C, Su J, Pighin D, Marton Y (eds) Proceedings of the 2015 conference on empirical methods in natural language processing, EMNLP 2015, The Association for Computational Linguistics, Lisbon, pp 643–653, https://doi.org/10.18653/v1/d15-1076

International Organization for Standardization (2010) Ergonomics of human-system interaction: part 210: human-centred design for interactive systems. ISO

Jansen P, Balasubramanian N, Surdeanu M, Clark P (2016) What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In: Calzolari N, Matsumoto Y, Prasad R (eds) COLING 2016, 26th international conference on computational linguistics, proceedings of the conference: technical papers, ACL, Osaka, pp 2956–2965, URL https://aclanthology.org/C16-1278/

Khosravi H, Shum SB, Chen G, Conati C, Tsai YS, Kay J, Knight S, Martinez-Maldonado R, Sadiq S, Gašević D (2022) Explainable artificial intelligence in education. Comput Educ: Artif Intell 3:100074. https://doi.org/10.1016/j.caeai.2022.100074

Liao QV, Varshney KR (2021) Human-centered explainable AI (XAI): from algorithms to user experiences. CoRR arXiv:/2110.10790

Liao QV, Gruen DM, Miller S (2020) Questioning the AI: informing design practices for explainable AI user experiences. In: Bernhaupt R, Mueller FF, Verweij D, Andres J, McGrenere J, Cockburn A, Avellino I, Goguey A, Bjøn P, Zhao S, Samson BP, Kocielnik R (eds) CHI '20: CHI conference on human factors in computing systems, ACM, Honolulu, pp 1–15, https://doi.org/10.1145/3313831.3376590

Lim BY, Dey AK, Avrahami D (2009) Why and why not explanations improve the intelligibility of context-aware intelligent systems. In: Jr DRO, Arthur RB, Hinckley K, Morris MR, Hudson SE, Greenberg S (eds) Proceedings of the 27th international conference on human factors in computing systems, CHI 2009, ACM, Boston, pp 2119–2128, https://doi.org/10.1145/1518701.1519023

Lundberg SM, Lee S (2017) A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information process-

ing systems 2017, Long Beach, pp 4765–4774, https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
Lundberg SM, Erion GG, Chen H, DeGrave AJ, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9
Madumal P, Miller T, Sonenberg L, Vetere F (2019) A grounded interaction protocol for explainable artificial intelligence. In: Elkind E, Veloso M, Agmon N, Taylor ME (eds) Proceedings of the 18th international conference on autonomous agents and multiagent systems, AAMAS '19, International foundation for autonomous agents and multiagent systems, Montreal, pp 1033–1041, http://dl.acm.org/citation.cfm?id=3331801
Martin R (2002) Agile software development: principles, patterns, and practices. Prentice Hall
Michael J, Stanovsky G, He L, Dagan I, Zettlemoyer L (2018) Crowdsourcing question-answer meaning representations. In: Walker MA, Ji H, Stent A (eds) Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, NAACL-HLT, Association for Computational Linguistics, vol 2 (Short Papers), New Orleans, pp 560–568, https://doi.org/10.18653/v1/n18-2089
Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007
Pyatkin V, Klein A, Tsarfaty R, Dagan I (2020) Qadiscourse: discourse relations as QA pairs: representation, crowdsourcing and baselines. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, Online, Association for computational linguistics, pp 2804–2819, https://doi.org/10.18653/v1/2020.emnlp-main.224
Rebanal JC, Combitsis J, Tang Y, Chen XA (2021) Xalgo: a design probe of explaining algorithms' internal states via question-answering. In: Hammond T, Verbert K, Parra D, Knijnenburg BP, O'Donovan J, Teale P (eds) IUI '21: 26th international conference on intelligent user interfaces, ACM, College Station, pp 329–339, https://doi.org/10.1145/3397481.3450676
Ribeiro MT, Singh S, Guestrin C (2016) "why should I trust you?": Explaining the predictions of any classifier. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R (eds) Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, San Francisco, pp 1135–1144, https://doi.org/10.1145/2939672.2939778
Ribera M, Lapedriza À (2019) Can we do better explanations? A proposal of user-centered explainable AI. In: Trattner C, Parra D, Riche N (eds) Joint proceedings of the ACM IUI 2019 workshops co-located with the 24th ACM conference on intelligent user interfaces (ACM IUI 2019), Los Angeles CEUR-WS.org, CEUR workshop proceedings, vol 2327, URL http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf
Sovrano F, Vitali F (2021a) From philosophy to interfaces: an explanatory method and a tool inspired by achinstein's theory of explanation. In: Hammond T, Verbert K, Parra D, Knijnenburg BP, O'Donovan J, Teale P (eds) IUI '21: 26th international conference on intelligent user interfaces, ACM, College Station, pp 81–91, https://doi.org/10.1145/3397481.3450655
Sovrano F, Vitali F (2021b) An objective metric for explainable AI: how and why to estimate the degree of explainability. CoRR arXiv:/2109.05327
Sovrano F, Vitali F (2022) Generating user-centred explanations via illocutionary question answering: from philosophy to interfaces. ACM Trans Interact Intell Syst. https://doi.org/10.1145/3519265
Sovrano F, Vitali F (2022b) How to quantify the degree of explainability: experiments and practical implications. In: 31th IEEE international conference on fuzzy systems, FUZZ-IEEE 2022, IEEE, Padova, pp 1–9
Sovrano F, Palmirani M, Vitali F (2020a) Legal knowledge extraction for knowledge graph based question-answering. In: Villata S, Harasta J, Kremen P (eds) Legal knowledge and information systems: JURIX 2020—the thirty-third annual conference, Frontiers in artificial intelligence and applications, Brno, IOS Press, vol 334, pp 143–153, https://doi.org/10.3233/FAIA200858
Sovrano F, Vitali F, Palmirani M (2020b) Modelling gdpr-compliant explanations for trustworthy AI. In: Ko A, Francesconi E, Kotsis G, Tjoa AM, Khalil I (eds) Electronic government and the information systems perspective: 9th international conference, EGOVIS 2020, Proceedings, Lecture notes in computer science, vol 12394, Springer, Bratislava, pp 219–233, https://doi.org/10.1007/978-3-030-58957-8_16
Sovrano F, Sapienza S, Palmirani M, Vitali F (2021) A survey on methods and metrics for the assessment of explainability under the proposed AI act. In: Erich S (ed) Legal knowledge and information systems:

JURIX 2021—the thirty-fourth annual conference, Frontiers in artificial intelligence and applications, vol 346, IOS Press, Vilnius, pp 235–242, https://doi.org/10.3233/FAIA210342
Sovrano F, Raymond A, Prorok A (2022) Explanation-aware experience replay in rule-dense environments. IEEE Robot Autom Lett 7(2):898–905. https://doi.org/10.1109/LRA.2021.3135927
Sovrano F, Sapienza S, Palmirani M, Vitali F (2022) Metrics, explainability and the European ai act proposal. J 5(1):126–138. https://doi.org/10.3390/j5010010
VanLehn K (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ Psychol 46(4):197–221. https://doi.org/10.1080/00461520.2011.611369
Vilone G, Longo L (2022) A novel human-centred evaluation approach and an argument-based method for explainable artificial intelligence. In: Maglogiannis I, Iliadis L, Macintyre J, Cortez P (eds) Artificial intelligence applications and innovations—18th IFIP WG 12.5 international conference, AIAI 2022, Hersonissos, Proceedings, part I, IFIP Advances in information and communication technology, vol 646, Springer, pp 447–460, https://doi.org/10.1007/978-3-031-08333-4_36
Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv J Law Technol. https://doi.org/10.2139/ssrn.3063289
Zwirn H, Delahaye JP (2013) Unpredictability and computational irreducibility. Springer, Berlin, Heidelberg, pp 273–295. https://doi.org/10.1007/978-3-642-35482-3_19