An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability

Francesco Sovrano and Fabio Vitali

Abstract

Numerous government initiatives (e.g. the EU with GDPR) are coming to the conclusion that the increasing complexity of modern software systems must be contrasted with some *Rights to Explanation* and metrics for the Impact Assessment of these tools, that allow humans to understand and oversee the output of Automated Decision Making systems. Explainable AI was born as a pathway to allow humans to explore and understand the inner working of complex systems. But establishing what *is* an explanation and *objectively* evaluating *explainability*, are not trivial tasks. With this paper, we present a new model-agnostic metric to measure the Degree of eXplainability of (correct) information in an *objective* way, exploiting a specific theoretical model from Ordinary Language Philosophy called the *Achinstein's Theory of Explanations*, implemented with an algorithm relying on deep language models for knowledge graph extraction and information retrieval. In order to understand whether this metric is actually behaving as *explainability* is expected to, we have devised a few experiments and user-studies involving more than 160 participants evaluating two realistic AI-based systems for *healthcare* and *finance* using famous AI technology including Artificial Neural Networks and TreeSHAP. The results we obtained are very encouraging, suggesting that our proposed metric for measuring the Degree of eXplainability is robust on several scenarios and it can be eventually exploited for a lawful Impact Assessment of an Automated Decision Making system.

Keywords: Objective Explainability Metric, XAI, Automated Impact Assessment, Degree of Explainability, Knowledge Graph based Question Answering

Preprint submitted to Journal of LaTeX Templates *December 9, 2021*

1. Introduction

Recent advances in Artificial Intelligence (AI) are enabling computer science and engineering to create machines that can learn from rough data, hence automating tasks previously thought to be accessible only by biological intelligence, such as autonomous control (i.e. of vehicles), sound, image or natural language processing. But these advances and results have a cost paid in terms of explainability, in fact it seems that the internal logic of the AI most effective in learning is, so far, not easily interpretable in symbolic terms [1, 2]. The paradigms that address this problem fall into the so-called eXplainable AI (XAI) field, which is broadly recognized as a crucial feature for the practical implementation of artificial intelligence models [3]. In fact, Automated Decision-Making systems (ADMs) are changing our society, so people and governments (e.g., EU's, California, etc.) have begun to be concerned about the impact that they may have on our lives. This concern gave birth to the so-called *Right to Explanation* [4], which was introduced in the EU legislation within the General Data Protection Regulation (GDPR), and further explored by the High-Level Expert Group on Artificial Intelligence (AI-HLEG) [5], established in 2018 by the EU Commission.

As a result, the EU indirectly posed an interesting challenge to the eXplainable AI (XAI) community, by demanding more transparent, user-centred, and accountable approaches to ADM that guarantee explainability of their working. More precisely, the GDPR art. 35 requires data controllers to prepare a Data Protection Impact Assessment (DPIA) for operations that are "likely to result in a high risk to the rights and freedoms of natural persons". To this end, Algorithmic Impact Assessment (AIA) can be intended as an instrument for ensuring certain minimal criteria [6] of explainability in ADMs, serving as an important "suitable safeguard" (Article 22) of individual rights.

This is certainly one of the reasons why we may be interested in any metric for automatically measuring the degree of explainability of information. In fact, controllers who use machine learning systems for processing of personal data should be able to argue in cases when Data Subjects or Data Protection Officers (DPOs)[1] quarrel that the logic of processing is explained way too vaguely, that they did what they could, providing an acceptable level of objective explainability of the respective algorithms.

In this paper, we propose a new model-agnostic approach and metric to *objectively* evaluate explainability, through knowledge graph extraction, in a manner that is mainly inspired by Ordinary Language Philosophy instead of Cognitive Science. Our approach is based on a specific theoretical model of explanation, called the *Achinstein's theory of explanations*, where explanations are the result of an *illocutionary* (i.e., broad yet pertinent and deliberate) act of pragmatically answering to a question. Accordingly, explanations are actually answers to many different basic questions (*archetypes*) each of which sheds a different light over the concepts being explained. As consequence, the more (archetypal) answers an ADM is able to give about the important aspects of its explanandum, the more it is explainable.

Therefore, we assert that it is possible to quantify the degree of explainability of a set of texts by applying the Achinstein-based definition of explanation proposed in [7]. Thus, drawing also from Carnap's criteria of adequacy of an explication [8], we frame the Degree of eXplainability (DoX) as the average *Explanatory Illocution* of information on the *Explanandum Aspects*[2]. More precisely, we hereby present an algorithm for measuring DoX by means of pre-trained *language models* for general-purpose question answer retrieval, as [9, 10], applied to a special knowledge graph of triplets automatically extracted from text to facilitate this type of information retrieval. Hence, we made the following hypothesis.

[1] See articles 37-39 of the GDPR for more details on what a DPO is.

[2] Carnap uses the term *explicandum* where we employ *explanandum*, but, by and large, we assume the two words can be used interchangeably. They both mean "what has to be explained" in Latin.

Hypothesis 1. *DoX measures explainability: the DoX can describe explainability, so that given the same explanandum, an higher DoX implies greater explainability and a lower DoX implies less explainability.*

To verify this hypothesis, we performed two experiments, both with the objective of showing that *explainability* changes in accordance with DoX. To conduct the experiments we considered two different XAI-based systems, respectively for the healthcare and finance domains:

- A *Heart Disease Predictor* based on XGBoost [11] and TreeSHAP [12].
- A *Credit Approval System* based on a simple Artificial Neural Network and on CEM [13].

The first experiment follows a *direct* approach, comparing the DoX of the XAI-based systems with their non-explainable counterpart. This approach is said to be *direct*, because the amount of *explainability* of a XAI-based system is, by design, clearly and explicitly dependent on the output of the underlying XAI. Therefore, by filtering away the XAI's output, the overall system can be forced to be not explainable enough, by construction.

On the other hand, the second experiment follows an *indirect* approach, analysing the expected effects of *explainability* on the explainees. In fact, if the hypothesis is correct, the lower is DoX, the less explanations can be extracted, the less effective (as per ISO 9241-210) is likely to be an explainee in achieving those explanatory goals that are not covered by the explanations. Hence, once fixed all the components that may affect effectiveness, including the presentation logic (the mechanism for re-elaborating explainable information into explanations) and the explanandum, an increase in DoX should always correspond to a proportional increase of effectiveness, at least in those tasks covered by the information provided by the increment of DoX. To show this, we borrowed the results of two independent user-studies, then measuring how DoX correlates with the effectiveness scores measured by the user-studies.

The first user-study [14] consisted of more than 90 participants, collected on the online platform Prolific[3]. These participants were asked to evaluate (through a quiz) the explanations generated by different versions of the Credit Approval System. While the second study [7] is an extension of the first one and consisted of more than 60 new respondents, that volunteered to a “call for participants” forwarded to the university students that subscribed to an interdisciplinary computer science class of the same university. Differently from the first, the respondents to the second user-study evaluated not just the Credit Approval System, but also the Heart Disease Predictor.

In all the experiments and user-studies the results were clear, showing that hypothesis 1 holds. Therefore, we believe that our technology for estimating the DoX might be used for an objective and lawful Algorithmic Impact Assessment (AIA), as soon as what is needed to be explained can be identified under the requirements of the law, in the form of a set of precise *Explanandum Aspects*. For guaranteeing the reproducibility of the experiments, we published[4] the source code of DoX, as well as the code of the XAI-based systems, the user-study questionnaires and the remaining data mentioned within this paper.

This paper is structured as follows. In Section 2 we study existing literature, comparing it to our proposed solution, and in Section 3 we give the necessary background information to properly introduce the theoretical models discussed in the remaining sections, including Achinstein’s theory of explanations. In Section 4.1 we show how it is possible to quantify the degree of explainability by defining *explaining* as an illocutionary act of question answering and by verifying Carnap’s criteria by means of deep language models applied on a knowledge graph of SVO triplets. Finally, in Section 5 we define our experiments, presenting the findings in Section 6 and discussing them in Section 7.

[3] `https://www.prolific.co`

[4] `https://github.com/Francesco-Sovrano/DoXpy`

2. Related Work

Being able to measure the quality of XAI tools is pivotal for claiming technological advancements, understanding existing limitations, developing better solutions and delivering XAI that can go into production. Not surprisingly, every good paper proposing a new XAI algorithm comes with some evidence or experiments to back up the underlying claims, usually relying on *ad hoc* or subjective mechanisms to measure the quality of explainability. In other words, it is very common to encounter explainability metrics that can work only with specific XAI models or that require to collect opinions/results generated by human subjects interacting with the system . For example, the metrics proposed by [15, 16, 17, 18] can only be used with specific types of XAI (i.e. prototype selection, feature attribution, etc.). While the metrics proposed by [19, 20] rely on usability tests and user-studies.

Interestingly, only [19] claim that their work is generic enough to be used to evaluate any XAI, proposing to measure explainability indirectly, by estimating the effects that the resulting explanations have on the subjects. More precisely, [19] 's metric is mainly inspired by the interpretation of explanations given by Cognitive Science, requiring to measure:

- the subjective goodness of explanations,
- whether users are satisfied by explanations,
- how well users understand the AI systems,
- how curiosity motivates the search for explanations,
- whether the user's trust and reliance on the AI are appropriate,
- how the human-XAI work system performs.

In other terms, the metric presented by [19] is heavily relying on subjective measurements. Differently, our DoX is a fully objective metric that can be used to evaluate the explainability of any textual information and to understand whether the amount of explainability is objectively poor, even if the resulting explanations are perceived as satisfactory and good by the explainees.

3. Background

In this section we provide enough background to understand and support the rest of the paper. Hereby we briefly summarise a number of recent and not-so-recent approaches to the theories of explanation, with a particular due focus on Achinstein's. After that, 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.

3.1. Definitions of Explainability

Considering the definition of "explainability" as "the potential of information to be used for explaining", we envisage that a proper understanding of how to measure explainability must pass through a thorough definition of what constitutes an explanation and the act of explaining. In 1948 Hempel and Oppenheim published their "Studies in the Logic of Explanation" [**?**], giving birth to what it is considered the first theory of explanation, the deductive-nomological model. After that date, many attempts followed to amend, extend or replace this first model, which was considered fatally flawed [**?** 21]. This gave birth to several competing and more contemporary theories of explanations [26]: i) Causal Realism, ii) Constructive Empiricism, iii) Ordinary Language Philosophy, iv) Cognitive Science, v) Naturalism and Scientific Realism. A summary of these definitions is shown in Table 1.

Interestingly, each one of these theories devises different definitions of "explanation". If we look at their specific characteristics we may find that all but *Causal Realism* are pragmatic. On the other hand, *Causal Realism* and *Constructive Empiricism* are rooted on causality, while the others not [5]. Nonethe-

[5] They study the act of explaining as an iterative process involving broader forms of question

Table 1: **Definitions of *explanation* and *explainable information*** for each theory of explanations.

Theory	**Def. of Explanation**	**Def. of Explainable Information**
Causal Realism [21]	It is a description of causality, as chains of causes and effects.	It can fully describe causality.
Constructive Empiricism [22]	It is contrastive information answering WHY questions, allowing one to calculate the probability of a particular event relative to a set of (possibly subjective) background assumptions.	It provides answers to contrastive WHY questions.
Ordinary Language Philosophy [23]	Explaining is pragmatically answering to (not just WHY) questions, with the explicit intent of producing understanding.	It can be used to pertinently answer questions about relevant aspects, in an illocutionary way.
Cognitive Science [24]	Explaining is a process triggered as response to predictive failures and it is about providing information to fix that failures in a mental model (sometimes intended as a hierarchy of rules).	It can fix failures in mental models.
Naturalism and Scientific Realism [25]	Explaining is an iterative process of confirmation of truth based on inference to the best explanation. An explanation increases understanding, not simply by being the correct answer to a particular question, but by increasing the coherence of an entire belief system (e.g. a subject).	It can be used to increase understanding, i.e. by answering to particular questions.

less, *Cognitive Science* and *Scientific Realism* are more focused on the effects that an explanation has on the explainee (the recipient of the explanation).

Importantly, with the present letter, we assert that whenever explaining is considered to be a pragmatic act, explainability differs from explaining. In fact, pragmatism in this sense is achieved when the explanation is tailored to the specific user, so that the same explainable information can be presented and re-elaborated differently across users. It follows that for each philosophical tradition, but Causal Realism, we have a definition of "explainable information" that slightly differs from that of "explanation", as shown in Table 1.

answering

3.2. Carnap's Criteria of Adequacy

In philosophy, the most important work about the central criteria of adequacy of *explainable information* is likely to be Carnap's [27]. Even though Carnap studies the concept of *explication* rather than that of *explainable information*, we assert that they share a common ground making his criteria fitting in both cases. In fact, *explication* in Carnap's sense is the replacement of a somewhat unclear and inexact concept (the explicandum) by a new, clearer, and more exact concept called explicatum, and that is exactly what information does when made explainable.

Carnap's central criteria of explication adequacy are [27]: *similarity*, *exactness* and *fruitfulness*[6]. *Similarity* means that the explicatum should be similar to the explicandum, in the sense that at least many of its intended uses, brought out in the clarification step, are preserved in the explicatum. On the other hand, *Exactness* means that the explication should, where possible, be embedded in some sufficiently clear and exact linguistic framework. While *Fruitfulness* means that the explicatum should be used in a high number of other *good* explanations (the more, the better).

Carnap's adequacy criteria seem to be transversal to all the identified definitions of explainability, possessing preliminary characteristics for any piece of information to be considered properly explainable. Interestingly, the property of *truthfulness* (being different from *exactness*) is not explicitly mentioned in Carnap's desiderata. That is to say that explainability and *truthfulness* are complementary, but different, as discussed also by [28]. In fact an explanation is such regardless its truth (wrong but high-quality explanations exist, especially in science). Vice-versa, highly correct information can be very poorly explainable.

[6] Carnap also discussed another desideratum, *simplicity*, but this criterion is presented as being subordinate to the others.

3.3. Achinstein's Theory of Explanations

In 1983, [23] 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 coming from a clear intention of producing new understandings in an explainee by providing a correct content-giving answer to an open question. Therefore, according to this view, answering by "filling the blank" of a pre-defined template answer (as most of One-Size-Fits-All approaches do) prevents the act of answering from being explanatory, by lacking illocution. These conclusions are quite clear and explicit in Achinstein's last works [29], consolidated after a few decades of public debates.

More precisely, according to Achinstein's theory, an explanation can be summarized as a pragmatically correct content-giving answer to questions of various kinds, not necessarily linked to causality. 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.

As consequence we can see a deliberate absence of a taxonomy of questions (helpful to categorize and better understand the nature of human explanations) to refer. This apparently results in a refusal to define a quantitative way to measure how pertinent an answer is to a question, justified by the important assertion that explanations have a pragmatic character, so that what exactly has to be done to make something understandable to someone may (in the most generic case) depend on the interests and background knowledge of the person seeking understanding [30].

In this sense, the strong connection of Achinstein's theory to natural language and natural users is quite evident, for example in the Achinsteinian concept of *elliptical understandings* as "understandings of what significance or importance X has in the present context" [29] or in the concept of *u-restrictions* where an utterance/explanation can be said to express a proposition if and only if it can appear in (many) contexts reasonably known by the explainee. But, despite this, Achinstein does not reject at all the utility of formalisms, hence suggesting the importance of following *instructions* (protocols, rules, algorithms) for correctly explaining some specific things within specific contexts.

3.4. Archetypal Questions

According to Achinstein's theory, explanations are the result of an *illocutionary* act of pragmatically answering to a question. In particular, it means that there is a subtle and important difference between simply "answering to questions" and "explaining", and this difference is *illocution.* It appears that an illocutionary act results from a clear intent of achieving the goal of such act, as a promise being "what it is" just because of the intent of maintaining it. So that illocution in explaining makes an explanation as such just because it is the result of an underlying and proper intent of explaining.

Notwithstanding this definition, *illocution* seems to be too abstract to be implementable into a concrete software. Nonetheless, recent efforts towards the automated generation of explanations [14, 7], have shown that it may be possible to define *illocution* in a more "computer-friendly" way. As stated by [14], illocution in explaining involves informed and *pertinent* answers not just to the main question, but also to other questions of various kinds, even unrelated to causality, that are relevant to the explanations. Such questions can be understood as instances of archetypes such as why, why not, how, what for, what if, what, who, when, where, how much, etc..

Definition 1 (Archetypal Question). *An archetypal question is an archetype applied on a specific aspect of the explanandum. 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". In other terms,* *archetypal questions identify generic explanations about a specific aspect to explain (e.g. a topic, an argument, a concept,etc.), in a given informative context.*

In other words, archetypal questions provide generic explanations on a specific aspect of the explanandum, in a given informative context, with a local or a global slant (i.e. linked or not to the specific computation as performed), which can precisely link the content to the informative goal of the person asking the question. For example, if the explanandum were "heart diseases", there would be many aspects involved including "heart", "stroke", "vessels", "diseases", "angina", "symptoms", etc. Some archetypal questions in this case might be "What is an angina?" or "Why a stroke?".

An answer to an archetypal question is said to be an *archetypal explanation.* Being an archetypal question a generic question requiring a generic answer, an archetypal explanation summarises the information given as answer to other punctual (non-archetypal) questions posed by (possibly) many different explainees, also in different moments in time.

3.5. XAI and Question Answering

If we assume that the interpretation of Achinstein's theory of explanations given by [14] is correct, then data or processes are said to be *explainable* when their informative content can adequately answer *archetypal questions.* The idea of answering questions as explaining is not new to the field of XAI [31] and it is also quite compatible with everyone's intuition of what constitutes an explanation. In fact, it is common to many works in the field [32, 33, 34, 35, 13, 36, 37, 38, 39] the use of generic (e.g. why, who, how, when, etc.) or more punctual questions to clearly define and describe the characteristics on explainability [31].

For example, [12] 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 [13] clearly state that they designed CEM (a method for the generation of counterfactuals and other contrastive explanations) to answer the question “why is input x classified in class y?”.

Also, [37] propose and studies an interactive approach where explaining is defined in terms of answering why-what-how questions. 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 [33], [34] and [35] that analysing XAI literature were able to hypothesise that a good explanation, about an automated decision-maker, answers at least the following questions:

- What did the system do?,
- Why did the system do P?,
- Why did the system not do X?,
- What would the system do if Y happens? ,
- How can I get the system to do Z, given the current context?
- What information does the system contain?

Nonetheless, despite its compatibility, practically none of the works in XAI ever explicitly mentioned Ordinary Language Philosophy’s theories, preferring to refer Cognitive Science’s [34, 19] instead. This is probably because Achinstein’s illocutionary theory of explanations is seemingly difficult to be implemented into a software, by being utterly pragmatic.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. Proposed Solution

In Section 2 we discussed how existing metrics for measuring (properties of) explainability are frequently either model-specific or subjective, raising the question of whether it is possible to objectively measure the degree of explainability with a fully automated software. With this paper we try to answer this question, by leveraging on an extension of Achinstein's theory of explanations proposed by [14] and presented in Sections 3.3 and 3.4. We do it by asserting that any algorithm for measuring the degree of explainability must pass through a thorough definition of what constitutes *explainability* and thus also an *explanation*. In fact, considering that *explainability* is fundamentally the *ability to explain*, it is clear that a proper definition of it requires a precise understanding of what is *explaining*. So, in this Section we discuss both the new theory behind our proposed solution for computing the Degree of eXplainability (DoX) and a concrete implementation we devised to measure the DoX in practice.

4.1. Theory: How to Quantify the Degree of Explainability

As discussed in Section 3.5, the informative contents of state-of-the-art XAI is clearly polarised towards answering "why", "what-if" or "how" questions. Considering that "why", "what-if" and "how" are different questions pointing to different types of information, which type is the best one? We assert that the correct answer to this question is: "none". In fact, depending on the needs of the explainee, its background knowledge, the context, and potentially many other factors, each one of these explanation archetypes may be equally needed, together with all the others we left out, including: "what", "who", "when", etc.. In other words, depending on the characteristics of the explainee (e.g. background knowledge, objectives, etc..), a combination of different XAI mechanisms may be required to really give meaningful and trustable insights on the inner logic of a black-box. Therefore, knowing the types of explainability that are covered by a XAI-based system can be of utmost importance for understanding how explainable is a piece of information. So, following this intuition, we started to study how to measure explainability in terms of (generic) questions.

Among the different theories mentioned in Section 3.1, the closest one to our intuition of explainability, as the ability to answer questions, is probably Achinstein's. In fact, as discussed in Section 3.3, Achinstein defines the act of explaining as an act of illocutionary question answering, stating that it is more than answering to a question, requiring some form of illocution. Nonetheless, without a precise and computer-friendly definition of illocution it is hard to go further than a philosophical and abstract understanding of an explanation. For this reason, as discussed in Section 3.4, [14] have suggested that illocution (or explanatory illocution) is indeed the process of answering multiple generic and primitive questions (i.e. Why? How? What? etc.) called *archetypal questions*.

We assert that, given these premises, we have everything we need to go further and concretely measure the degree of explainability of information in a quantitative way. More precisely, we propose that the degree of explainability of information depends on the amount of *archetypal questions* to which it can answer properly. In other words, we hypothesise that it is possible to estimate the degree of explainability of a piece of information by measuring the relevance with which it can answer a (pre-defined) set of archetypal questions. Indeed, although it may seem that numerically calculating the relevance of an answer to a question is a daunting task, recent developments in modern artificial intelligence have shown that it is possible to create tools capable of that, i.e. [40, 41, 42, 43, 9]. Hence, our theoretical contribution consists in a precise, formal, definition of:

- Explanandum Aspects Coverage
- Explanatory Illocution
- DoX,
- Averaged and Weighted DoX.

4.1.1. Explanatory Illocution and Degree of Explainability

Assuming that the content of a given piece of information is correct, *explainability* is a property that information possesses and it can be measured in terms of *Explanatory Illocution*. We formally define it as follows:

Definition 2 (Explanandum Aspects Coverage). *Let I be the set of aspects contained in that piece of information and A the set of relevant aspects to be explained about an explanandum, then the Explanandum Aspects Coverage is the set $A \cap I$ of explanandum aspects that are covered by that information, while the inverse-coverage is the set $A - I$ of uncovered aspects.*

Definition 3 (Explanatory Illocution). *The Explanatory Illocution is an estimate of how pertinently and how in detail a given piece of information can answer to a set of pre-defined archetypal questions on an explanandum aspect. Let D be the set $\{\forall a \in A | D_a\}$ and D_a be the set of all the details contained in that information about an aspect $a \in A$, let Q be the set of all possible archetypes q, let q_a be the archetypal questions obtained by applying the archetype q to an aspect $a \in A$, and let $p\left(d, q_a\right) \in [0, 1]$ be the pertinence of a detail $d \in D_a$ to q_a. Let also t be a pertinence threshold in $[0, 1]$, and let $P_{D_a,q_a} = \sum_{d \in D_a, p(d,q_a) \geq t} p\left(d, q_a\right)$ be the cumulative pertinence of D_a to q_a, then the Explanatory Illocution for a is the set $\{\forall q \in Q | \left\langle q, P_{D_a,q_a} \right\rangle\}$.*

Consequently we have that:

Definition 4 (Degree of eXplainability). *The Degree of eXplainability (DoX) is the average Explanatory Illocution per archetype, on the whole set A of relevant aspects to be explained. In other terms, let $R_{D,q,A} = \frac{\sum_{a \in A} P_{D_a,q_a}}{|A|}$ be the average cumulative pertinence of D to q and A, then the DoX is the set $\{\forall q \in Q | \left\langle q, R_{D,q,A} \right\rangle\}$.*

Importantly, the DoX, as we defined it, is akin to Carnap's *central* criteria of adequacy of explanation (introduced in Section 3.2). Although, differently from Carnap, our understanding of *exactness* is not that of adherence to standards of formal concept formation[7] [44], but rather that of being precise or pertinent enough as an answer to a given question. In other terms, the DoX is an estimate

[7] Actually, Carnap did not specify what he means by "exactness", regardless that is often viewed as either lack of vagueness or adherence to standards of formal concept formation.

of the *fruitfulness* of D that combines in one single score the *similarity* of D to A and the *exactness* of D with respect to Q.

In fact, the number of *relevant aspects* $(A \cap I)$ covered by a given piece of information, and the *amount of details* (D) that are pertinent about it, roughly say how much *similar* that information is to the explanandum. More precisely, the formula used for computing P_{D_a,q_a} sums the contribution of each single detail according to its pertinence to the aspects $a \in A$, telling us how much D_a is similar to a, so that if $p\left(d, q_a\right)$ is close to zero for all $q \in Q$, then a detail d has nothing to do with an aspect a. Furthermore, $R_{D,q,A}$ contains information about the *exactness* of multiple archetypal explanations, being an aggregation of pertinence scores. As result, by measuring $R_{D,q,A}$ for all the $q \in Q$ we obtain also an estimate of how D is *fruitful* for the formulation of many other different explanations intended as the result of an illocutionary act of answering questions.

We believe that this construction of DoX in terms of Carnap's central criteria of adequacy is probably the most important theoretical contribution we do here, because it allows us to move to the next step: implementing an algorithm to experimentally verify hypothesis 1.

4.1.2. Averaged Degree of Explainability

Despite all the good properties DoX has, it cannot by itself help to judge whether one collection of information has a higher degree of explainability than another, because it is a multidimensional estimate of different archetypes. This characteristic makes it harder to tell if one DoX is greater than another. To overcome this issue, a mechanism is required for combining the pertinence of the DoX into a single score representing *explainability*. Hence, we propose to summarise the DoX by simply averaging its pertinence scores. Hence, the resulting Averaged DoX can act as a metric to judge whether the *explainability* of a system is greater than, equal to, or lower than another.

Definition 5 (Averaged Degree of eXplainability). *The Averaged DoX is the average of the pertinences of each archetype composing the DoX. In other* *terms, the Averaged DoX is* $\frac{\sum_{q \in Q} R_{D,q,A}}{|Q|}$.

The Averaged DoX represents a naive approach to quantify explainability with one single score, because it implies that all the archetypal questions and aspects have the same weight, although this may not be always true. In fact, as suggested by Section 3.5 and [31], it appears that in literature there is a shared understanding of "why" explanations, as the most important in XAI, sometimes followed by "how", "what for", "what if" or, more generally, "what". For example, according to [45], local "why" explanations are more effective in exposing fairness discrepancies between different cases, while global "how" explanations seem to render more confidence in understanding the model and generally enhance the fairness perception, reducing communication costs [46]. In other words, the relevance of an explanation can be estimated by the ability to effectively answer the most relevant (archetypal) questions for the objectives of the stakeholders.

Therefore, an alternative to the Averaged DoX might be the Weighted DoX, that is a weighted combination of the pertinence scores of a DoX, where the weights are pre-defined for each archetypal question and aspect, depending on the main goals of the system. Therefore, the resulting Weighted DoX is said to be goal-dependent and, similarly to the Averaged DoX, it can act as a metric to judge whether the *explainability* of a system is greater than, equal to, or lower than another.

Definition 6 (Weighted Degree of eXplainability). *The Weighted DoX is a weighted sum of the pertinences of each archetype composing a DoX. The archetype weights* W *used for weighting the DoX are a set of real numbers in* $[0, 1]$*, one per archetype. In other terms, let* W_q *be the weight of archetype* q*, then the Weighted DoX is:* $\sum_{q \in Q} W_q \cdot R_{D,q,A}$.

4.2. Practice: An Algorithm for Computing the Degree of Explainability

Given definition 4, we argue that it is possible to write an algorithm that can approximate a quantification of the Degree of eXplainability (DoX) of information representable with a *natural language*, i.e. English. Let's suppose we want to measure the DoX of a set of texts called *explanandum support material*, containing correct textual information (in English) about a given explanandum. For example, if the *explanandum* were “heart diseases”, there would be many aspects involved including “heart”, “stroke”, “vessel”, “diseases”, “angina”, “symptoms”, etc. Hence a reasonable *support material* for it would probably be a book describing all these aspects and more (if deemed relevant by the author), or a set of web-pages (i.e. those published by the *U.S. Centers for Disease Control and Prevention*[8]), or any other kind of corpus written in natural language.

In order to implement an algorithm capable of computing the (averaged) DoX as we defined it in Definition 4, we need to identify:

- the set A of explanandum aspects, as in Definition 2,
- the set of all possible archetypes Q and the set D of details contained in the support material, as in Definition 3,
- a mechanism to identify D for each $a \in A$,
- the function p to compute the pertinence of a detail d to an archetypal question q_a about an aspect a, as in Definition 3,

While the set of aspects A is task-dependent and needs to be defined for every explanandum (i.e. by manually listing all the aspects, or by automatically extracting with a tokenizer the list of aspects from a textual description of the explanandum), we believe that the set of archetypes Q, the pertinence function p and the mechanism for extracting D and D_a (out of the support material) can be always the same for all the explananda.

Indeed, by leveraging on existing pre-trained deep language models, i.e. [47, 40], capable of converting snippets of text (e.g. questions and answers) into numerical representations, in the following sections we show how to concretely implement an algorithm capable of estimating the DoX score of any arbitrary piece of textual information with the pipeline shown in Figure 1. More precisely,

[8] `https://www.cdc.gov`

this pipeline relies on a mechanism for the automatic extraction of a knowledge graph from the explanandum support material in order to identify the details needed by the DoX estimator to generate a score according to Definition 4.

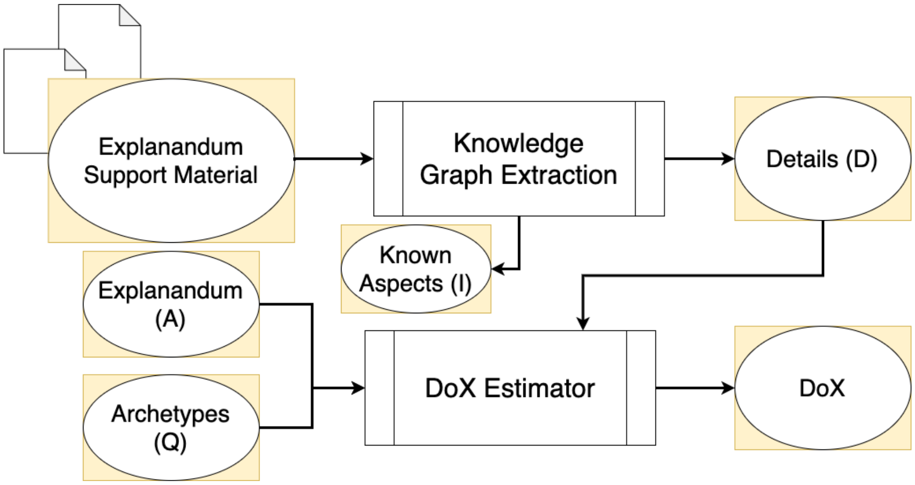


Figure 1: **The DoX Pipeline**: The pipeline starts with the extraction of a knowledge graph from the *explanandum support material* that is then converted into a set of details D. The set of details is then used in combination with the explanandum A and the set of archetypes Q to compute the DoX. To do so, we use some deep language models for Question-Answer retrieval.

4.2.1. Details Extraction and Pertinence Estimation

First of all, Definition 4 requires a mechanism to identify the set D of details contained in the support material, as well as a mechanism to identify the subsets $D_a \subseteq D$ for every $a \in A$. The set A of explanandum aspects is a collection of lemmatised words/syntagms to which it would be easy to associate a Uniform Resource Identifier (URI). On the other hand, given Definition 4, a detail d is a snippet of text with some specific characteristics. In fact, a detail is what we call *information unit*: a relatively small sequence of words about one or more aspects (i.e. a sub-set of A) that is usually extracted from a more complex information bundle (i.e. a paragraph, a sentence, etc.) comprising several information units. In other terms, these details should carry enough information to describe different parts of an aspect a (possibly connected to many other aspects), so that we can use them to answer some (archetypal) questions about a and to correctly estimate a *level of detail*, as required by Definition 4.

Considering the aforementioned characteristics of D and A, we believe that the most natural representation of them might be a knowledge graph. Indeed, knowledge graphs are sets of triplets connecting two different nodes (i.e. a subset of aspects in A) with some kind of relation or edge (i.e. a detail $d \in D$), hence any such (knowledge) graph representation of D and A would automatically give us a mechanism to identify a $D_a \subseteq D$ for every $a \in A$. Therefore, we believe that the easiest way to identify the set of details D (and possibly also A) might pass through some mechanism for the extraction of a (knowledge) graph of *information units* from the explanandum support material. Thus, an approach like the one used by [14, 43], for archetypal question answering, might be suitable to our ends, allowing for the identification of meaningful *information units* and (importantly) suggesting also a mechanism for the estimation of *pertinence*. In fact, the algorithm proposed by [43] relies on an automated mechanism that is capable of decomposing sentences into dependency trees, converting them into a special knowledge graph of Subject-Template-Object triplets (or template-triplets in short) specifically designed to facilitate (archetypal) question answering through state-of-the-art algorithms for Question Answer Retrieval [42, 43, 9]. Specifically, these algorithms can estimating the pertinence p of a detail $d \in D$ to a question q by generating a numerical embedding of both the question and the answer, so that the inner product (or other similarity metrics, i.e. the cosine similarity) between these embeddings is a measure of the pertinence of the latter to the first.

Therefore, template-triplets are used instead of normal Subject-Verb-Object triplets (or verb-triplets in short) to cope with the limitations of modern Question-Answer Retrieval. In fact, existing algorithms for Question Answer Retrieval [42, 9]: have, usually, constraints on the size of their inputs and outputs; are trained on natural language snippets of text and not verb-triplets. More in detail, these template-triplets are a sort of function, where the template is the body of the function and the object and the subject are the parameters, so that obtaining a natural language representation of these template-triples is straightforward by design, by replacing the instances of the parameters in the body. Differently from the verb in the verb-triplets, the template can be any snippet of text, possibly containing multiple verbs or referring to external concepts that are not the subject or the object. Also considering that serialising natural language into verb-triples is a challenging open-problem, template-triplets have the potential to fully harness the expressive power of existing deep language models. An example of template-triple (in the form subject, predicate, object) is: "the applicable law", "Surprisingly {subj} is considered to be clearly more related to {obj} rather than to something else.", "that Member State".

Importantly, as *information units*, to form the aforementioned template-triplets, [14] use meaningful decompositions of grammatical dependency trees, so to empower the units with the smallest granularity of information. As consequence, using such sub-trees as *information units* guarantees:

- a disentanglement of complex information bundles, into the most simple units, so to be able to correctly estimate the *level of detail* covered by the information pieces, as required by Definition 4.
- a better identification of duplicated units scattered throughout the information pieces, so to avoid an over-estimation of the *level of detail.*
- an easy way to understand whether an answer is invalid, as being totally contained in the question, hence forcing its *pertinence* to be zero.

All these properties meet the requirements that a good detail $d \in D$ should possess to be used for the generation of a DoX score, supporting our decision to re-use inside our pipeline the technology adopted by [14].

4.2.2. Archetypes Selection

According to Definition 1, an archetypal question is a very generic question characterised by one or more interrogative formulas. Literature is full of different examples of such archetypal questions, and many of them are used to classify both semantic and discourse relations [48, 49, 50, 51]. Interestingly, it is possible to identify a sort of hierarchy or taxonomy of such archetypes, ordered by their intrinsic level of specificity. For example, the simplest interrogative formulas (made only of an interrogative particle, i.e. what, why, when, who, etc.) can be seen as the most generic archetypes. While the more complex and composite is the formula (i.e. what-for, what-cause, etc.), the more specific is the question.

Hence, we decided to consider as set Q of main *archetypes* the most generic interrogative formulas used by literature [48, 49, 50, 51] to classify semantic relations within discourse. The main *archetypes* coming from Abstract Meaning Representation theory [50] are: What? Who? How? Where? When? Which? Whose? Why?
We refer to these archetypes as the *primary* ones because they consist only of interrogative particles. While the main *archetypes* coming from PDTB-style discourse theory [51] (also called *secondary archetypes* because they make use of the *primary archetypes*) are: In what manner? What is the reason? What is the result of it? What is an example of it? After what? While what? In what case? Despite what? What is contrasted with it? Before what? Since when? What is similar to it? Until when? Instead of what? What is an alternative to it? Except when? Unless what?

Despite the fact that many other archetypes may be identified (i.e. “Where to?” or “Who by?”), we believe that the list of questions we provided is rich enough to be generally representative for any other question[9], whereas more specific questions can be always framed by using the interrogative particles (i.e. why, what, etc.) we considered. In fact, *primary archetypes* can be used to represent any fact and abstract meaning [52], while the *secondary archetypes* can cover all the discourse relations between them (at least according to the

[9] For concrete examples of how all these questions (especially the primary ones) are related to XAI algorithms, we point the reader to this recent survey by IBM Research [31] or to Section 3.5.

PDTB theory).

5. Experiments

In Section 4.1 we argued that the degree of explainability of any collection of text (i.e. the output of a XAI-based system) can be measured in terms of DoX on a set of chosen *Explanandum Aspects*. In order to verify this assertion and hypothesis 1, we have to show that there is a strong correlation between our DoX and the perceived amount of *explainability*. To this end, we devised two experiments using some XAI-based systems:

- a Heart Disease Predictor based on XGBoost [11] and TreeSHAP [12];
- a Credit Approval System based on a simple Artificial Neural Network and on CEM [13].

More precisely, with the first experiment we measured explainability *directly*, while with the second we performed *indirect* measurements obtained through a few user-studies with human subjects.

Measuring explainability *directly* is not possible without a metric like the one we propose (DoX), except for a few naive cases. One of these cases is surely when a simple XAI-based system is considered. In fact, in a standard XAI-based system, the amount of *explainability* is by design, clearly and explicitly dependent on the output of the underlying XAI, for the black-box not being explainable by nature. So that, by masking the output of the XAI, the overall system can be forced to be not explainable enough. This characteristic can be exploited to partially verify hypothesis 1, but not in a generic way because this type of verification is based on a comparison with lack of explainability and not different degrees of it.

This is why we also need to measure explainability *indirectly*, with a second experiment, to understand whether our DoX correlates with the expected effects of explainability on human subjects. In fact, if the hypothesis is correct, the lower is DoX, the less explanations can be extracted, the less effective (as per

ISO 9241-210) is likely to be an explainee in achieving those explanatory goals that are not covered by the explanations. Hence, once fixed all the components that may affect effectiveness, including the presentation logic (the mechanism for re-elaborating explainable information into explanations) and the explanandum, an increase in DoX should always correspond to a proportional increase in effectiveness, at least in those tasks covered by the information provided by the increment of DoX. To show this, we borrowed and extended the results of two independent user-studies [14, 7], studying how DoX correlates with the effectiveness scores measured by these studies.

Therefore, in the following sections we are discussing more in detail:

- What are the two XAI-based systems object of these experiments.
- Which pertinence functions p and threshold t we considered for computing the DoX scores and why.
- How we conducted the two experiments, i.e. how we identified a set A of Explanandum Aspects in each experiment.

5.1. XAI-Based Systems

The XAI-Based Systems we considered for the two experiments of this paper are a Credit Approval System and a Heart Disease Predictor, respectively on finance and healthcare topics. Both these two systems are an example of Normal XAI Explainer (NXE), a One-Size-Fits-All explanatory mechanism providing the bare output of the 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.

5.1.1. Finance: Credit Approval System

The Credit Approval System (or CA in short) is the same used also in [14, 7], designed by IBM to showcase its XAI library: AIX360 [53]. This explanandum is about finance and the system is used by a bank. The bank deploys an Artificial Neural Network to decide whether to approve a loan request, and it uses the Contrastive Explanations Method (CEM) [13] algorithm to create post-hoc contrastive explanatory information. This information is meant to help the customers, showing them what minimal set of factors is to be manipulated for changing the outcome of the system from denial to approval (or vice-versa).

The Artificial Neural Network was trained on the “FICO HELOC” dataset [54]. The FICO HELOC dataset contains anonymized information about Home Equity Line Of Credit (HELOC) applications made by real homeowners. While a HELOC is a line of credit typically offered by a US bank as a percentage of home equity. Importantly, the Artificial Neural Network is trained to properly answer the following question: “What is the decision on the loan request of applicant X?”.

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. This is why the output of CEM is designed to answer the questions:

- What are the easiest factors to consider in order to change the result of applicant X’s application?
- How factor F should be modified in order to change the result of applicant X’s application?
- What is the relative importance of factor F in changing the result of applicant X’s application?

Nonetheless many other relevant questions might be to answer before the user is satisfied, reaching its goals. These questions include: “How to perform those minimal actions?”, “Why are these actions so important?”, etc..

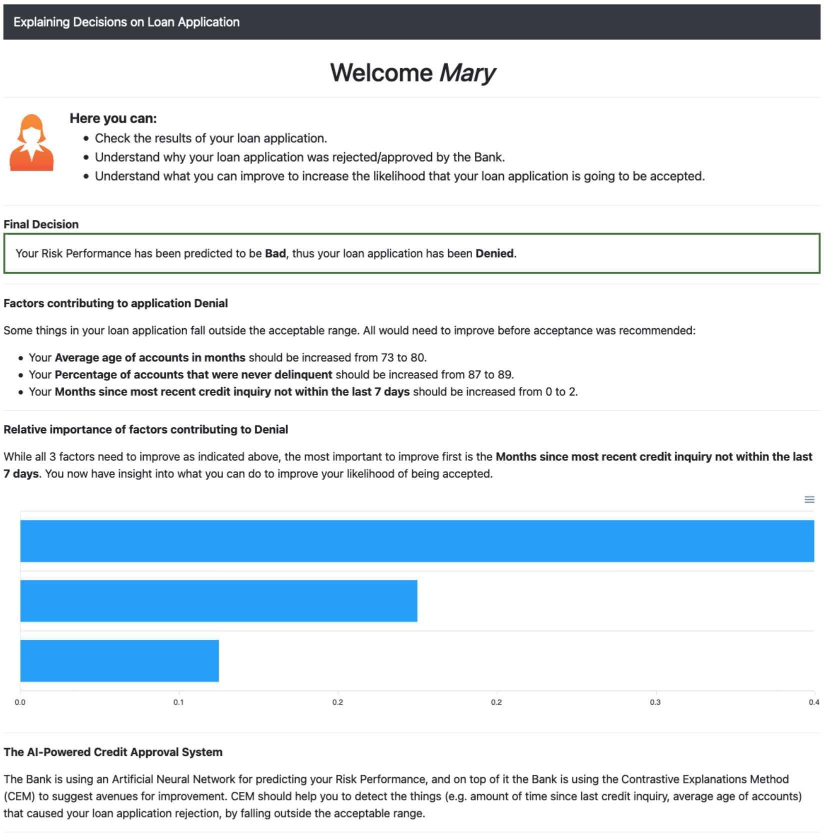
More precisely, the output of the Credit Approval System is composed by:

- Context: a titled heading section kindly introducing Mary (the user) to the system.
- AI Output: the decision of the Artificial Neural Network for the loan

application. This decision normally can be “denied” or “accepted”. For Mary it is: “denied”.

- XAI Output: a section showing the output of CEM. This output consists in a minimal ordered list of factors that are the most important to change for the outcome of the AI to switch.

A screenshot of this Credit Approval System is shown in figure 2.



Explaining Decisions on Loan Application

Welcome *Mary*

Here you can:

- Check the results of your loan application.
- Understand why your loan application was rejected/approved by the Bank.
- Understand what you can improve to increase the likelihood that your loan application is going to be accepted.

Final Decision

Your Risk Performance has been predicted to be **Bad**, thus your loan application has been **Denied**.

Factors contributing to application Denial

Some things in your loan application fall outside the acceptable range. All would need to improve before acceptance was recommended:

- Your **Average age of accounts in months** should be increased from 73 to 80.
- Your **Percentage of accounts that were never delinquent** should be increased from 87 to 89.
- Your **Months since most recent credit inquiry not within the last 7 days** should be increased from 0 to 2.

Relative importance of factors contributing to Denial

While all 3 factors need to improve as indicated above, the most important to improve first is the **Months since most recent credit inquiry not within the last 7 days**. You now have insight into what you can do to improve your likelihood of being accepted.

The AI-Powered Credit Approval System

The Bank is using an Artificial Neural Network for predicting your Risk Performance, and on top of it the Bank is using the Contrastive Explanations Method (CEM) to suggest avenues for improvement. CEM should help you to detect the things (e.g. amount of time since last credit inquiry, average age of accounts) that caused your loan application rejection, by falling outside the acceptable range.

Figure 2: **Screenshot of the Credit Approval System**

5.1.2. Health: Heart Disease Predictor

Similarly to the Credit Approval System, also the Heart Disease Predictor (or HD in short) comes from [7]. This explanandum is about health and the system is used by a first level responder of a help-desk for heart disease prevention. The system uses XGBoost [11] to predict the likelihood of a patient having a heart disease given its demographics (gender and age), health (diastolic blood pressure, maximum heart rate, serum cholesterol, presence of chest-pain, etc.) and the electrocardiographic (ECG) results. This likelihood is classified into 3 different risk areas: low (probability p of heart disease below 0.25), medium ($0.25 < p < 0.75$) or high. XGBoost is used to answer the following questions:

- How is likely that patient X has a heart disease?
- What is the risk of heart disease for patient X?
- What is the recommended action, for patient X to cure or prevent a heart disease?

The dataset used to train XGBoost is the “UCI Heart Disease Data”[55, 56]. TreeSHAP[12], a famous XAI algorithm specialised on tree ensemble models (i.e. XGBoost) for post-hoc explanations, is used to understand what is the contribution of each feature to the output of the model (that is XGBoost). TreeSHAP can be used to answer the following questions:

- What would happen if patient X would have factor Y (e.g. chest-pain) equal to A instead of B?
- What are the most important factors contributing to the predicted likelihood of heart disease, for patient X?
- How factor Y contributes to the predicted likelihood of heart disease, for patient X?

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. In other words, 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 a caller. To this end, the questions answered by TreeSHAP are quite useful, but many other important questions should probably 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..

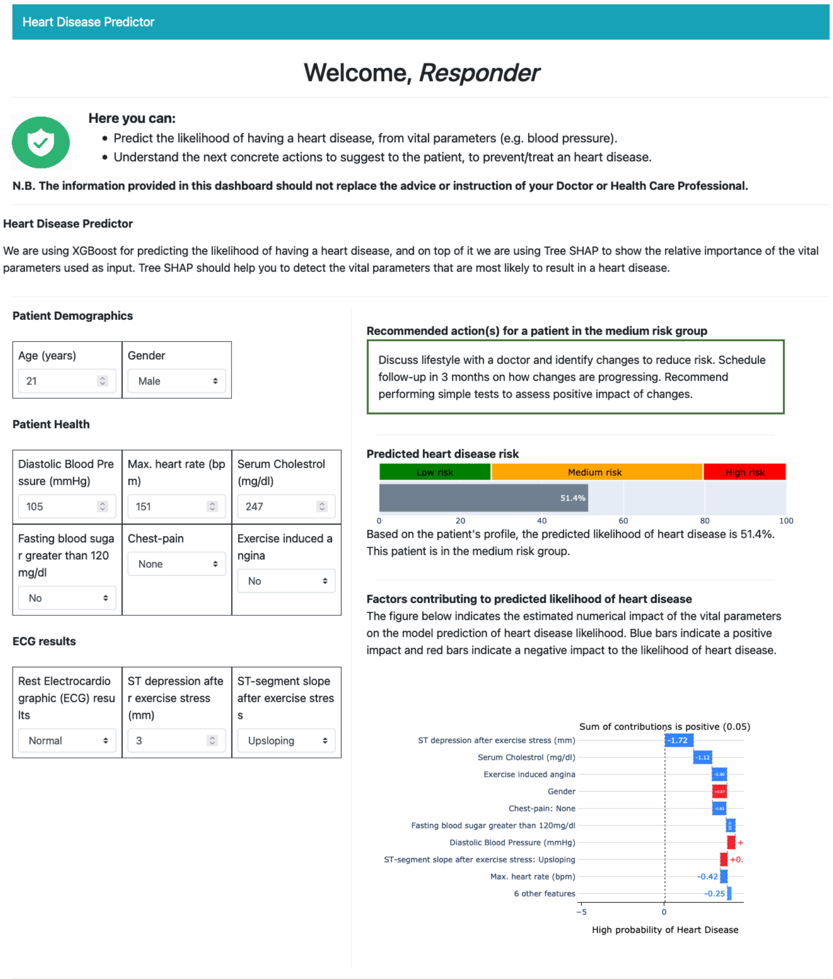
More precisely, the output of the Heart Disease Predictor is composed by:

- Context: a titled heading section kindly introducing the responder (the user) to the system.
- AI Inputs: a panel for inserting the patient's parameters.
- AI Outputs: a section displaying the likelihood of heart disease estimated by XGBoost and a few generic suggestions about the next actions to suggest.
- XAI Outputs: a section showing the contribution (positive or negative) of each parameter to the likelihood of heart disease, generated by TreeSHAP.

A screenshot of this Heart Disease Predictor is shown in figure 3.

5.2. Pertinence Functions and Thresholds

According to Definition 4, we need to define a pertinence function p and pick a threshold t in order to compute the DoX. As discussed in Section 4.2, we are going to use as pertinence function p a deep language model for Question-Answer Retrieval. The point is that many different deep language models exist for this task, i.e. [42, 43, 9], and each one of them has different characteristics producing different pertinence scores. So, which model is the right one for computing the DoX? Can we use any model?



Heart Disease Predictor

Welcome, *Responder*

Here you can:

- Predict the likelihood of having a heart disease, from vital parameters (e.g. blood pressure).
- Understand the next concrete actions to suggest to the patient, to prevent/treat an heart disease.

N.B. The information provided in this dashboard should not replace the advice or instruction of your Doctor or Health Care Professional.

Heart Disease Predictor

We are using XGBoost for predicting the likelihood of having a heart disease, and on top of it we are using Tree SHAP to show the relative importance of the vital parameters used as input. Tree SHAP should help you to detect the vital parameters that are most likely to result in a heart disease.

Patient Demographics

Age (years) 21	Gender Male

Patient Health

Diastolic Blood Pressure (mmHg) 105	Max. heart rate (bpm) 151	Serum Cholestrol (mg/dl) 247
Fasting blood sugar greater than 120 mg/dl No	Chest-pain None	Exercise induced angina No

ECG results

Rest Electrocardiographic (ECG) results Normal	ST depression after exercise stress (mm) 3	ST-segment slope after exercise stress Upsloping

Recommended action(s) for a patient in the medium risk group

Discuss lifestyle with a doctor and identify changes to reduce risk. Schedule follow-up in 3 months on how changes are progressing. Recommend performing simple tests to assess positive impact of changes.

Predicted heart disease risk

Based on the patient's profile, the predicted likelihood of heart disease is 51.4%. This patient is in the medium risk group.

Factors contributing to predicted likelihood of heart disease

The figure below indicates the estimated numerical impact of the vital parameters on the model prediction of heart disease likelihood. Blue bars indicate a positive impact and red bars indicate a negative impact to the likelihood of heart disease.

Sum of contributions is positive (0.05)
ST depression after exercise stress (mm)
-1.72
Serum Cholestrol (mg/dl)
-1.12
Exercise induced angina
-0.90
Gender
+0.87
Chest-pain: None
-0.83
Fasting blood sugar greater than 120mg/dl
-0.55
Diastolic Blood Pressure (mmHg)
+
ST-segment slope after exercise stress: Upsloping
+0.
Max. heart rate (bpm)
-0.42
6 other features
-0.25
−5
0
High probability of Heart Disease

Figure 3: **Screenshot of the Heart Disease Predictor**

To answer these questions, during our experiments we decided to study the behaviour of more than one deep language model, as pertinence function p. In fact, assuming that these models get good results on state-of-the-art benchmarks for *pertinence estimation*, i.e. [9, 40], we believe that the results of the computation of DoX should be consistent across them. Hence the models we considered are:

- FB: published by [9] and [47], and trained on the combination of the following datasets: Natural Questions [57], TriviaQA [58], WebQuestions [59], and CuratedTREC [60].

- TF: or Multilingual Universal Sentence Encoder [40] and trained on the Stanford Natural Language Inference (SNLI) corpus [10].

Furthermore, we found that different pertinence thresholds t had to be considered for TF and FB. We experimentally found on the two XAI-based systems presented in Section 5.1 that for FB a good pertinence threshold is $t = 0.55$, while for TF is $t = 0.15$.

5.3. 1st Experiment: Direct Evaluation on Normal XAI-generated Explanations

The 1st experiment is meant to shed more light on how a few changes to the explainability of a system affect the estimated DoX. Specifically, XAI-based systems are considered for this experiment, instead of other AI-based systems, because their amount of *explainability* is by design, clearly and explicitly dependent on the output of the underlying XAI. So that, by masking the XAI output, the overall system can be forced to be less explainable. Hence, this characteristic can be exploited to (at least partially) verify hypothesis 1, in a very simple but effective way. With this experiment we compare the DoX of the output of a Normal XAI Explainer (NXE) with the DoX of that same information without the XAI (namely a NAE, as Normal AI-based Explanation). We expect the (averaged) DoX of the NXE to be clearly higher than NAE.

For this experiment, we used the XAI-based systems defined in Section 5.1. In fact, both the Credit Approval System and the Heart Disease Predictor are examples of NXE. Therefore, by simply removing the output of the XAI (respectively CEM and TreeSHAP) from these systems we obtain a NAE. In order to compare the (averaged) DoX of a NXE to that of its NAE, as set of *Explanandum Aspects* we take those targeted by the XAI of the NXE and by the AI of both the NAE and the NXE. More precisely, the main *Explanandum Aspects* targeted by XGBoost [11] and TreeSHAP [12] in the Heart Disease Predictor (HD) are 5:

- The recommended action for patient X
- The most important factors that contribute to predict the likelihood of heart disease
- The likelihood of heart disease
- The risk R of having a heart disease
- The contribution of Y to predict the likelihood of heart disease for patient X

While the main *Explanandum Aspects* targeted by the Artificial Neural Network and CEM [13] in the Credit Approval System (CA) are 4:

- The easiest factors to consider for changing the result
- The relative importance of factor F in changing the result of applicant X's application
- Applicant X's risk performance
- The result of applicant X's application

For computing the DoX, we used the pipeline described in Section 4 to extract from the textual representations of the NXE and the NAE different knowledge graphs of details D. After properly converting the images produced by the NXE to textual explanations, the resulting *Explanandum Aspects Coverage* of NXE for both HD and CA is 100%, while that of NAE is 60% for HD and 50% for CA. After extracting the knowledge graphs, we used the set of *Explanandum Aspects* A to select from them all the details D_a. We did it for each $a \in A$, checking every a against the nodes of the graph, exploiting the properties of the template-triplets (see Section 4.2) to identify every detail $d \in D_a$. More precisely, we were able to understand whether a template-triplet is likely to be related to an $a \in A$ by using the algorithm[10] described in [61] and used also

[10] This algorithm simply computes the similarity between a and the subject/object of the triplet. If the similarity is above a given threshold, then the triplet is said to be related to a.

by [43]. Finally, we were able to compute the DoX scores in accordance with Definition 4 by using:

- the set of archetypes Q described in Section 4.2.2,
- the pertinence functions p and the thresholds t presented in Section 5.2,
- the aforementioned set of details D_a of each $a \in A$.

Results are presented and discussed in Section 6.

5.4. 2nd Experiment: Explainability vs Effectiveness

Differently from the first one, this second experiment aims to understand if changing the DoX implies a change also in the effects of explainability on the explainees. If hypothesis 1 is correct, we would expect that an increment in the DoX of some information i (i.e. the explanandum support material) would seemingly correspond to an increment of effectiveness of the explanations generated from i by an explainer. This is why for this experiment we borrowed and extended the user-studies published by [14, 7], involving more than 160 human subjects. Importantly, these user-studies consider the same XAI-based systems used during the first experiment and described in Section 5.1. More in detail, each user-study analyses the usability (in terms of effectiveness, efficiency, satisfaction) of the explanations given by the XAI-systems, when changing i or the way i is re-elaborated into explanations.

Importantly, we considered two user-studies instead of just one because they are performed on two different user pools. The benefit of considering different pools is that the conclusions and the final claim might be strengthened, for the results being more generic and statistically relevant. In the following subsections we will present the two user-studies more in detail.

5.4.1. 1st User-Study

The first user-study comes from [14]. [14] present a novel mechanism (that we call OBE, in short, for Overview-Based Explainer) to explain large collections of heterogeneous documents (i.e. more than 50 web-pages) about the

Credit Approval System, in a user-centred and interactive way. This is done by organising knowledge as a graph of abstract aspects whose related explanations are ordered by relevance and simplicity, according to a set of pre-defined archetypal questions (what, how, when, why, etc.). To show that OBE makes more usable explanations, the authors compare the usability scores of the simple Credit Approval System (a NXE) with a version of it enhanced by OBE. The usability scores are generated by the users interacting with the system and answering to:

- a quiz on the Credit Approval System comprising 7 different questions, used to estimate the effectiveness and efficiency (in terms of time to complete the quiz) of explanations;
- a System Usability Scale (SUS) questionnaire [62], used to estimate satisfaction.

For the user-study 103 different participants were recruited (57 males, 44 females, 2 unknowns, ages 18-55) on the online platform Prolific [63]. All the participants were recruited among those who: 1. are resident in UK, US or Ireland; 2. have a Prolific's acceptance rate greater or equal to 75%[11]. Participants were randomly allocated to test only one of the two versions of the Credit Approval System: either NXE or OBE. In the end, 51 participants evaluated NXE and 52 evaluated OBE. For more details about the evaluation, please read [14].

Importantly, the results of the user-study show that OBE produces significantly greater effectiveness scores than NXE. Indeed, the difference between NXE and OBE is twofold. First of all, the explanations of OBE are interactive and more user-centred, while those of NXE are not. Secondly, NXE considers for its explanations a smaller amount of explainable information than OBE, in fact OBE builds its explanations using more than 50 extra web-pages that NXE does not see. This last difference between NXE and OBE is exactly what allows us to exploit the user-study to verify hypothesis 1. In fact, the amount

[11] Mainly because they are unlikely to answer poorly/randomly to questions.

of information handled by NXE is roughly $\frac{1}{100}$ of OBE and this is enough to say that the explainability of NXE is not none and it is probably lower than OBE. In other terms, if hypothesis 1 holds, then NXE should have a lower DoX than OBE (at least in those *Explanandum Aspects* covered by the additional information used by OBE).

In order to use this first user-study, to show that an increment in DoX causes an increment in the effectiveness of the explanations, we have to compute the DoX scores of NXE and OBE as in the first experiment. To do so, we identified the set of *Explanandum Aspects* A from the quiz used to generate the effectiveness scores, applying on it the same technique for knowledge graph extraction mentioned in Section 4.2. In fact, the quiz defines exactly what the users should know in order to be effective, indirectly defining also what is important for the system to explain: the *Explanandum Aspects*. The resulting DoX scores are given in Section 6.

5.4.2. 2nd User-Study

The second user-study comes from [7]. Differently from the first one, this user-study is on both the Credit Approval System and the Heart Disease Predictor and it analyses an extension of OBE called YAI4Hu together with other two explainers: 2EC and HWN. The 2EC explainer is static, as the NXE, in fact it is made of the output NXE directly connected to a 2nd (non-expandable) level of information consisting in an exhaustive and verbose set of autonomous static explanatory resources. The 2EC is organized therefore as a very long text document (more than 50 pages per system, when printed), structured in titled sections and prefixed with a table of content with hypertext links. On the other hand, the HWN explainer is exactly like the OBE but it uses only the archetypes "why" and "how" for generating an explanation. Also YAI4Hu is similar to OBE, but it adds a mechanism for users to ask their own questions to the system. Importantly, YAI4Hu, HWN and 2EC use exactly the same explanandum support material used for OBE during the first user-study.

For the user-study, [7] recruited 64 different participants among the university students of a few different courses of study[12]:

- Bachelor Degree in Computer Science
- Bachelor Degree in Management for Informatics
- Master Degree in Digital Humanities
- Master Degree in Artificial Intelligence

Similarly to the first user-study, each participant evaluated the XAI-based systems by answering to a quiz per system and a SUS questionnaire. For more details about the evaluation, please read [7].

The 64 participants were randomly allocated to test only one of the three types of explainer (YAI4Hu, HWN or 2EC) on both the Credit Approval System and the Heart Disease Predictor, so that each participant evaluated both the XAI-based systems. In the end, there were approximatively 20 participants per explainer. The results of the user-study published by [7] show that YAI4Hu produces significantly greater effectiveness scores than HWN, and that HWN produces better results than 2EC as well. Though, considering that YAI4Hu, HWN and 2EC share the same explanandum support material (of OBE), these results tell us only that (unsurprisingly) changing the explainer might change the quality of the explanations. In other words, differently from the first user-study, these results do not show any improvement in effectiveness due to changes in the explainability of the explanandum support material, as we would need to make our point. Considering that, we decided to extend the user-study recruiting 19 more participants[13] from the same pool of users, asking them to answer the same quizzes and questionnaires but on NXE. Indeed, NXE has a smaller explanandum support material than YAI4Hu, HWN, 2EC and OBE.

[12] All the courses of study were of an Italian university, and only the master degrees were international, with English teachings and students from countries other than Italy.

[13] We made sure that none of our 19 extra participants was involved in the user-study published by [7].

After this modification, also the second user-study can be used to check whether there is a non-spurious correlation between the DoX scores and the perceived effectiveness score. That is because we can compare the scores of NXE against the others, having a situation where it is possible to study the effects on effectiveness of different explanandum support materials. To do so, we have to compute the DoX scores of NXE, YAI4Hu, HWN and 2EC. We did it also by identifying the set of *Explanandum Aspects A* from the quizzes used to generate the effectiveness scores, as in the first user-study. As result we were able to identify 82 *Explanandum Aspects* for the Heart Disease Predictor and 40 for the Credit Approval System.

6. Results

In this Section we will present and discuss the results of the two experiments defined in Section 5.

6.1. 1st Experiment

Computing the DoX for the first experiment, we got the results displayed in Table 2, where for simplicity we show only the *primary archetypes.* As expected, on both the XAI-based systems, the results of the first experiment neatly show that the averaged DoX of NXE is way higher than NAE, regardless the adopted *deep language model.*

Nonetheless, considering that in this 1st experiment we arbitrarily picked a simple set of *Explanandum Aspects*, what would happen if we would consider different and more complex explicanda and explanatory contents? Furthermore, the result of the experiment is based on the comparison of the DoX of a non-explainable system (as NAE) with an explainable system, and this is a very peculiar and naive case to consider. Therefore, in order to fully verify hypothesis 1 we need to understand whether DoX is behaving properly also when explainability is present in different, non-zero, amounts. To do so, we envisage that explainability can be measured *indirectly*, by studying the effectiveness of

Table 2: **Experiment 1 - Degree of eXplainability**: in this table DoX and Averaged DoX are shown for the Credit Approval System (CA) and the Heart Disease Predictor (HD). As columns we have the different explanatory mechanisms used for experiment 1: NAE and NXE. As rows we have different explainability estimates using different deep language models for computing pertinence: FB and TF. For simplicity, with DoX we show only the *primary archetypes*.

		CA		HD	
		NAE	NXE	NAE	NXE
Avg DoX	FB	0.65	1.79	3.05	4.53
	TF	24.00	32.02	27.66	40.12
DoX	FB	"how": 0.65	"whose": 1.89	"which": 5.29	"what": 6.72
		"which": 0.64	"how": 1.84	"what": 4.59	"which": 6.53
		"whose": 0.63	"why": 1.829	"how": 4.35	"how": 5.63
		"what": 0.62	"which": 1.821	"whose": 2.43	"whose": 4.16
		"who": 0.617	"where": 1.598	"when": 2.3	"why": 3.87
		"when": 0.614	"when": 1.597	"why": 2.12	"where": 3.81
		"where": 0.6	"what": 1.57	"where": 2.09	"when": 3.5
		"why": 0.58	"who": 1.38	"who": 2.08	"who": 3.25
	TF	"when": 25.22	"which": 32.66	"whose": 28.232	"what": 41.52
		"which": 24.03	"when": 32.32	"what": 28.231	"which": 41.31
		"what": 22.64	"why": 31.10	"how": 28.13	"how": 40.96
		"why": 22.22	"how": 30.65	"which": 27.93	"whose": 40.80
		"whose": 22.06	"what": 30.23	"why": 27.78	"why": 40.26
		"how": 21.97	"whose": 30.21	"where": 27.4	"where": 39.75
		"where": 21.49	"where": 29.54	"when": 27.3	"when": 39.54
		"who": 20.93	"who": 29.52	"who": 27.13	"who": 39.35

the resulting explanations. In fact, we have that more explainability implies a greater ability to explain, therefore more explanations.

In short, the lower the DoX, the less explanations can be produced, the less effective is likely to be an explainee on the tasks related to the explanandum. So, if hypothesis 1 is correct, once fixed all the components that may affect effectiveness (including the presentation logic (the mechanism for re-elaborating explainable information into explanations) and the explanandum) an increase in the DoX of the (explanatory) system should always correspond to a proportional increase of its effectiveness, at least in those tasks covered by the information provided by the increment of DoX.

6.2. 2nd Experiment

During this second experiment we studied if there is a correlation between the effectiveness scores and the DoX scores. We did it considering 5 different types of explainer, 2 different user pools and 2 different XAI-based systems. The 2 user pools come from 2 separate user-studies, as discussed in Section 5. Both the user-studies compare NXE to different types of explainers using larger explanandum support materials. The first user-study comes from [14], while the second is an extension to the user-study presented by [7]. The approach we followed to extend this latter study is described in Section 5.4.2.

The results we got by extending the 2nd user-study are summarised in the box-plots of Figure 4, where the median effectiveness score of NXE is seen to be significantly lower than 2EC (this implies that is also lower than HWN and YAI4Hu). More precisely, in figure 4 is shown a consistent increment in median effectiveness of 2EC on those questions which aspects are *not covered* by the information presented with NXE. This increment in effectiveness is of 13.33% in HD and 10% in CA, and it is higher in HD probably because the difference between the averaged DoX of NXE and 2EC is larger, for both FB and TF, than CA.

We performed a one-sided Mann-Whitney U-Test (a non-parametric version of the t-test for independent samples) under the alternative hypothesis that NXE effectiveness is stochastically less than 2EC. The results confirmed that, at least for HD, there is significant statistical evidence to support the fact that 2EC scores are greater than NXE not just by chance. In fact we obtained a p-value equal to 0.007[14] with $U = 45$ for HD, and $p = 0.12$ with $U = 118.5$ for CA.

[14] A $p < 0.05$ is normally considered to be a significant statistical evidence.

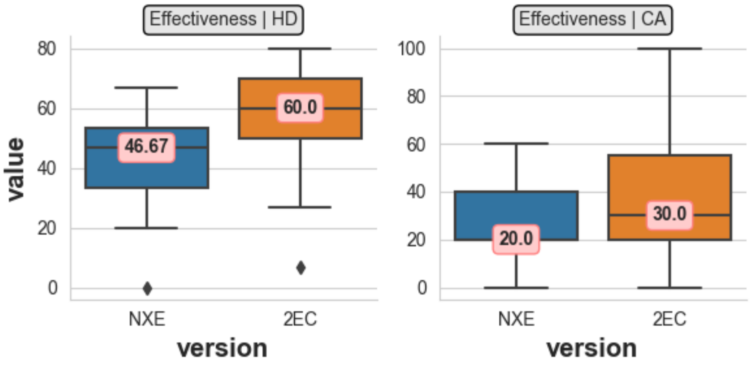


Figure 4: **NXE vs 2EC - Effectiveness Scores on the Questions *not covered* by NXE**: Comparison of NXE's results (the blue ones) with 2EC's (the orange ones), only on those questions which aspects are *not covered* by the information presented with NXE. Results are shown in the form of box plots (25th, 50th, 75th percentile, and whiskers covering all data and outliers). The numerical value of medians is shown inside pink boxes. The 1st column is for the Heart Disease Predictor, while the 2nd for the Credit Approval System.

Unexpectedly, we can see in figure 5 an important increment in effectiveness of 2EC also on those questions (questions 1 and 6 in the CA quiz and questions 1, 2 and 3 in the HD quiz of [7]) which aspects are specifically *covered* by the information of NXE. This increment is even higher than the previous one and it is of 33% in HD and 25% in CA.

Given both the results of the first and the second user-study we can conclude that all the explainers with a larger explanandum support material (OBE, 2EC, HWN, YAI4Hu) have greater effectiveness scores than NXE. This is true even for 2EC, despite the fact that it is not interactive and it does not re-organise information to make it simpler and easier to access. Indeed, 2EC dumps on the users hundreds of pages of contents making very challenging the task of identifying the right information to answer the questions of the quizzes used to measure effectiveness. Now, if hypothesis 1 is true, we expect that the higher is DoX, the higher is the effectiveness of an explainer. Importantly, the opposite is not necessarily true, in fact, two explainers (with different presentation logics) might have different effectiveness scores despite having the same DoX, i.e. 2EC

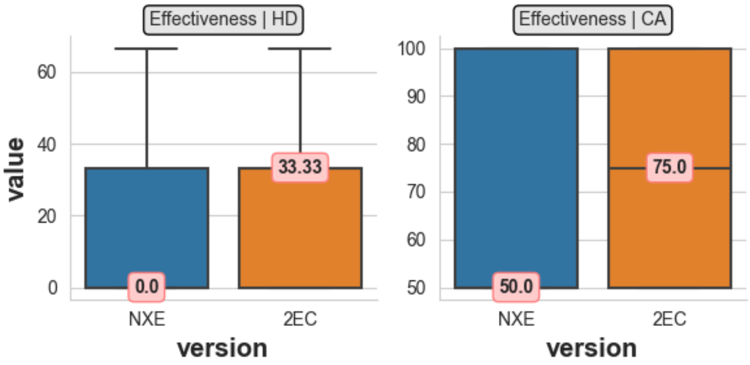


Figure 5: **NXE vs 2EC - Effectiveness Scores on the Questions *covered* by NXE**: Comparison of NXE's results (the blue ones) with 2EC's (the orange ones), only on those questions (questions 1 and 6 in the CA quiz and questions 1, 2 and 3 in the HD quiz) which aspects are specifically *covered* by NXE's information. 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 column is for the Heart Disease Predictor, while the 2nd for the Credit Approval System.

and YAI4Hu.

Computing the DoX scores for the second experiment we got the results shown in Table 3. These results confirm our expectations for them. In fact, they show that 2EC, OBE, HWN and YAI4Hu have higher DoX scores than NXE.

7. Discussion

The results of the first experiment tell us that whenever new information about different aspects to be explained is added to the explanandum support material (see Section 4.1 for a definition of what it is), the DoX scores increase, and this is true also when changing the set of Explanandum Aspects, as we did with the second experiment. Furthermore, the results of the second experiment tell us that whenever the DoX scores increase, the overall effectiveness of the explanations generated from the explanandum support material increase as well. To strengthen the outcomes of the second experiment we also have that

Table 3: **Experiment 2 - Degree of eXplainability**: in this table DoX and Averaged DoX are shown for the Credit Approval System (CA) and the Heart Disease Predictor (HD). As columns we have the different explanatory mechanisms used for experiment 2: NXE and 2EC/HWN/OBE/YAI4Hu (the “others”). As rows we have different explainability estimates using different deep language models for computing pertinence: FB and TF. For simplicity, with DoX we show only the *primary archetypes*.

		CA		HD	
		NXE	Others	NXE	Others
Avg DoX	FB	3.18	14.63	1.21	12.25
	TF	28.73	41.52	35.42	38.85
DoX	FB	"which": 3.75	"why": 15.51	"which": 1.15	"why": 12.83
		"when": 3.7	"how": 14.95	"why": 1.13	"when": 12.64
		"how": 3.55	"when": 14.91	"what": 1.12	"which": 12.13
		"why": 3.19	"who": 14.32	"how": 0.87	"how": 12.04
		"what": 2.82	"which": 14.12	"whose": 0.58	"what": 11.68
		"who": 2.71	"whose": 13.9	"when": 0.56	"who": 11.31
		"whose": 2.13	"where": 13.71	"who": 0.557	"whose": 11.02
		"where": 1.71	"what": 11.74	"where": 0.555	"where": 10.0
	TF	"when": 30.24	"why": 43.08	"why": 36.12	"why": 39.4
		"which": 29.48	"when": 41.43	"whose": 35.51	"when": 39.3
		"why": 27.87	"how": 40.99	"which": 35.47	"who": 38.02
		"whose": 27.77	"whose": 40.68	"how": 35.25	"what": 37.85
		"where": 27.23	"who": 40.24	"who": 35.24	"how": 37.66
		"who": 26.93	"which": 40.23	"what": 34.65	"whose": 36.82
		"how": 26.67	"what": 40.2	"where": 34.54	"which": 36.43
		"what": 25.27	"where": 39.75	"when": 34.45	"where": 36.42

the user-studies involved more than 160 participants and were consistent across two different and sufficiently broad pools, producing statistically significant results. Therefore, considering that *explainability* is fundamentally the *ability to explain*, the two experiments combined together tell us that our (averaged) DoX can quantitatively approximate the degree of explainability of information. In fact, we showed in Section 4.1 that the DoX scores are computed by aggregating together information about the level of detail, exactness and fruitfulness of information, quantifying how this information explains in terms of answers to different (archetypal) questions.

Importantly, in both the experiments no inconsistencies were found across the considered pertinence functions (TF and FB; see Section 5.2). This suggests that the alignment of DoX with explainability may be independent from the chosen deep language model, at least in all the experiments we considered. We believe that this is happening because both TF and FB, in average, perform reasonably well on the same benchmarks for evaluating Question-Answer Retrieval. In other terms, it could be that if the Averaged DoX aggregates enough archetypes and the number of considered aspects and details is also enough, then different pertinence functions performing in similar ways on some good benchmarks may produce similar Averaged DoX scores despite their differences (i.e. the archetype with the best explanatory illocution in TF is "what", while in FB is "which"). Anyway, this does not exclude the fact that there might be a deep language model that is better than the others for computing the DoX, or that multiple standardised deep language models should be adopted for a thorough estimate of the DoX. We leave this analysis for future work.

We believe that this new metric we propose to measure the amount of explainability may have a large impact in all those applications where it is important to objectively evaluate explainability, i.e. for an impact assessment or for generating more user-centred explanations. The benefits of using DoX over a normal user-study are manifold, in fact:

- it removes the costs normally sustained during subject-based evaluations;
- it allows to directly measure the degree explainability of any piece of information that has a meaningful textual representation written in a natural language (i.e. English);
- it disentangles the evaluation of the explanandum support material from that of the explainer (or presentation logic) and the interface.

In other terms, DoX could be used to understand whether a piece of information is enough to explain something. Indeed, our DoX is a fully objective metric that can evaluate the explainability of any textual information and understand whether the amount of explainability is objectively poor, even if the resulting explanations are perceived as satisfactory and good by the explainees. We deem that this characteristic of DoX is very important, in fact if explanations are built over explainable information, a poor degree of explainability objectively implies poor explanations, no matter how good the adopted explanatory process is (perceived): "Users also do not necessarily perform better with systems that they prefer and trust more. To draw correct conclusions from empirical studies, explainable AI researchers should be wary of evaluation pitfalls, such as proxy tasks and subjective measures" [64].

Though, there are a few characteristics of our DoX that require some extra discussion in order to fully understand the potential and also the limitations of this technology. First of all, in order to compute DoX a set of Explanandum Aspects A is needed, as per Definition 4. It is clear that this set of aspects is task specific, changing from explanandum to explanandum. In other words, for computing the DoX a precise definition of what has to be explained is required, without it we could not compute any score summarising the degree of explainability of information.

Despite the fact that A might be (manually) specified by a subject, the final score is still measured objectively with respect to any A. On the other hand, identifying a proper A is not enough, for estimating the DoX also a set of archetypes Q is needed. Considering the impressive and possibly infinite amount of archetypal questions that our language can conceive, it would appear that also the choice of Q might be a source of subjectivity. But questions and archetypes have been studied for a very long time in linguistics, resulting in many theories capable of organising our understanding of what constitutes a discourse and a representation of knowledge. This is why we assert that instead of relying on subjective choices of Q, we can exploit the plethora of (what we call) archetypal questions, identified by linguistic theories as those discussed in Section 4.2.2.

8. Conclusions and Future Work

The long-term goal of this paper is to change and improve the interaction between organisations and individuals, by the automated assessment of the Degree of eXplainability (DoX) of AI-based systems or (more generally) explainable information. This is why we described an algorithm for objectively quantifying the DoX of information, by estimating the number and quality of the explanations it could generate on the most important aspects to be explained.

In order to understand whether the DoX is actually behaving as *explainability* is expected to, we designed a few experiments on two realistic AI-based systems for heart disease prediction and credit approval, involving famous AI technology as Artificial Neural Networks, TreeSHAP [12], XGBoost [11] and CEM [13]. The results we obtained show that the DoX is aligned to our expectations, and it is possible to actually quantify *explainability* in natural language information.

Surely this does not imply that an estimate of the DoX, alone, is enough for a thorough impact assessment under the law. For example, starting from the point that explainable information (e.g. an explanation) can be incorrect, our definition of DoX does not consider the degree of correctness of information, assuming that truth is given and that it is a different thing from explainability. Anyway, we believe that this technology might be used for an Algorithmic Impact Assessment (AIA), as soon as a set of relevant *Explanandum Aspects* can be identified under the requirements of the law. Therefore, being able to select a reasonable threshold of *explainability* for law-compliance is certainly one of the next challenges we envisage for a proper standardisation of *explainability* in the industrial panorama.

References

References

[1] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, et al.,

Interpretability of deep learning models: A survey of results, in: 2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI), IEEE, 2017, pp. 1–6.

[2] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (5) (2019) 206–215.

[3] 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, Information Fusion 58 (2020) 82–115.

[4] J. Mazur, Right to access information as a collective-based approach to the gdpr's right to explanation in european law, Erasmus L. Rev. 11 (2018) 178.

[5] A. HLEG, Ethics guidelines for trustworthy ai (2019).

[6] M. E. Kaminski, G. Malgieri, Algorithmic impact assessments under the gdpr: producing multi-layered explanations, U of Colorado Law Legal Studies Research Paper (19-28). doi:10.2139/ssrn.3456224.

[7] F. Sovrano, F. Vitali, Generating user-centred explanations via illocutionary question answering: From philosophy to interfaces, arXiv preprint arXiv:2110.00762.
URL http://arxiv.org/abs/2110.00762

[8] C. D. Novaes, E. Reck, Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization, Synthese 194 (1) (2017) 195–215. doi:10.1007/s11229-015-0816-z.

[9] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answeringdoi:10.18653/v1/2020.emnlp-main.550.

[10] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inferencedoi:10.18653/v1/D15-1075.

[11] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794. doi:10.1145/2939672.2939785.

[12] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable ai for trees, Nature machine intelligence 2 (1) (2020) 56–67. doi:10.1038/s42256-019-0138-9.

[13] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, P. Das, Explanations based on the missing: Towards contrastive explanations with pertinent negatives, in: Advances in neural information processing systems, 2018, pp. 592–603. doi:10.5555/3326943.3326998.

[14] F. Sovrano, F. Vitali, From philosophy to interfaces: an explanatory method and a tool based on achinstein's theory of explanation, in: Proceedings of the 26th International Conference on Intelligent User Interfaces, 2021. doi:10.1145/3397481.3450655.

[15] G. Villone, L. Rizzo, L. Longo, A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence, 2020.

[16] A.-p. Nguyen, M. R. Martínez, On quantitative aspects of model interpretability, arXiv preprint arXiv:2007.07584.

[17] H. Lakkaraju, E. Kamar, R. Caruana, J. Leskovec, Interpretable & explorable approximations of black box models, arXiv preprint arXiv:1707.01154.

[18] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques, arXiv preprint arXiv:2103.01035.

[19] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects (2018).
URL `https://arxiv.org/abs/1812.04608`

[20] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (scs), KI-Künstliche Intelligenz (2020) 1–6.

[21] W. C. Salmon, Scientific explanation and the causal structure of the world, Princeton University Press, 1984. `doi:10.1515/9780691221489`.

[22] B. C. Van Fraassen, et al., The scientific image, Oxford University Press, 1980. `doi:10.1093/0198244274.001.0001`.

[23] P. Achinstein, The Nature of Explanation, Oxford University Press, 1983.
URL `https://books.google.it/books?id=0XI8DwAAQBAJ`

[24] J. Holland, K. Holyoak, R. Nisbett, P. Thagard, Induction: Processes of Inference, Learning, and Discovery, Bradford books, MIT Press, 1989.
URL `https://books.google.it/books?id=Z6EFBaLApE8C`

[25] W. S. Sellars, Philosophy and the scientific image of man, in: R. Colodny (Ed.), Science, Perception, and Reality, Humanities Press/Ridgeview, 1962, pp. 35–78.

[26] G. R. Mayes, Theories of explanation (2001).
URL `https://iep.utm.edu/explanat/`

[27] H. Leitgeb, A. Carus, Rudolf carnap (2021).
URL `https://plato.stanford.edu/archives/sum2021/entries/carnap/`

[28] D. J. Hilton, Mental models and causal explanation: Judgements of probable cause and explanatory relevance, Thinking & Reasoning 2 (4) (1996) 273–308.

[29] P. Achinstein, Evidence, Explanation, and Realism: Essays in Philosophy of Science, Oxford University Press, USA, 2010.
URL https://books.google.it/books?id=0oM8DwAAQBAJ

[30] I. Douven, Peter achinstein: Evidence, explanation, and realism: Essays in philosophy of science (2012). doi:10.1007/s11191-011-9405-9.

[31] Q. V. Liao, D. Gruen, S. Miller, Questioning the ai: informing design practices for explainable ai user experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–15.

[32] M. Ribera, A. Lapedriza, Can we do better explanations? a proposal of user-centered explainable ai., in: IUI Workshops, 2019.
URL http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf

[33] 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 SIGCHI Conference on Human Factors in Computing Systems, 2009, pp. 2119–2128. doi:10.1145/1518701.1519023.

[34] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligencedoi:10.1016/j.artint.2018.07.007.

[35] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89. doi:10.1109/DSAA.2018.00018.

[36] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gpdr, Harv. JL & Tech. 31 (2017) 841. doi:10.2139/ssrn.3063289.

[37] J. Rebanal, J. Combitsis, Y. Tang, X. Chen, Xalgo: a design probe of explaining algorithms' internal states via question-answering, in: Proceedings of the 26th International Conference on Intelligent User Interfaces, ACM, 2021. doi:10.1145/3397481.3450676.

[38] P. Jansen, N. Balasubramanian, M. Surdeanu, P. Clark, What's in an explanation? characterizing knowledge and inference requirements for elementary science exams, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 2956–2965. URL https://aclanthology.org/C16-1278

[39] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence (2019) 1033–1041doi:10.5555/3306127.3331801.

[40] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, et al., Multilingual universal sentence encoder for semantic retrievaldoi:10.18653/v1/2020.acl-demos.12.

[41] U. Roy, N. Constant, R. Al-Rfou, A. Barua, A. Phillips, Y. Yang, Lareqa: Language-agnostic answer retrieval from a multilingual pooldoi:10.18653/v1/2020.emnlp-main.477.

[42] M. Guo, Y. Yang, D. Cer, Q. Shen, N. Constant, Multireqa: A cross-domain evaluation for retrieval question answering models, arXiv preprint arXiv:2005.02507.

[43] F. Sovrano, M. Palmirani, F. Vitali, Legal knowledge extraction for knowledge graph based question-answering, in: Legal Knowledge and Information Systems: JURIX 2020. The Thirty-third Annual Conference, Vol. 334, IOS Press, 2020, pp. 143–153.

[44] G. Brun, Explication as a method of conceptual re-engineering, Erkenntnis 81 (6) (2016) 1211–1241. doi:10.1007/s10670-015-9791-5.

[45] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, C. Dugan, Explaining models: an empirical study of how explanations impact fairness judgment, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 275–285. `doi:10.1145/3301275.3302310`.

[46] A. Raymond, M. Malencia, G. Paulino-Passos, A. Prorok, Agree to disagree: Subjective fairness in privacy-restricted decentralised conflict resolution (2021).
URL `https://arxiv.org/pdf/2107.00032.pdf`

[47] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks`doi:10.18653/v1/D19-1410`.

[48] L. He, M. Lewis, L. Zettlemoyer, Question-answer driven semantic role labeling: Using natural language to annotate natural language, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 643–653. `doi:10.18653/v1/D15-1076`.

[49] N. FitzGerald, J. Michael, L. He, L. Zettlemoyer, Large-scale qa-srl parsing`doi:10.18653/v1/P18-1191`.

[50] J. Michael, G. Stanovsky, L. He, I. Dagan, L. Zettlemoyer, Crowdsourcing question-answer meaning representations`doi:10.18653/v1/N18-2089`.

[51] V. Pyatkin, A. Klein, R. Tsarfaty, I. Dagan, Qadiscourse–discourse relations as qa pairs: Representation, crowdsourcing and baselines`doi:10.18653/v1/2020.emnlp-main.224`.

[52] J. Bos, Expressive power of abstract meaning representations, Computational Linguistics 42 (3) (2016) 527–535.

[53] IBM, Ai explainability 360 - demo, `https://aix360.mybluemix.net/explanation_cust`, online; accessed 29-Mar-2020 (2019).

[54] S. Holter, O. Gomez, E. Bertini, Fico explainable machine learning challenge (2019).

URL https://fico.force.com/FICOCommunity/s/explainable-machine-learning-challenge?tabset-3158a=a4c37

[55] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, V. Froelicher, International application of a new probability algorithm for the diagnosis of coronary artery disease, The American journal of cardiology 64 (5) (1989) 304–310. doi:10.1016/0002-9149(89)90524-9.

[56] R. Alizadehsani, M. Roshanzamir, M. Abdar, A. Beykikhoshk, A. Khosravi, M. Panahiazar, A. Koohestani, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, A database for using machine learning and data mining techniques for coronary artery disease diagnosis, Scientific data 6 (1) (2019) 1–13. doi:10.1038/s41597-019-0206-3.

[57] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al., Natural questions: a benchmark for question answering research, Transactions of the Association for Computational Linguistics 7 (2019) 453–466. doi:10.1162/tacl_a_00276.

[58] M. Joshi, E. Choi, D. S. Weld, L. Zettlemoyer, Triviaqa: A large scale distantly supervised challenge dataset for reading comprehensiondoi:10.18653/v1/P17-1147.

[59] J. Berant, A. Chou, R. Frostig, P. Liang, Semantic parsing on freebase from question-answer pairs, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1533–1544.
URL https://aclanthology.org/D13-1160

[60] P. Baudiš, J. Šedivỳ, Modeling of the question answering task in the yodaqa system, in: International Conference of the cross-language evaluation Forum for European languages, Springer, 2015, pp. 222–228. doi:10.1007/978-3-319-24027-5_20.

[61] F. Sovrano, M. Palmirani, F. Vitali, Deep learning based multi-label text classification of unga resolutions, in: Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance, 2020, pp. 686–695.

[62] J. Brooke, Sus: a retrospective, Journal of usability studies 8 (2) (2013) 29–40.

[63] S. Palan, C. Schitter, Prolific. ac—a subject pool for online experiments, Journal of Behavioral and Experimental Finance 17 (2018) 22–27.

[64] Z. Buçinca, P. Lin, K. Z. Gajos, E. L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 454–464. `doi:10.1145/3377325.3377498`.