Highlights

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

Francesco Sovrano,Fabio Vitali

- Presentation of a model-agnostic and deterministic metric for explainability: DoX.

- DoX is the first explainability metric based on Ordinary Language Philosophy.

- DoX can quantify Carnap's central criteria of explication adequacy.

- Presentation of an open-source software implementation of DoX called DoXpy.

- Evaluation of DoX with two user studies and more than 190 participants.

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

Francesco Sovrano[a,*], Fabio Vitali[a]

[a]*DISI, University of Bologna, Mura Anteo Zamboni 7, Bologna, 40126, Italy*

ARTICLE INFO

ABSTRACT

Explainable AI was born as a pathway to allow humans to explore and understand the inner working of complex systems. However, establishing what *is* an explanation and *objectively* evaluating *explainability* are not trivial tasks. This paper presents a new model-agnostic metric to measure the Degree of Explainability of information in an *objective* way. We exploit 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. To understand whether this metric can measure explainability, we devised a few experiments and user studies involving more than 190 participants, evaluating two realistic systems for *healthcare* and *finance* using famous AI technology, including Artificial Neural Networks and TreeSHAP. The results we obtained are statistically significant (with P values lower than .01), suggesting that our proposed metric for measuring the Degree of Explainability is robust in several scenarios, and it aligns with concrete expectations.

1. Introduction

Recent advances in Artificial Intelligence (AI) enable computer science and engineering to create machines that can learn from rough data, automating tasks previously thought to be accessible only by biological intelligence. However, these advances and results seem to come at a cost in terms of explainability, so the most effective machine learning techniques are, so far, not easily interpretable in symbolic terms [17, 59].

The paradigms that address this explainability 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 [4]. Recently we are seeing a growing demand for explainability in AI applications, motivated by the growing realization that transparency is critical for fairness and legality.

More precisely, in the European Union (EU), we now have several laws in force which establish obligations of explainability based on who uses AI (e.g., public authorities, private companies) and the degree of automation of the decision-making process (e.g., fully or partially automated) [10]. As a result, the EU is indirectly posing an exciting challenge to the Explainable AI (XAI) community by calling for more transparent, user-centered, and accountable *automated decision-making systems* to ensure the explainability of their workings.

In a recent attempt to capture the "legal requirements on explainability in machine learning", Bibal et al. [10] have identified four primary explainability requisites for Business-to-Consumer and Business-to-Business. In particular, Bibal et al. assert that, for Business-to-Consumer and Business-to-Business, explanations about a solely-automated decision-making system should at least provide information about:

- the main features used in a decision taken by the AI;

- all features processed by the AI;

- the specific decision taken by the AI;

- the underlying logical model followed by the AI.

Therefore, with the present paper, we want to expand further the work of Bibal et al., trying to understand whether it is possible to objectively quantify how much of the information required by the law is explained by an AI.

In this paper, we propose a new model-agnostic approach and metric to *objectively* evaluate explainability in a manner 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 answers to many basic questions (*archetypes*), each of which sheds a different light on the concepts being explained. As a consequence, the more (archetypal) answers an *automated decision-making system* can give about the important aspects of its explanandum[1], 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 [66]. Thus, drawing also from Carnap's criteria of adequacy of an explication [50], we frame the Degree of Explainability (DoX) as the average *explanatory illocution* of information

*Corresponding author
✉ francesco.sovrano2@unibo.it (F. Sovrano); fabio.vitali@unibo.it (F. Vitali)
ORCID(s): 0000-0002-6285-1041 (F. Sovrano); 0000-0002-7562-5203 (F. Vitali)

---

[1]The word *explanandum* means "what is to be explained", in Latin.

on a set of *explanandum aspects*[2]. More precisely, we hereby present an algorithm for measuring explainability through pre-trained *deep language models* for general-purpose answer retrieval (e.g., [36, 12]) applied to a particular graph of triplets automatically extracted from text to facilitate this type of information retrieval.

Hence, we made the following hypothesis.

Hypothesis 1. *DoX scores measure explainability: a DoX score can describe explainability, so that, given the same explanandum, a higher DoX implies greater explainability and a lower DoX implies smaller explainability.*

To verify this hypothesis, we devised and implemented a pipeline of algorithms called DoXpy to compute *DoX* scores. We also performed a few experiments to show that *explainability* changes in accordance with varying DoX scores. Notably, the results of all our experiments clearly and undoubtedly showed that Hypothesis 1 holds.

This paper is structured as follows. In Section 2 we give the necessary background information to introduce the theoretical models properly discussed subsequently (i.e., Achinstein's theory), while in Section 3, we discuss existing literature, comparing it to our proposed solution. In Section 4, we show how a metric for quantifying the degree of explainability is possible by defining *explaining* as an illocutionary act of question-answering and by verifying Carnap's criteria employing deep language models. In Section 5, we describe our experiments, discussing the results and some possible limitations of DoX in Section 6. Finally, we point to future work and conclusions in Section 7.

To guarantee the reproducibility of the experiments, we publish the source code[3] of the algorithm for computing DoX scores, as well as the code of the systems used for the experiment, the full details of our user studies and the complete set of data mentioned in this paper.

2. Background

This section provides some background to justify and support the rest of the paper. Hereby we briefly summarise several recent and less recent approaches to the theories of explanation, with a particular 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 profound the connection between answering questions and explaining is in this field.

2.1. Adequacy of Explainability: Carnap's Criteria

In philosophy, the most important work about the criteria of adequacy of *explainable information* is likely to be Carnap's [16]. 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. *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, the *explicatum*[4], and this is precisely what information does when made explainable.

Carnap's main criteria of explication adequacy[16] are *similarity*, *exactness* and *fruitfulness*[5]. *Similarity* means that the explicatum should be *detailed* about the explicandum, in the sense that at least many of the intended uses of the explicandum, brought out in the clarification step, are preserved in the explicatum. On the other hand, *exactness* means that the explication should be embedded in some sufficiently *clear* and exact linguistic framework, while *fruitfulness* implies that the explicatum should be *useful* and usable in a variety of other *good* explanations (the more, the better).

Carnap's adequacy criteria are transversal to all the identified definitions of explainability, possessing preliminary characteristics for any information to be adequately considered 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 also discussed by [29]. An explanation is such regardless of its truth (high-quality but ultimately false explanations exist, especially in science). Vice versa, highly correct information can be inferior at explaining.

2.2. 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 of the act of explaining.

In 1948 Hempel and Oppenheim published their "Studies in the Logic of Explanation" [28], giving birth to what is considered the first theory of explanations: the deductive-nomological model. After that work, many amended, extended, or replaced this model, which came to be considered fatally flawed [13, 60]. Several more modern and competing theories of explanations resulted from this criticism.

Summarising our full analysis [63], the five most important theories of explanation in contemporary philosophy are: Causal Realism, Constructive Empiricism, Ordinary Language Philosophy, Cognitive Science, Naturalism and Scientific Realism. Consequently, there are at least five definitions of "explanation", one per theory. A summary of these definitions is shown in Table 1, highlighting that there is no complete agreement between them on the nature of explanations.

In particular, Hempel's, Salmon's (Causal Realism), and Van Fraassen's (Constructive Empiricism) theories frame the act of explaining more as a *locutionary act* [5], whereby

---

Table 1
Philosophical definitions of explanation and explainable information. In this table, we summarise the definitions of *explanation* and *explainable information* for each one of the identified theories of explanations.

Theory	Explanations	Explainable Information
Causal Realism [60]	Descriptions of causality, expressed as chains of causes and effects.	What can fully describe causality.
Constructive Empiricism [24]	Contrastive information that answers why questions, allowing one to calculate the probability of a particular event relative to a set of (possibly subjective) background assumptions.	What provides answers to contrastive why questions.
Ordinary Language Philosophy [1]	Answers to questions (not just why ones) given with the explicit intent of producing understanding in someone, i.e., the result of an illocutionary act.	What can be used to pertinently answer questions about relevant aspects with *illocutionary force*.
Cognitive Science [31]	Mental representations resulting from a cognitive activity. They are information which fixes failures in someone's mental model.	What can have a *perlocutionary effect*, fixing failures in someone's mental model.
Naturalism and Scientific Realism [61]	Information which increases the coherence of someone's belief system, resulting from an iterative process of confirmation of truths aimed at improving understanding.	What can have a *perlocutionary effect*, increasing coherence of someone's belief system.

an explanation is such because it utters something. Differently, Achinstein's theory (from Ordinary Language Philosophy) explicitly frames explaining as an *illocutionary act* [5] so that an explanation is such because of the intention to explain. The theories of Holland (Cognitive Science) and Sellars (Naturalism/Scientific Realism), on the other hand, frame explaining more as a *perlocutionary act* [5], thus with an explanation being such because of the effects it produces in the interlocutor.

Notably, we notice that whenever explaining is considered to be an act that has to satisfy someone's needs, then explainability differs from explaining. In fact, in this context, pragmatically satisfying someone (i.e., user-centrality) is achieved when explanations are tailored to a specific person so that the same explainable information can be presented and re-elaborated differently across different individuals. It follows that in each philosophical tradition except Salmon's Causal Realism [60], we have a definition of "explainable information" that slightly differs from that of "explanation", as described in [63]. For example, in Ordinary Language Philosophy *explainable information* can be understood as "what can be used to pertinently answer questions about relevant aspects, in an illocutionary way".

2.3. Explainability According to Ordinary Language Philosophy

According to Achinstein's theory, explanations result from an *illocutionary* act of pragmatically answering a question. In particular, it means that there is a subtle and essential 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.

Despite this definition, *illocution* seems too abstract to implement inside an actual software application. Nonetheless, recent efforts towards the automated generation of explanations [64, 66], have shown that it may be possible to define *illocution* in a more "computer-friendly" way. Indeed, as stated in [66], 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. These 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 to a specific aspect of the explanandum. Examples of archetypes are the interrogative particles (e.g.,* why, how, what, who, when, where*), or their derivatives (e.g.,* why not, what for, what if, how much*), or also more complex interrogative formulas (e.g.,* what reason, what cause, what effect*). 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".*

Thus, archetypal questions provide generic explanations on a specific aspect of the explanandum in a given informative context, 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". Some archetypal questions, in this case, are "What is angina?" or "Why a stroke?".

2.4. Explainable AI and Question Answering

Suppose we assume that the interpretation of Achinstein's theory of explanations given by [66] is correct. In that case, 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 [41], and it is also compatible with our intuition of what constitutes an explanation. It is common to many works in the field [57, 42, 46, 25, 21, 73, 55, 34, 44] the use of generic (e.g., why, who, how, when) or more punctual questions to clearly define and describe the characteristics of explainability [41].

For example, Lundberg et al. [43] assert that the local explanations produced by their TreeSHAP (an *additive feature attribution* method for feature importance) may "help human experts understand *why* the model made a specific recommendation for high-risk decisions". On the other hand, Dhurandhar et al. [21] 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*?". Furthermore, Rebanal et al. [55] propose and studies an interactive approach where explaining is defined in terms of answering why, what and how questions. These are just some examples, among many, of how Achinstein's theory of explanations is already implicit in existing XAI literature. They highlight how deep the connection between answering questions and explaining is in this field.

Nonetheless, despite the compatibility, practically none of the works in XAI explicitly mentions any theory from Ordinary Language Philosophy, preferring to refer to Cognitive Science [46, 30] instead. This is probably because Achinstein's illocutionary theory of explanations is challenging to implement into software by being utterly pragmatic. *User-orientedness* is challenging and sometimes not connected to the primary goal of XAI: "opening the black box" (e.g., understanding how and why an opaque AI model works).

3. Related Work

Measuring the quality of explanations and 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 evidence and experiments backing up their claims and none other, usually relying on *ad hoc* or subjective mechanisms for measuring the quality of their explainability. This makes it very hard to perform meaningful comparisons.

In other words, as also suggested by literature reviews (e.g., [71], and especially [63], which reports in Table 2 its main results), it is common to encounter explainability metrics that work only with a specific XAI model or prove their usefulness by collecting human-generated opinions/results after interacting with the studied system and no other.

For example, the metrics proposed in [3, 58, 72, 49, 39, 37] can only be used with specific types of XAI approaches (e.g., prototype selection or feature attribution). Instead, the metrics proposed in [30, 32, 22] rely on user studies, as

many other works [64, 48, 74, 70, 15, 52], based on classical usability metrics (i.e., effectiveness, efficiency, satisfaction).

Only one work among those examined, [30], claims its proposed metric is model-agnostic and thus generic enough to be compatible with any XAI. In particular, this is possible because the work measures explainability *indirectly* by estimating the effects of explanations on human subjects. More precisely, [30] is mainly inspired by the interpretation of explanations given by Cognitive Science, requiring measuring: i) the subjective goodness of explanations; ii) whether users are satisfied by explanations; iii) how well users understand the AI systems; iv) how curiosity motivates the search for explanations; v) whether the user's trust and reliance on the AI are appropriate; vi) how the human-XAI work system performs.

Indeed, the metric presented in [30] is non-deterministic and heavily relies on subjective measurements, despite being model-agnostic. The metric we propose here, DoX, is objective, deterministic, and model-agnostic[7]. It 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 explanations are perceived as satisfactory and sound by the explainees.

Furthermore, only DoX and [39] appear to measure all three main Carnap's desiderata. More specifically, Lakkaraju et al. [39] evaluate Carnap's criteria separately, while with DoX, we propose a single metric that combines all of them.

Finally, as suggested in [63], all existing explainability metrics can be aligned to different interpretations of explainability coming from complementary theories of explanations. As shown in Table 2, most of these metrics seem aligned with Causal Realism and Cognitive Science. In contrast, DoX is the first metric based on Ordinary Language Philosophy.

4. Degrees of Explanation (DoX)

In Section 3, 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 measure the degree of explainability with fully automated software objectively. With this paper, we try to answer this question by leveraging on an extension of Achinstein's theory of explanations as proposed in [66] and summarized in Section 2.3. 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 *explanation*. 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*.

In this section, we discuss the theory behind DoX and a concrete implementation to measure DoX in practice.

---

[6]This table extends a similar one in [63].

[7]DoX is model-agnostic only under the assumption that any explanation or bit of explainable information can be represented or described in natural language, e.g., English.

Table 2
Comparison of Different Explainability Metrics[6]: The column "Sources" points to referenced papers, while column "Metrics" points to the names of the metrics. Elements in bold are column-by-column better than the rest.

Source	Model & Information Format	Closest Supporting Theory	Subject - based	Measured Carnap's Criteria	Metrics
[58]	Rule-based	Causal Realism	**No**	Exactness, Fruitfulness	Performance Difference, Number of Rules, Number of Features, Stability
[72]	Rule-based	Causal Realism	**No**	Similarity, Fruitfulness	Fidelity, Completeness
[49]	Feature Attribution	Causal Realism	**No**	Exactness, Fruitfulness	Monotonicity, Non-sensitivity, Effective Complexity
[39]	Rule-based	Causal Realism	**No**	**Similarity, Exactness, Fruitfulness**	Fidelity, Unambiguity, Interpretability, Interactivity
[32]	**Any**	Causal Realism, Cognitive Science, Naturalism & Co.	Yes	Exactness, Fruitfulness	System Causability Scale
[30]	**Any**	Cognitive Science, Naturalism & Co.	Yes	Exactness, Fruitfulness	Satisfaction, Trust, Mental Models, Curiosity, Performance
[22, 64, 48, 74, 70, 15, 52]	**Any**	Cognitive Science, Naturalism & Co.	Yes	Exactness, Fruitfulness	Usability: Effectiveness, Efficiency, Satisfaction
[3]	Heatmap	Constructive Empiricism	**No**	Similarity, Exactness	Relevance Mass Accuracy, Relevance Rank Accuracy
[37]	Prototype-based	Constructive Empiricism	**No**	Exactness	Proximity, Sparsity, Adequacy (Coverage)
[49]	Prototype-based	Constructive Empiricism	**No**	Similarity, Fruitfulness	Non-Representativeness, Diversity
This Paper	**Any** (Natural Language Text)	Ordinary Language Philosophy	**No**	**Similarity, Exactness, Fruitfulness**	Degree of Explainability

4.1. Quantifying the Degree of Explainability

As discussed in Section 2.4, the informative contents of state-of-the-art XAI are 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". Depending on the needs of the explainees, their background knowledge, the context, and potentially many other factors, each archetype may be equally important.

In other words, depending on the characteristics of the explainee (e.g., background knowledge, objectives, context), a combination of different XAI mechanisms may be necessary to obtain a minimum *understanding of the internal logic of a black-box AI*. Therefore, knowing the types of explainability covered by a system using XAI can be of the utmost importance in understanding how explainable it is. Hence, following this intuition, we started to study how to measure explainability in terms of (generic) questions.

Among the different approaches mentioned in Section 2.2, the closest one to our intuition of explainability is probably Achinstein's theory, coming from Ordinary Language Philosophy. Achinstein defines the act of explaining as an act of illocutionary question-answering, stating that *explaining* is more than *answering a question* because it requires 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 such a concept. For this reason, as discussed in Section 2.3, [64] suggested that illocution (or, better, *explanatory illocution*) is, in fact, the process of answering multiple generic and primitive questions (e.g., why, how, what) called *archetypal questions*.

For example, if someone is asking "How are you doing?", an answer like "I am good" would not be considered an explanation. Differently, the answer "I am happy because I just got a paper accepted at this important venue, and [...]" would instead be normally considered an explanation because it answers other *archetypal questions* together with the main question.

We are convinced that, under these premises, we can concretely measure the degree of explainability of information quantitatively. More precisely, we propose that the degree of explainability of the information depends on the number of *archetypal questions* to which it can adequately

answer. In other words, we estimate the degree of explainability of a piece of information by measuring its relevance to answering a (pre-defined) set of archetypal questions.

Therefore, our theoretical contribution, set out in the following subsections, consists of the precise and formal definition of: *cumulative pertinence*, *explanatory illocution*, *Degree of Explainability (DoX)*, and *average DoX*. We will first provide formal definitions and then explain them further with some examples.

4.1.1. Cumulative Pertinence, Explanatory Illocution and DoX

Assuming the correctness of a given piece of information, explainability is a property of that information. Explainability can be measured in terms of *explanatory illocution*. In order to understand what explanatory illocution is, we have to define the concept of *cumulative pertinence* first.

Definition 2 (Cumulative Pertinence). *The cumulative pertinence is an estimate of how pertinently and how in detail a given piece of information Φ can answer a question about an aspect a of an explanandum Δ. Let A be the set of relevant aspects to be explained about Δ. Let D_a be the subset of all the details (e.g., sentences, grammatical clauses[8], paragraphs) in Φ that are about an aspect $a \in A$. Let q_a be a question about an aspect $a \in A$. Let $p(d, q_a) \in [0, 1]$ be the pertinence of a detail $d \in D_a$ to q_a. Let also t be a pertinence threshold in the $[0, 1]$ range. Then, the cumulative pertinence of D_a to q_a is $P_{D_a,q_a} = \sum_{d \in D_a, p(d,q_a) \geq t} p(d, q_a)$.*

Definition 3 (Explanatory Illocution - Formal Definition). *The explanatory illocution is a set of cumulative pertinences for a pre-defined set of archetypal questions. Let Q be a set of archetypes q and q_a be the question obtained by applying the archetype q to an aspect $a \in A$. Then the explanatory illocution of Φ to an aspect $a \in A$ is the set of tuples $\{\forall q \in Q| < q, P_{D_a,q_a} >\}$[9].*

Consequently, we define DoX as follows.

Definition 4 (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, where $D = \{\forall a \in A, \forall d \in D_a | d\}$, then the DoX is the set $\{\forall q \in Q| < q, R_{D,q,A} >\}$.*

However, DoX alone cannot help in judging whether some collections of information have higher degrees of explainability than others. This is because DoX is a set, and sets are not sortable. Thus we combine the set of pertinence scores composing DoX into a single score representing explainability, called *average DoX*. So, the resulting *average*

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 (Average Degree of Explainability). *The Average DoX is the average of the pertinence of each archetype composing the DoX. In other terms, the Average DoX is $\frac{\sum_{q \in Q} R_{D,q,A}}{|Q|}$.*

The average DoX represents a naive approach to quantify explainability with a single score, as it implies that all the archetypal questions and aspects have the same weight. However, this may not necessarily be true. As suggested by Liao et al. [41], it seems that there is a shared understanding that `why` explanations are the most important in XAI, sometimes followed by `how`, `what for`, `what if` and, possibly, `what`. In other words, the relevance of an explanation can be estimated by the ability to effectively answer the most relevant (archetypal) questions for the stakeholders' objectives. Nonetheless, defining which (archetypal) question is the most relevant is challenging and somewhat subjective. Therefore we believe that average DoX is probably the only objective solution to this dispute.

We will now discuss some examples of applying the formulas mentioned earlier. We will also demonstrate how these formulas can measure Carnap's adequacy criteria.

4.1.2. Interpreting DoX in Terms of Carnap's Criteria

Suppose the sentence "I am happy that my article has been accepted in this prestigious journal" is given as Φ and the set of relevant aspects *{heart, stroke, vessel, disease, angina, symptom}* as A. In this case, the set of details D contains the following details:

- "I am happy";

- "my article has been accepted in this prestigious journal";

- "I am happy that my article has been accepted".

However, none of the details above is about the explanandum. Thus $D_a = \emptyset, \forall a \in A$, because nothing in Φ is related to A. Hence, the average cumulative pertinence would be equal to 0 for every archetype $q \in Q$, forcing the DoX score to be equal to 0, as expected. In other words, no detail of Φ would explain anything about A. Therefore the explainability of Φ for A would be zero.

On the contrary, we would not have a null DoX for A when using the sentence "angina happens when some part of your heart does not get enough oxygen" as Φ. That is because the new Φ contains details about at least two relevant aspects in A: "angina", "heart". Such details would score a higher average cumulative pertinence $R_{D,q,A}$ for q equal to `why` because they are about causality.

Eventually, when computing the DoX of the new Φ for this set of explanandum aspects A with the DoXpy algorithm presented in Section 4.2[10], the *average DoX* is 0.29. In

---

[8]A typical *clause* consists of a subject and a syntactic predicate, the latter typically a verb phrase composed of a verb with any objects and other modifiers.

[9]The operator $< x, y >$ is used here to represent tuples.

[10]When using the MiniLM pertinence estimator introduced in Section 4.2.2.

particular, as expected, the archetypes with the best score are the ones related to causality (i.e., what effect has a score of 0.59; in what case, why and how have a score of 0.57). In contrast, most of the other archetypes have a null score (i.e., who, when).

Given Definition 4, we can say that DoX is an estimate of the *fruitfulness* of D that combines in one single score the *similarity* of D to A and the *exactness* of D for Q. For these reasons, DoX is akin to Carnap's *central* criteria of adequacy of explanation (introduced in Section 2.1). Although, differently from Carnap, our understanding of *exactness* is not that of adherence to standards of formal concept formation[11] [14], but rather that of being precise or pertinent enough as an answer to a given question.

The number of relevant explanandum aspects covered by a given piece of information, and the number of details that are pertinent about it (i.e., $|\{\forall a \in A, \forall d \in D_a | d\}|$), roughly say how much *similar* that information is to the explanandum. More precisely, the formula used for computing the cumulative pertinence P_{D_a,q_a} sums the contribution of every single detail according to its pertinence to the aspects $a \in A$, telling us how much D_a is similar to a. Thus, if *pertinence* $p(d, q_a)$ would close to zero for all archetypes $q \in Q$, then a detail d would have nothing to do with an aspect a. Furthermore, the average cumulative pertinence $R_{D,q,A}$ contains information about the *exactness* of multiple answers, aggregating pertinence scores. As a result, by measuring $R_{D,q,A}$ for all the $q \in Q$, we also obtain an estimate of how D is *fruitful* for the formulation of many other different explanations intended as the result of an illocutionary act of pragmatically answering questions.

This construction of DoX in terms of Carnap's main criteria (cf. Section 2.1) of adequacy is crucial because it allows implementing an actual algorithm to quantify explainability as shown by the experimental results presented in Section 5.

4.2. The DoXpy Algorithm

Throughout this section, we will explain why and how to use existing algorithms for answer retrieval and information extraction to implement DoXpy, an algorithm for computing DoX. We publish the DoXpy source code at https://github.com/Francesco-Sovrano/DoXpy for reproducibility purposes.

Given Definition 4, we argue that it is possible to write an algorithm that can approximately quantify the Degree of Explainability of information representable with *natural language* (e.g., English) by adapting existing technology for question-answering. In particular, according to Definition 3, in order to implement an algorithm capable of computing the (average) DoX of Φ, we need to:

- define a set A of *explanandum aspects*;

- identify the set of all possible archetypes Q;

- define a mechanism to identify the set D of details contained in Φ and the subset D_a for every $a \in A$;

- define the question-answering process: the function p to compute the pertinence of an individual detail d to an archetypal question q_a.

Interestingly, the set of aspects A is task-dependent and must be defined for each explanandum (e.g., manually listing all aspects or automatically extracting the list of aspects from a textual description of the explanation with a tokenizer). Instead, the set of archetypes Q, the pertinence function p, and the mechanism for extracting D and D_a from Φ *may always be the same* for all explananda. Specifically, the set A of *explanandum aspects* is a collection of (lemmatized) words, and it can be different from the set I of aspects explained by Φ. What is of utmost importance for a Φ to be a good *explanandum support material* is that $A \subseteq I$.

4.2.1. Details Extraction and Pertinence Estimation

Definition 4 requires a mechanism to identify the set D of details contained in the *explanandum support material* Φ, as well as a mechanism to identify the sub-sets $D_a \subseteq D$ for every $a \in A$.

A detail d is a snippet of text with some specific characteristics, also called *information unit*. It is a relatively small sequence of words about one or more aspects (i.e., a subset of I) that is usually extracted from a more complex information bundle (i.e., a paragraph, a sentence). In other terms, these details should carry enough information to describe different parts of an aspect (possibly connected to many other aspects). So we can use them to answer some (archetypal) questions about an $a \in A$ and to correctly estimate a *level of detail*, as required by Definition 4.

Considering the characteristics of D and I mentioned above, their most natural representation is a (knowledge) graph. A graph is a set of nodes (i.e., I) connected by a set of edges (i.e., D). Therefore, we believe that the simplest way to identify the set of details D may be to extract a graph of *information units* from Φ on which efficient question-answering could be performed.

The task of answering questions using an extensive collection of documents about diverse topics or from different domains is called open-domain question-answering [18, 33]. There are at least three main software architectures for open-domain question-answering: the retriever-reader architecture, the retriever-generator, and the generator-only architecture. The first two architectures combine information retrieval techniques and neural reading comprehension or text generation models. In particular, the latter does not involve classical information retrieval, thus being completely end-to-end. A famous example of generator-only architecture could be OpenAI's ChatGPT, an adaptive and intelligent dialogue system. This type of question-answering algorithm usually relies on huge (i.e., with hundreds of billions of parameters) deep neural networks trained in an unsupervised manner to memorize facts and in a supervised manner to answer questions in a meaningful and coherent way. Even though generator-only architectures are capable of impressive results, they tend to write plausible-sounding but incorrect or nonsensical answers. One of the reasons for

---

[11]Actually, Carnap did not specify what he means by "exactness". Regardless, in this context, "exactness" is often viewed as either lack of vagueness or adherence to standards of formal concept formation.

this problem is that this type of architecture is fully end-to-end and needs to perform fact-checking.

On the other hand, the retriever-generator and retriever-reader architectures circumvent the latter problem by relying on a system capable of retrieving plausible answers from a knowledge base (or graph) of verified contents. The retriever-reader and retriever-generator models usually have an asymptotic time complexity that grows linearly with the number of answers considered for retrieval. That number does not necessarily has to be equal to the number of all the retrievable texts. In other words, the time complexity of the answer retrieval system can be intelligently controlled by making it fit the memory and time constraints of a personal computer, e.g., by filtering out all texts unrelated to a question. In particular, the retriever-generator rewrites and reprocesses the retrieved information, while the retriever-reader limits itself to extracting it (as it is) and reclassifying it properly. While the retriever-generator may still slightly suffer from the problem of generating hallucinated answers, the retriever-reader does not, at the cost of producing less cohesive answers.

For example, a retriever-reader like the one used in [64, 62] for (archetypal) question-answering could be suitable for DoXpy, allowing the identification of meaningful *information units* and also suggesting a mechanism for estimating *pertinence* by extracting from Φ a graph of D and I designed for answer retrieval. Indeed, the aforementioned answer retrieval algorithm consists of a pipeline of AI tools for the extraction from Φ of a graph of D and I specifically designed to measure the pertinence of D to a set of (archetypal) questions Q on A.

More specifically, this graph is extracted by detecting, with a dependency parser, all the clauses within the *explanandum support material* that stand as an edge of the graph. In practice, these clauses are represented as special triplets of subjects, templates, and objects called template-triplets. Specifically, the templates are composed of the ordered sequence of tokens connecting a subject and an object. On the other hand, the subject and the object are represented in these templates by the placeholders *"{subj}"* and *"{obj}"*. An example of template-triple is:

- Subject: *"angina pectoris"*

- Template: *"In particular, {subj} happens when some part of your heart does not get enough {obj}."*
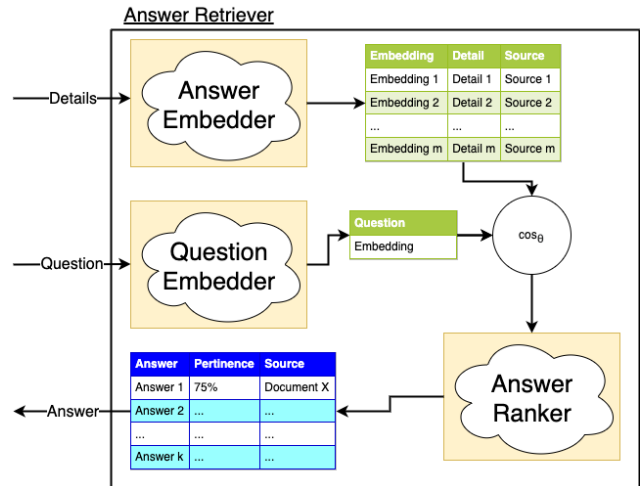
- Object: *"oxygen"*

Hence, the resulting template-triplets are a sort of function where the predicate is the body, and the object and subject are the parameters. Obtaining a natural language representation (i.e., a detail $d \in D$) of these template-triplets is straightforward by design by replacing the instances of the parameters in the body. This natural representation is then used as a possible answer for retrieval by measuring the (cosine) similarity (or pertinence p) between its embedding (obtained through deep language models such as [26, 36]) and the embedding of a question q.

Notably, as *information units*, Sovrano and Vitali [64, 62] use grammatical clauses (meaningful decompositions of grammatical dependency trees) to ensure that the units represent the smallest granularity of information.

As a consequence, using this type of *information units* for DoX guarantees:

- a disentanglement of complex information bundles into the most simple units, to correctly estimate the *level of detail* covered by the information pieces, as per Definition 4;

- a better identification of duplicated units scattered throughout the information pieces, to avoid an over-estimation of the *level of detail*.

All these properties satisfy the requirements that a good detail $d \in D$ should possess for generating a DoX score. This motivates our decision to use the answer retrieval algorithm of [64, 62] as the main component of the DoXpy pipeline.
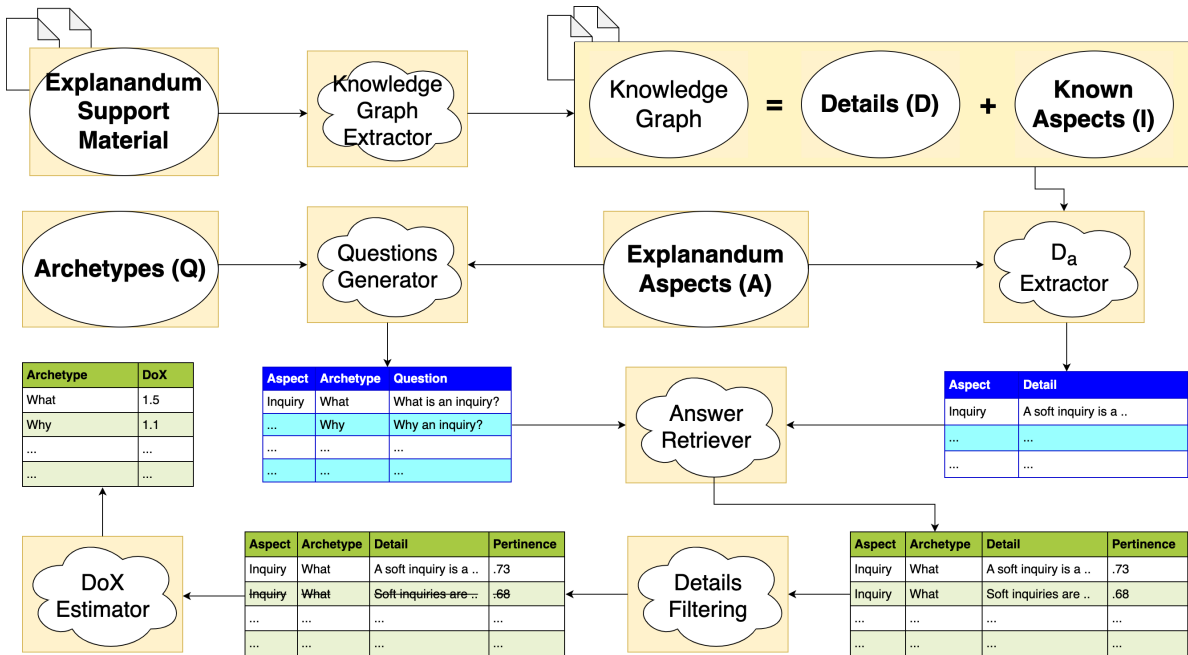

Figure 1: Flow Diagram of the Answer Retriever used in the pipeline of DoXpy. This figure summarises the answer retrieval algorithm used by DoXpy. A question q_a is given as input to the retriever together with a set of details D_a. Then, the details in D_a and q_a are embedded, and their cosine similarity (θ) is used to rank details according to their pertinence to the question.

As shown in Figure 1, the retriever-reader used by DoXpy relies on mechanisms for embedding questions and answers in dense numerical representations so that the cosine similarity between the embedding of a question and an answer is a measure of the latter's relevance to the former.

In particular, these embeddings can be obtained (for example) through deep neural networks (i.e., the pertinence function p) specialized in answer retrieval and pre-trained on ordinary English to associate similar vectorial representations to a question and its correct answers. Examples of these pre-trained deep language models are discussed in the following subsection.

More specifically, let a be the explanandum aspect of a question q_a, and $m = < s, t, o >$ be a template-triplet,

Figure 2: DoXpy Pipeline. The pipeline starts with extracting a graph from the *explanandum support material* Φ 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 this, we use some deep language models for answer retrieval.

and $d = t(s, o)$ be the natural language representation of *m* also called *information unit*, and *z* the context (i.e., a paragraph, a sentence) from which *m* was extracted. DoXpy performs answer retrieval by retrieving the set D_a of all the template-triplets about *a* and selecting amongst the natural language representations *d* of the retrieved template-triplets that are likely to be an answer to q_a. The probability that *d* pertinently answers q_a can be estimated as the similarity between the embedding of $d + z$ (i.e., *d* concatenated with *z*) and the embedding of q_a. So that if $d + z$ is similar enough to q_a, then *z* is said to be an answer to q_a for *information unit d*. Therefore, in practice, the algorithm can retrieve an unbounded number of details (i.e., answers).

In particular, a detail is said to be redundant (i.e., duplicated) whenever it contains information that answers an archetypal question $q_a \in Q$ in a manner too similar to that of other (more pertinent) details. For example, the detail "*P is the probability of having a heart disease*" is different but similar to "*the score P is the probability of having a disease*". However, the former detail is more precise (it speaks of *heart disease* instead of generic diseases) and relevant than the latter in answering the archetypal question "*What is probability P?*". Therefore, the second detail must be discarded as redundant to prevent DoXpy from considering two details that express the same information differently. To do this, DoXpy uses the same deep neural networks used for retrieval to compute the similarity between two answers, discarding those with the lowest relevance scores that share a similarity greater than a threshold *r*.

Consequently, as shown in Figure 2, the pipeline of DoXpy consists of the following four steps. First, a knowledge graph is extracted from the explanandum support material Φ using the algorithm described in [62], thus defining the set of details *D* and the set of known aspects *I*. Secondly, a given set of explanandum aspects *A* and archetypes *Q* is used to generate a set of questions q_a for each $a \in A$ and $q \in Q$ and to identify all $D_a \subseteq D$. Third, the answer retriever of [64, 62] is used to associate a pertinence score with each $d \in D_a$ for each q_a and (importantly) to identify and filter out duplicate answers. Fourth, the formulas in Section 4.1 are used to aggregate the relevance scores and estimate the (average) DoX without considering duplicate details.

4.2.2. Pertinence Functions and Thresholds

According to Definition 4, we need to define a pertinence function *p* and pick a threshold *t* to compute the DoX. As previously discussed, we will use as pertinence function *p* a deep neural network for answer retrieval. The point is that many different deep neural networks exist for this task, i.e., [26, 62, 36], 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?

To answer these questions, we decided to study the behavior of more than one deep language model as pertinence function *p*. Assuming that these models get good results on state-of-the-art benchmarks for *pertinence estimation*, we believe that the results of the computation of DoX should be consistent across them. Hence the models we considered are:

- MiniLM: published by [36, 56] and trained on Natural Questions [38], TriviaQA [35], WebQuestions [8], and CuratedTREC [7].

- Multilingual Universal Sentence Encoder: published by [76] and trained on the Stanford Natural Language Inference corpus [12].

We experimentally found on the two systems presented in Section 5.1 that for both the aforementioned language models, a good pertinence threshold can be $t = 0.15$.

4.2.3. Archetypal Questions

According to Definition 4, we need to define a set of archetypal questions Q to compute the explanatory illocution of a snippet of text correctly. According to Definition 1, an archetypal question is a generic question characterized by one or more interrogative formulas. Casting the semantic annotations of individual propositions as narrating an archetypal question-answer pair recently gained increasing attention in computational linguistics [27, 23, 45, 54], especially in *discourse theory* and the *theory of sentential meaning representations*.

On the one hand, *discourse theory* is a branch of linguistics that studies how coherence and cohesion make up a text to form a discourse. So that discourse is said to be coherent if all of its pieces belong together, while it is said to be cohesive if its elements have some common thread. In recent years, many different discourse models have been spelled out, each with different pros and cons. Amongst them, we cite the model of the Penn Discourse TreeBank (PDTB for short) [47, 53, 75] because it is considered one of the most generic models of discourse. In fact, with little or no change in the model, PDTB appears to be usable for representing discourses of natural languages belonging to different families [77], e.g., Chinese, Arabic, Hindi.

The central assumption behind PDTB is that "the meaning and coherence of a discourse result partly from how its constituents relate to each other". Specifically, these relations between constituents, called discourse relations, are defined as semantic relations between abstract objects, called elementary discourse units, connected by implicit or explicit connectives, e.g., "but", "then", "for example", "although". In PDTB, elementary discourse units are spans of text denoting a single event serving as a complete and distinct unit of information that the surrounding discourse may connect to [69]. What is of interest to us is that according to Pyatkin et al. [54], all discourse connectives can be represented as questions: `in what manner`, `what is the reason`, `what is the result`, `after what`, `what is an example`, `while what`, `in what case`, `since when`, `what is contrasted with`, `before what`, `despite what`, `what is an alternative`, `unless what`, `instead of what`, `what is similar`, `except when`, `until when`.

On the other hand, the *theories of sentential meaning representation* are grammatical theories that study the relationships between predicates and arguments in a sentence. In particular, predicate-argument relationships support answering basic questions such as *who did what to whom*, and they

can be captured with models to separate a sentence's meaning from its syntactic representation. Amongst these models, we mention the theory of abstract meaning representations [6, 40], which can be used to represent whole sentences as (directed and acyclic) graphs of predicates and arguments that can be exploited for tasks such as machine translation (e.g., the conversion of sentences into symbolic knowledge representations, for example, a piece of software written in Prolog that can be used for inference by an automated reasoner), natural language generation and understanding. In particular, according to Michael et al., [45], all the abstract meaning representations can be encoded as pairs of questions and answers involving the following *archetypes*: `what`, `who`, `how`, `where`, `when`, `which`, `whose`, `why`.

Interestingly, it is possible to identify a hierarchy or taxonomy of these archetypes, ordered by their intrinsic level of generality or specificity. For example, the simplest interrogative formulas, such as those used by the theory of abstract meaning representations, can be seen as the most generic archetypes since they consist of only one interrogative particle: `what`, `why`, `when`, `who`, etc. We will refer to these archetypes as the *primary* ones. On the other hand, the archetypes used by the PDTB model (e.g., `what is the reason`, `what is the result`) are more complex and specific, building over the primary archetypes. For this reason, we will refer to them as *secondary* archetypes.

Even though many more archetypes could be devised (e.g., `where to` or `who by`), we believe that the list of questions we provided earlier is already rich enough to be generally representative of any other question, whereas more specific questions can always be framed by using the interrogative particles we considered (e.g., `why`, `what`). *Primary archetypes* can be used to represent any fact and abstract meaning [11]. In contrast, the *secondary archetypes* can cover all the discourse relations between them (at least according to the PDTB theory).

5. Evaluation of DoXpy: Experiments and Results

In Section 4.1 we argued that the degree of explainability of any collection of text (e.g., the output of an XAI) could 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 systems making use of XAI (also called XAI-based systems). In particular, with the first experiment, we measure explainability *directly*, while with the second, we perform *indirect* measurements obtained through 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 undoubtedly 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. Thus, by masking the output of the XAI, the overall system can be forced to be not explainable enough. This characteristic can be used to partially verify Hypothesis 1, but not in a generic way because this type of verification is based on a comparison with a total lack of explainability and not with different degrees of it.

This is why we decided to measure explainability also *indirectly* with a second experiment, to understand whether DoX correlates with the expected effects of explainability on human subjects. In other terms, we have to compare DoX to existing metrics for explainability based on Cognitive Science (e.g., usability, effectiveness) as shown in Table 2.

If Hypothesis 1 is correct, the lower the DoX score, the fewer explanations can be extracted, and the less effective (as per ISO 9241-210) an explainee is likely to be in achieving explanatory goals that are not covered by the explanations. More specifically, *effectiveness* here is defined as "accuracy and completeness with which users achieve specified goals". In our case, it is measured through multiple-choice domain-specific quizzes.

We expect an increment in DoX always corresponds to an increment in effectiveness, at least on those tasks covered by the information provided by the increment of DoX. To show this, we borrowed the results of two independent user studies [64, 65], observing how DoX correlates with the effectiveness scores measured by these studies.

5.1. XAI Systems Considered for the Experiments

The two systems making use of XAI that we considered as case studies are:

- a heart disease predictor based on XGBoost [19] and TreeSHAP [43], concerning healthcare;

- a credit approval system based on a simple Artificial Neural Network and CEM [21], concerning finance.

Both these systems are an example of normal *XAI-based explainer*, a one-size-fits-all explanatory mechanism providing the bare output of the XAI as a 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: the Credit Approval System

The credit approval system is the same also used by [64, 66], and it has been designed by IBM to showcase AIX360[12]. In particular, this credit approval system uses an artificial neural network to predict a customer's credit risk (and thus decide whether to approve a loan) together with an XAI (called CEM [21]) to provide post-hoc (static) explanations of the neural network's predictions. These explanations aim at helping customers understand whether they have been treated fairly, providing insights into ways to improve their qualifications so as the likelihood of future acceptance can be increased.

A typical use case of this system is the following. A customer (e.g., John) applies for a loan from the bank. The bank collects sufficient information about the customer. It transmits it to the artificial neural network, which uses it to work out how likely the customer is to repay the loan. If the customer's credit risk is low, the loan application is approved, but if the credit risk is too high, the system uses the CEM to explain why.

The artificial neural network behind this credit approval system is trained on the "FICO HELOC" dataset[13], containing anonymized information about loan applications made by real homeowners, to answer the following question: "What is the decision on the loan request of applicant X?".

Given the specific characteristics of this credit approval system, we assume that its users' main goal is to understand the causes behind a loan rejection and what to do to get a loan accepted. This is why CEM is deployed to answer the following questions:

- What are the factors to consider to change the result of the application of applicant X?

- How should factor F be modified to change the result of the application of applicant X?

- What is the relative importance of factor F in changing the result of the application of applicant X?

Nonetheless, many other relevant questions might be answered before the user is satisfied and reaches his/her objective. These questions may be: "How to perform those minimal actions?", "Why are these actions so important?", etc.

Indeed, interpreting the internal parameters and complex calculations of an AI model such as this credit approval system is complicated. For example, a layperson trying to obtain a loan might undoubtedly be interested to know that her/his application was rejected (by the AI) mainly due to a high number of credit inquiries on his/her accounts (as CEM can tell). However, this information alone might not be sufficient to achieve her/his goals. These objectives may be beyond the reach of the AI, such as understanding how to effectively reduce the number of inquiries to obtain the loan, what type of credit inquiries may affect his status and the difference between a hard and a soft inquiry.

To summarise, the output of the credit approval system is composed by:

- Context: a titled heading section kindly introducing the user to the system.

- AI Output: the decision taken by the artificial neural network for the loan application (i.e., "denied" or "accepted").

- XAI Output: a section showing the output of the CEM. This output consists of a minimally ordered list of factors deemed to be the most important to change

---

[12]https://aix360.mybluemix.net/explanation_cust

[13]https://fico.force.com/FICOCommunity/s/explainable-machine-learning-challenge?tabset-3158a=a4c37
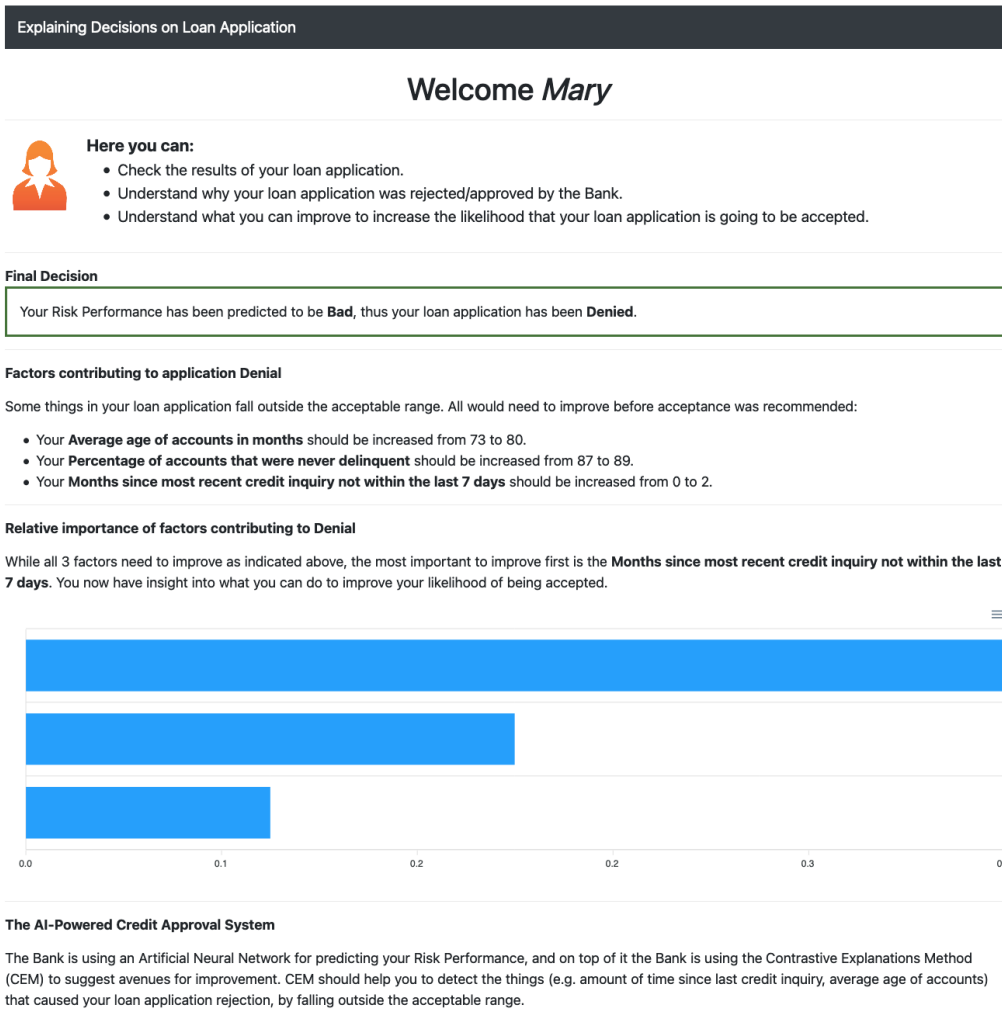
Figure 3: Screenshot of the credit approval system.

for the outcome of the artificial neural network to be different.

A screenshot of a web application implementing this credit approval system is shown in Figure 3.

5.1.2. Healthcare: the Heart Disease Predictor

Similarly to the credit approval system, the heart disease predictor comes from [66]. In particular, the explanandum of the heart disease predictor is about health, and the system is used by a first-level responder of a help desk for heart disease prevention. More specifically, a first-level responder is responsible for handling the requests for assistance of a patient, forwarding them to the correct 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. So, they quickly use the heart disease predictor to determine what to answer the callers and the subsequent 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 responders' goal is to answer the questions of a caller in the most efficient and effective way.

The considered heart disease predictor uses an AI called XGBoost [19] 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 three different risk areas: low (probability p of heart disease below 0.25), medium ($0.25 < p < 0.75$), or high. Therefore, XGBoost is used to answer the following questions:

- How likely is it that patient X has heart disease?

- What is the risk of heart disease for patient X?

- What is the recommended action for patient X to treat or prevent heart disease?

In particular, the dataset used to train XGBoost is the "UCI Heart Disease Data" [20, 2].

On top of XGBoost, the heart disease predictor uses TreeSHAP [43], a famous XAI algorithm specialized in tree ensemble models (e.g., XGBoost) for post-hoc explanations. In particular, TreeSHAP is used to understand the contribution of each input feature to the output of XGBoost. Therefore, TreeSHAP is used to answer the following questions:

- What would happen if patient X had 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?

However, many other important questions should be answered. These include "What is the easiest thing the patient could 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 from raise?".

Finally, to summarise, 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 entering the patient's biological parameters.

- AI Outputs: a section displaying the likelihood of heart disease estimated by XGBoost and a few generic suggestions about the next actions to take.

- XAI Outputs: a section showing each biological parameter's contribution (positive or negative) to the likelihood of heart disease generated by TreeSHAP.

A screenshot of a web application implementing this heart disease predictor is presented in Figure 4.

5.2. 1st Experiment: Direct Evaluation on Normal XAI-generated Explanations

In order to verify Hypothesis 1, we have to show that there is a strong correlation between DoX and the perceived amount of *explainability*. To this end, we devised two experiments.

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 because their amount of *explainability* is, by design, clearly and explicitly dependent on the output of the underlying XAI. So, by masking the output of the XAI, the overall system can be forced to be less explainable. Hence, this characteristic can be exploited to verify the hypothesis in a straightforward but effective way.

In other words, a XAI-based system is composed of a black-box AI system wrapped by a XAI. So, with this experiment, we compare the DoX of a normal XAI-based

explainer with that of the same system without the XAI, also called *normal AI-based explainer*. As a result, we expect the (average) DoX of the XAI-based explainer to be higher than its wrapped AI.

For this experiment, we used the XAI-based systems defined in Section 5.1. Therefore, by simply removing the output of the XAI (respectively CEM and TreeSHAP) from these systems, it is possible to obtain the *normal AI-based explainers* we need.

In order to compute the (average) DoX of these systems, we take as a set of *explanandum aspects* those targeted by the credit approval system and the heart disease predictor. More precisely, the main *explanandum aspects A* targeted by XGBoost [19] and TreeSHAP [43] in the heart disease predictor are five:

- the recommended action for patient X;

- the most important factors Y that contribute to predicting 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 A* targeted by the Artificial Neural Network and CEM [21] in the credit approval system are four:

- the factors F to consider for changing the result;

- the relative importance of factors F in changing the result;

- the risk performance of applicant X;

- the result of the application of applicant X.

Eventually, after properly converting the images produced by the *XAI-based explainers* to textual explanations, the resulting *explanandum aspects coverage* (i.e., the ratio of $|A \cap I|$ to $|A|$) of both the heart disease predictor and the credit approval system is 100%. In contrast, that of their *AI-based explainers* is 48% and 43% respectively.

By calculating the DoX through DoXpy, we obtained the results shown in Table 3. As expected, for both the heart disease predictor and the credit approval system, the experiment results indicate that the (average) DoX of all XAI-based explainers is significantly higher than that of AI-based explainers, regardless of the *deep language model* adopted. Although, we can see that MiniLM and the Universal Sentence Encoder (the two adopted *language models*) produce comparable but slightly different DoX scores, suggesting that the choice of the pertinence function p could sensibly impact the value of DoX.

---

[14] The numerical values in this table are different from those reported in [67] because we used DoXpy v3.0 instead, which includes several improvements in the information retrieval algorithm that prevent details duplication, as described in Section 4.2.

Figure 4: Screenshot of the heart disease predictor.

Considering that in this first experiment, we arbitrarily chose a simple set of *explanandum aspects*, what would happen if we considered different and more complex explananda and explanatory contents? Furthermore, the result of this experiment is b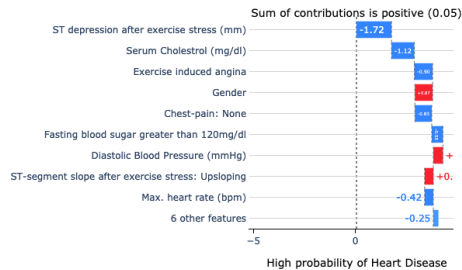ased on comparing the DoX of an unexplained system (i.e., the normal *AI-based explainers*) with that of a more explainable system, and this is an exceptional and naive case to consider. Therefore, to thoroughly test Hypothesis 1, we must understand whether DoX behaves as expected even when explainability is present in different and non-zero quantities. To this end, explainability can be measured *indirectly* by studying the effectiveness of the resulting explanations on human subjects, as shown in Section 5.3.

5.3. 2nd Experiment: A Study of the Effects of Explainability on Human Subjects

The second experiment aims to show whether there is a correlation between DoX and the effects of explainability on human subjects. We have that a higher explainability implies a greater capacity to explain, hence a greater number of explanations. In short, the lower the DoX, the fewer explanations can be produced, and the less effective the explainer is in explanandum-related tasks.

So, if Hypothesis 1 were true, an increase in the DoX of the (explanatory) system would always correspond to a proportional increase of its effectiveness, at least on those tasks covered by the information provided by the increment of DoX. Therefore, to verify this point, we borrowed two

Table 3
Results of the 1st Experiment on DoXpy[14]. In this table, DoX and average (Avg) DoX are shown for the credit approval system (CA) and the heart disease predictor (HD). As columns, we have the normal AI-based explainers (AI, for short) and the normal XAI-based explainers (XAI, for short). As rows, we have different explainability estimates using MiniLM (ML) and the Universal Sentence Encoder (TF). For simplicity, for DoX, we show only the *primary archetypes*.

		CA		HD	
		AI	XAI	AI	XAI
Avg DoX	ML	0.46	1.52	0.53	1.49
	TF	0.23	0.86	0.27	0.84
DoX	ML	"what": 0.49 "how": 0.48 "which": 0.47 "who": 0.47 "why": 0.46 "whose": 0.45 "when": 0.45 "where": 0.44	"which": 1.61 "how": 1.60 "what": 1.59 "why": 1.53 "when": 1.51 "where": 1.46 "who": 1.45 "whose": 1.41	"why": 0.60 "which": 0.55 "what": 0.54 "how": 0.54 "whose": 0.49 "when": 0.47 "where": 0.47 "who": 0.46	"why": 1.63 "which": 1.60 "what": 1.52 "how": 1.52 "whose": 1.51 "who": 1.40 "when": 1.38 "where": 1.38
	TF	"what": 0.26 "when": 0.24 "which": 0.22 "how": 0.19 "where": 0.18 "who": 0.18 "why": 0.16 "whose": 0.15	"what": 0.94 "when": 0.87 "which": 0.77 "where": 0.73 "how": 0.73 "who": 0.68 "why": 0.64 "whose": 0.55	"what": 0.30 "when": 0.25 "which": 0.23 "who": 0.22 "how": 0.22 "where": 0.22 "why": 0.22 "whose": 0.18	"what": 0.97 "when": 0.78 "how": 0.74 "which": 0.72 "who": 0.68 "where": 0.68 "why": 0.66 "whose": 0.56

user studies published by [64] and [65], involving more than 190 human subjects.

Notably, these user studies considered the same explanandum support materials and AI systems used during the first experiment and described throughout Section 5.1, analyzing the effectiveness of explanations given by other explainers when changing the *explanandum support material* and the way explanations are presented to the explainee.

The effectiveness of explanations was measured by giving the explanations to the participants of the user studies and asking them questions to see whether the given explanations helped them to understand the explananda. In particular, two domain-specific multiple-choice quizzes (one per explanandum) were used to measure effectiveness, each consisting of questions representing plausible information goals for the system's users. Being impossible and unfeasible to identify all the possible questions a real user would ask to reach its goals, only a few representative questions were considered for the sake of the study. It appears from preliminary studies, such as the one by Liao et al. [41], that users are interested in asking a variety of different questions about an AI-based system, pointing to complex and heterogeneous needs for explainability that go beyond the output of a single XAI.

5.3.1. 1st User Study

On the one hand, the first user study comes from [64], where a novel mechanism, called *overview-based explainer*, is used to explain extensive collections of heterogeneous documents (i.e., more than 50 web pages) about the credit approval system, in a user-centered and interactive way. This is done by organizing knowledge as a graph of explanandum aspects whose related explanations are ordered by relevance and simplicity according to a set of pre-defined archetypal

Table 4
Quiz of the Credit Approval System. This table contains the quiz used for evaluating the effectiveness of tools explaining the credit approval system. In this table, XAI is the normal *XAI-based explainer* (i.e., the webpage shown in Figure 3) and "OBE" is the overview-based explainer. Column *Steps* indicates the minimum number of steps (in terms of links to click, overviews to open, or questions to pose) required by each explanatory tool to provide the correct answer. Negative *steps* means that the correct answer cannot be found. In contrast, 0 *steps* means that the answer is immediately available in the initial explanans (i.e., the content of the webpage shown in Figure 3). Column "Archetype" indicates which interrogative particles represent the question. Many questions are polyvalent in that they can be rewritten using different archetypes.

Question	Archetype	Steps XAI	Steps OBE
What did the credit approval system decide for Mary's application?	what, how	0	0
What is an inquiry (in this context)?	what	-1	1
What type of inquiries can affect Mary's score, the hard or the soft ones?	what, how	-1	1
What is an example of hard inquiry?	what	-1	1
How can an account become delinquent?	how, why	-1	1
Which specific process was used by the Bank to automatically decide whether to assign the loan?	what, how	0	0
What are the known issues of the specific technology used by the Bank (to automatically predict Mary's risk performance and to suggest avenues for improvement)?	what, why	-1	1

questions (e.g., what, how, when, why). In particular, a user can carry out *overviewing* from the initial explanation shown in Figure 3 by clicking on the annotated words for which an explanation is needed. An example of *overview* is shown in Figure 5.

The external resources used by the *overview-based explainer* for the credit approval system consist of 58 webpages, 50 of which come from the website of MyFICO[15], while the remaining come from Forbes[16], Wikipedia, AIX360[17], and BankRate[18].

This first user study compares the effectiveness scores of the credit approval system with and without the possibility for users to perform *overviewing* to demonstrate that the *overview-based explainer* generates more effective explanations than the baseline. Specifically, effectiveness scores are generated by users interacting with the system and answering a multiple-choice quiz (shown in Table 4) on the credit approval system. In particular, each question of the multiple choice quiz has 4 to 8 plausible answers, of which only one is (the most) correct. At the end of the quiz, answers are automatically scored as correct (score 1) or not (score 0), and the resulting scores are added together to
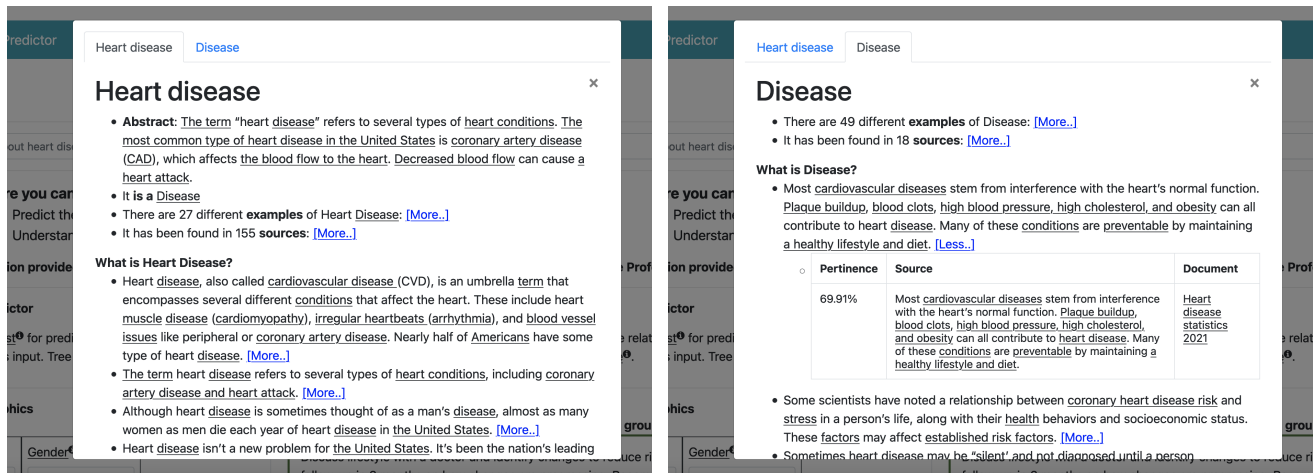
---

[15] https://www.myfico.com
[16] https://www.forbes.com
[17] http://aix360.mybluemix.net
[18] https://www.bankrate.com

Figure 5: Example of Overview: this figure shows an example of interactive *overview* displaying relevant information about important concepts for the heart disease predictor. Clicking on any underlined word would open a new *overview* in a new tab, as shown. Furthermore, every given answer is linked to its source document.

form the effectiveness score. For example, for the question "What did the Credit Approval System decide for Mary's application?", the correct answer is "It was rejected", and the wrong answers are "Nothing" or "I do not know".

For this user study, 103 participants were recruited (57 males, 44 females, and two unknowns, ages 18-55) on the online platform Prolific [51]. All the participants were recruited among those who: 1. are resident in UK, US, or Ireland; 2. have a Prolific acceptance rate greater or equal to 75%[19]. Participants were randomly assigned to use the credit approval system with or without *overview-based explainer* in a between-subjects test. The credit approval system without *overview-based explainer* is also called normal *XAI-based explainer* because it only explains through the output of an XAI.

In the end, 51 participants evaluated the normal *XAI-based explainer*, and 52 evaluated the *overview-based explainer*. For more details about this user study, read [64].

5.3.2. 2nd User Study

The second user study comes instead from [65] and concerns the heart disease predictor. Similar to the first, this study compares the effectiveness of an *overview-based explainer* called YAI4Hu and that of a normal *XAI-based explainer*. In addition, this second study also investigates the effectiveness of two other explainers: a *two-level explainer* and a *how-why explainer*.

The *two-level explainer* is static, as the normal *XAI-based explainer*. It is made of the output of the XAI (shown in Figure 4) directly connected to a second (non-expandable) layer of information consisting of an exhaustive and verbose set of autonomous static explanatory resources. The *two-level explainer* 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 *how-why explainer* is like the *overview-based explainer*, but it uses only the archetypes why and how for generating explanations. Furthermore, also YAI4Hu is an extension of the *overview-based explainer* that instead adds a mechanism (called *open-ended questioning*) for users to ask their questions to the system. More specifically, *open-ended questioning* can be performed by asking questions in English through a search box that uses the graph-based answer retrieval mechanism described in Section 4.2. Importantly, YAI4Hu, the *how-why explainer* and the *two-level explainer* share the same *explanandum support material*.

Such *explanandum support material* is composed by the contents shown in Figure 4 and a set of external resources carefully selected to cover the topics of the heart disease predictor that consists of 103 webpages, 75 of which come from the website of the U.S. Centers for Disease Control and Prevention[20], while the remaining from the American Heart Association[21], Wikipedia, MedlinePlus[22], MedicalNewsToday[23] and other minor sources.

For this second user study, 89 different participants were recruited amongst the university students of the following courses of study[24]: bachelor's degree in computer science; bachelor's degree in management for informatics; master's degree in digital humanities; master's degree in artificial intelligence. The 89 participants were randomly allocated for testing only one of the three types of explainers. In other words, similarly to the first user study, also this second study

---

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

[20]https://www.cdc.gov
[21]https://www.heart.org
[22]https://medlineplus.gov
[23]https://www.medicalnewstoday.com
[24]All the courses of study were of an Italian university, and only the master's degrees were international, i.e., with English teachings and students from countries other than Italy.

Table 5
Quiz of the Heart Disease Predictor. This table contains the quiz used for evaluating the effectiveness of tools explaining the heart disease predictor. XAI is the normal *XAI-based explainer* (i.e., the webpage shown in Figure 4), HWN is the *how-why explainer*, and 2EC is the *two-level explainer*. Column *Steps* indicates the minimum number of steps (in terms of links to click, overviews to open, or questions to pose) required by each explanatory tool to provide the correct answer. Negative *steps* means that the correct answer cannot be found. In contrast, 0 *steps* means that the answer is immediately available in the initial explanans (i.e., the content of the webpage shown in Figure 4). Column "Archetype" indicates which interrogative particles represent the question.
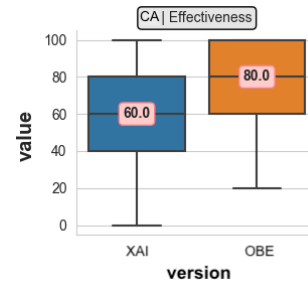
Question	Archetype	Steps			
		XAI	2EC	HWN	YAI4Hu
What are the most important factors leading that patient to medium risk of heart disease?	what, why	0	0	0	0 (no OQ)
What is the easiest thing the patient could do to change his heart disease risk from medium to low?	what, how	0	0	0	0 (no OQ)
According to the predictor, what serum cholesterol level is needed to shift the heart disease risk from medium to high?	what, how	0	0	0	0 (no OQ)
How could the patient avoid raising bad cholesterol, preventing his heart disease risk from shifting from medium to high?	how	-1	1	2	2
What tests can be done to measure bad cholesterol levels in the blood?	what, how	-1	1	-1	1
What are the risks of high cholesterol?	what, why not	-1	1	2	1
What is LDL?	what	-1	1	2	1
What is Serum Cholesterol?	what	-1	1	1	1
What types of chest pain are typical of heart disease?	what, how	-1	1	1	1
What is the most common type of heart disease in the USA?	what	-1	1	1	1
What are the causes of angina?	what, why	-1	1	2	1
What kind of chest pain do you feel with angina?	what, how	-1	1	1	1
What are the effects of high blood pressure?	what, why not	-1	1	1	1
What are the symptoms of high blood pressure?	what, why, how	-1	1	1	1
What are the effects of smoking on the cardiovascular system?	what, why not	-1	1	3	1
How can the patient increase his heart rate?	how	-1	1	3	1
How can the patient try to prevent a stroke?	how	-1	1	3	2
What is a Thallium stress test?	what, why	-1	1	3	1

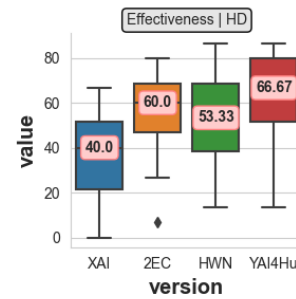followed a between-subjects design. In the end, there were approximately 20 participants per explainer.

Each participant evaluated the effectiveness of the four explainers by taking the multiple-choice quiz shown in Table 5. At the end of the effectiveness quiz, answers were automatically scored as correct (score 1) or not (score 0), and the resulting scores were added together to form the effectiveness score. For further details about this user study, read [66].

5.3.3. Results of 2nd Experiment

Both the results of the (first) user study involving 89 participants and the (second) user study involving 103 participants indicate that a better (i.e., more explainable) *explanandum support material* implies an explainer capable of producing more effective explanations. As also shown in Figure 6, according to a one-sided Mann-Whitney U-Test, there is enough statistical evidence to claim that the instance

Figure 6: 1st User Study: Effectiveness Scores on Questions that cannot be answered with the information provided by the XAI-based explainer. This figure shows a comparison of the median effectiveness scores obtained on the credit approval system (CA) with the normal *XAI-based explainer* (XAI; the blue one) and YAI4Hu without open-ended question-answering (called *overview-based explainer* or OBE for short; the orange one) on those questions whose answer is not provided by the XAI-based explainer. Results are shown as box plots (25th, 50th, 75th percentile, and whiskers covering all data and outliers). The numerical value of the medians is shown inside pink boxes. Differently from [64], here effectiveness scores are normalised in [0, 100].

Figure 7: 2nd User Study: Effectiveness Scores on Questions that cannot be answered with the information provided by the XAI-based explainer. Comparison of the results achieved on the heart disease predictor (HD) with the normal *XAI-based explainer* and the other explainers, only on those questions whose aspects are *not covered* by the information presented by the XAI. In particular, the other explainers are the *two-layered explainer* (2EC), the *how-why explainer* (HWN), and YAI4Hu. For more details about interpreting this figure, read the caption of Figure 6. Effectiveness scores are normalized in [0, 100].

of YAI4Hu considered for the second user study is more effective in credit approval system (U=849.5, p=.007) than the *XAI-based explainer* on those questions that cannot be answered by the XAI (i.e., questions number 2, 3, 4, 5 and 7 in Table 4).

Moreover, as shown in Figure 7, the same can be said for the heart disease predictor in the first user study. As expected, also in this case, we see the median effectiveness score of the normal XAI-based explainer being significantly lower than the other explainers on the questions that the XAI cannot answer (i.e., the questions with negative steps in Table 5). More precisely, according to some one-sided

Mann-Whitney U-tests, there is enough statistical evidence to claim that YAI4Hu is better than the XAI-based explainer (U=40, p=.0002) on those questions. The same can be said about the two-level explainer (U=48, p=.003) and the how-why explainer (U=65.5, p=.02).

Indeed, the difference between a normal XAI-based explainer and the other explainers is twofold. First of all, the explanations produced by YAI4Hu and the how-why explainer are interactive and more user-centered, while those of the normal XAI-based system are not. Secondly, the normal XAI-based explainer considers a smaller amount of explainable information. YAI4Hu, the how-why explainer, and the two-level explainer produce their explanations using more than 50 extra web pages that the XAI-based explainer does not see. This last difference allows us to exploit these user studies to test Hypothesis 1 further. The amount of information the normal XAI-based explainer handles is $\frac{1}{100}$ of all the other explainers.

In order to show that an increment in DoX causes a consequent increment in the effectiveness of explanations, we have to compute the DoX scores of the normal XAI-based explainer and the DoX scores of the other explainers involved in the user study. To do so, we identified the set of *explanandum aspects A* from the quizzes used to generate the effectiveness scores (see Table 4 and Table 5). These quizzes define what the users should know to be effective, indirectly defining what is essential for the system to explain: the *explanandum aspects*.

Eventually, if Hypothesis 1 were true, we would expect that the greater DoX is, the greater the effectiveness of an explainer. Notably, the opposite is not necessarily correct. Two explainers (with different presentation logics; e.g., the *two-layered explainer* and YAI4Hu) might have different effectiveness scores despite having the same DoX.

Computing the DoX scores for this second experiment, we got the results shown in Table 6. Importantly, these results confirmed our expectations for them. They indicate that the *two-level explainer*, the *overview-based explainer*, the *how-why explainer*, and YAI4Hu have higher DoX scores than the normal *XAI-based explainer* regardless their presentation logic.

6. Discussion and Analysis of Empirical Results: How to Use DoX for Assessing Law Compliance

The results of all experiments and user studies showed that Hypothesis 1 is valid. We see that DoX increases whenever a black-box AI is enclosed in a XAI and that an increase in DoX corresponds to a statistically significant increase in the effectiveness of the explanatory system. Therefore, we believe that our technology for estimating the DoX might be used for an objective and lawful algorithmic explainability assessment, 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*. To guarantee the reproducibility of the experiments, we published the source

Table 6

Results of the 2nd Experiment on DoXpy. The scores in this table are different from those of the first experiment (Table 3) because a different explanandum is considered for the second experiment. In this table, DoX and average (Avg) DoX are shown for the credit approval system (CA) and the heart disease predictor (HD). As columns, we have the normal XAI-based explainers (XAI, for short) and the other explainers, i.e., YAI4Hu, the two-level explainer, and the how-why narrator. For more details about interpreting this table, read the caption of Table 3.

		CA		HD	
		XAI	Others	XAI	Others
Avg DoX	ML	1.19	18.75	0.21	21.59
	TF	0.72	12.86	0.16	17.55
DoX	ML	"which": 1.26 "how": 1.26 "when": 1.25 "what": 1.24 "who": 1.18 "why": 1.16 "where": 1.13 "whose": 1.12	"how": 20.32 "what": 19.78 "when": 19.59 "which": 19.04 "why": 19.00 "whose": 17.34 "where": 17.19 "who": 17.09	"which": 0.23 "how": 0.21 "why": 0.21 "whose": 0.21 "what": 0.21 "when": 0.20 "where": 0.20 "who": 0.20	"why": 24.01 "which": 22.90 "how": 22.90 "what": 21.76 "whose": 21.37 "when": 21.01 "where": 19.91 "who": 19.61
	TF	"what": 0.88 "when": 0.75 "how": 0.69 "which": 0.67 "who": 0.66 "where": 0.64 "why": 0.57 "whose": 0.55	"what": 15.97 "when": 13.70 "how": 12.15 "who": 11.89 "where": 11.33 "which": 11.06 "why": 9.68 "whose": 9.32	"what": 0.19 "how": 0.16 "when": 0.15 "who": 0.15 "which": 0.14 "where": 0.14 "why": 0.14 "whose": 0.12	"what": 20.67 "when": 17.42 "how": 16.45 "who": 15.95 "which": 15.90 "why": 15.30 "where": 15.25 "whose": 13.35

code of DoXpy[25], as well as the code of the XAI-based systems, the user study questionnaires, and the remaining data mentioned within this paper.

In particular, 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*, the DoX scores increase, and this is also true 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* increases as well. This is true even for the *two-level explainer*, even though it is not interactive and does not re-organize information to make it simpler and easier to access, dumping on the user dozens of pages of content.

Our user studies involved more than 190 participants and were consistent across two somewhat different and broad user pools, producing statistically significant results (with p-values lower than 0.05). Therefore, considering that *explainability* is fundamentally the *ability to explain*, the two experiments combined tell us that our (average) DoX can quantitatively approximate the degree of explainability of information. In other words, we conclude from our experiments that DoX *can* be used as a proxy for measuring the explainability of an explanatory system, as long as a set of *explanandum aspects* can be defined. DoX is deterministic and entirely objective, and it could be used as a cheaper alternative to expensive non-deterministic user studies.

We are convinced that DoX may have a role in all applications where it is crucial to evaluate explainability

---

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

objectively. Indeed, the main benefit of DoX is that it works with any set of *explanandum aspects A*. Therefore it can be used to quantify how the explanations given by an AI are aligned with any of the Business-to-Business, and Business-to-Consumer requirements identified by Bibal et al. [10].

In particular, for each Business-to-Business and Business-to-Consumer requirement we may have the following set of *explanandum aspects A*:

- *Providing the main features used in a decision by the AI*: *A* can be the set of main feature labels used for a decision. This list can be generated with a XAI like CEM, TreeSHAP, or others.

- *Providing all features processed by the AI*: in this case, *A* is the set of all the feature labels considered by the AI.

- *Providing a comprehensive explanation of a specific decision taken by the AI*: *A* can be the set of aspects deemed relevant to the decision of the AI, i.e., what is the AI, what are the known issues of the AI, or all the other aspects discussed in [68].

- *Providing the underlying logical model followed by the AI*: in this case, *A* can be the set of all the nouns or noun/verbal phrases used in the textual description of the logical model of the AI.

Hence, the benefits of using DoX over a normal user study are manifold, in fact:

- DoX reduces testing costs normally sustained during subject-based evaluations.

- DoX allows the direct measurement of the degree of explainability of any piece of information for which a meaningful textual representation is written in a natural language (i.e., English).

- DoX disentangles the evaluation of the *explanandum support material* from that of the explainer (or presentation logic) and the interface.

In other words, DoX is a fully objective metric that could be used to understand whether a piece of information is sufficient to explain something regardless of whether the resulting explanations have been perceived as satisfactory and good by the explainees. We deem this characteristic of DoX to be very important: a poor degree of explainability objectively implies poor explanations, no matter how good the adopted explanatory process is (or how it 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" [15].

Despite all the good properties supported by both theory and empirical results, we found that DoX may have limitations that we plan to address in future works.

First of all, the results of the second experiment show that explanatory systems with the same DoX could be usable and effective in different ways. Indeed, this points to the fact that DoX should not be considered as a total replacement to user studies but rather as a cheaper alternative to consider while developing complex explanatory systems. In other words, DoX cannot fully replace subjective metrics (i.e., usability) if one wants to evaluate the user-centrality of an explanatory system or interface. On the other hand, DoX is probably better than subjective metrics if one wants to objectively evaluate the contents of an explanatory system to understand how many questions can be adequately answered: the higher DoX, the greater the chances to adequately explain to a variety of users.

Secondly, the numerical differences between the DoX scores shown in table 3 and 6 suggest that our algorithm for computing DoX scores may be sensitive to the choice of a deep language model for pertinence estimation. In fact, on the one hand, we see that the difference in terms of DoX between the normal *XAI-based explainers* and the other explainer tend to differ from MiniLM to the Universal Sentence Encoder slightly. Nonetheless, we also see that in all the considered experiments, the DoX scores increase as expected, with both MiniLM and the Universal Sentence Encoder, suggesting that the alignment of DoX to *explainability* is independent of the chosen deep language model. This intuition is supported by the fact that the deep language models, on average, perform reasonably well on existing benchmarks for evaluating answer retrieval algorithms. In other words, if the average DoX aggregates enough archetypes, aspects, and details, then different pertinence functions performing similarly on standard benchmarks may produce proportionally similar scores. This does not exclude the fact that some deep language models might be better than others for computing DoX scores or that multiple standardized deep language models should be adopted for a thorough estimate of the DoX. We leave this analysis for future work.

Another possible limitation of DoX is that its scores cannot be easily normalized in a $[0, 1]$ range. In fact, according to Definition 4, DoX is computed by performing a sum (called *cumulative pertinence*) over the set of details *D* extracted from an *explanandum support material*, so that DoX can measure the similarity of the *explanandum support material* to the explanandum. Unfortunately, it is impossible to know the total number of details of any possible *explanandum support material*. Therefore, it is impossible to normalize the score by dividing the *cumulative pertinence* by such number. It is worth noting that such a sum is necessary. Indeed, suppose the *cumulative pertinence* were a mean instead of a sum. In that case, the resulting score for an *explanandum support material* could not be compared to that of any larger (in terms of the number of details) *explanandum support material*, making pointless the use of DoX in the first place.

Furthermore, it is essential to mention that DoX, alone, is not sufficient for a thorough quantification of how much of the information is explained by an AI. Our definition of

DoX does not consider the correctness of information of the *explanandum support material*, assuming that truth is given and that it is different from explainability. In other words, DoX should always be used with other metrics that can evaluate the correctness of available information.

Finally, although DoX can be used to verify many of the requirements defined by [10], it is still unclear how to apply DoX to verify also Government-to-Citizen legal requirements. Selecting a reasonable threshold of DoX scores for law compliance is undoubtedly one of the challenges we envisage for a proper standardization of *explainability* in the industrial context. We also leave these analyses for future work.

7. Conclusions

In this paper, we proposed a new metric for explainability called DoX that could objectively quantify how much of the information is explained by an AI. For instance, DoX can be used to verify the satisfaction of Business-to-Business, and Business-to-Consumer requirements as defined by Bibal et al. [10].

DoX is based on the intuition coming from Achinstein's theory of explanations that explaining is an act of illocutionary question-answering. Specifically, DoX frames explanations as answers to many simple questions (*archetypes*), shedding light on the concepts being explained so that the more (archetypal) answers a corpus can give about essential aspects of an explanandum, the more that corpus is explainable. Thus DoX is the first explainability metric based on Ordinary Language Philosophy. It is a model-agnostic and deterministic approach that can work with any corpus of explainable information represented in natural language (i.e., English).

In particular, DoX quantifies the three main criteria of explainability adequacy defined by Carnap: similarity, exactness, and fruitfulness. In this sense, our contribution is a mechanism for quantifying Carnap's criteria and aggregating them together in one single score called average DoX, used to compare the degree of explainability of different explanatory systems. DoX can quantify the degree of explainability of a corpus of information by estimating how adequately that corpus could answer an arbitrary set of archetypal questions about the concepts of an explanandum.

Throughout the paper, we also presented a concrete implementation of DoX called DoXpy.

In order to understand whether the DoX is behaving as expected, we designed a few experiments on two realistic systems for heart disease prediction and credit approval, involving state-of-the-art AI technologies such as Artificial Neural Networks, TreeSHAP [43], XGBoost [19], and CEM [21]. The results show that the DoX is aligned with our expectations and can be used to quantify *explainability* in natural language information corpora.

Although DoX cannot be used directly on a black-box model to understand how much of it can be explained, it can be used on the output of an ensemble of XAI algorithms or

any other explainable information (e.g., documentation, papers, books) to understand how that information can be used to explain. In this sense, DoX is the most useful when used to evaluate extensive collections of explainable information (e.g., the output of an ensemble of XAI algorithms).

Another context for applying DoX could be education. Not surprisingly, many would argue that explanations are one of the primary artifacts through which humans understand reality and learn to solve complex problems [9]. Therefore, *explaining* is central to XAI and education, and these are two contexts where our technology and understanding of explanations could be of utmost importance.

References

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

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

[3] Arras, L., Osman, A., Samek, W., 2022. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. Inf. Fusion 81, 14–40. URL: https://doi.org/10.1016/j.inffus.2021.11.008, doi:10.1016/j.inffus.2021.11.008.

[4] Arrieta, A.B., Rodríguez, N.D., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115. URL: https://doi.org/10.1016/j.inffus.2019.12.012, doi:10.1016/j.inffus.2019.12.012.

[5] Austin, J., Urmson, J., Sbisà, M., 1975. How to Do Things with Words. William James lectures, Clarendon Press. URL: https://books.google.it/books?id=XnRkQSTUpmgC.

[6] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N., 2013. Abstract meaning representation for sembanking, in: Dipper, S., Liakata, M., Pareja-Lora, A. (Eds.), Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, LAW-ID@ACL 2013, August 8-9, 2013, Sofia, Bulgaria, The Association for Computer Linguistics. pp. 178–186. URL: https://aclanthology.org/W13-2322/.

[7] Baudis, P., Sedivý, J., 2015. Modeling of the question answering task in the yodaqa system, in: Mothe, J., Savoy, J., Kamps, J., Pinel-Sauvagnat, K., Jones, G.J.F., SanJuan, E., Cappellato, L., Ferro, N. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings, Springer. pp. 222–228. URL: https://doi.org/10.1007/978-3-319-24027-5_20, doi:10.1007/978-3-319-24027-5_20.

[8] Berant, J., Chou, A., Frostig, R., Liang, P., 2013. Semantic parsing on freebase from question-answer pairs, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL. pp. 1533–1544. URL: https://aclanthology.org/D13-1160/.

[9] Berland, L.K., Reiser, B.J., 2009. Making sense of argumentation and explanation. Science Education 93, 26–55.

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

[11] Bos, J., 2016. Expressive power of abstract meaning representations. Comput. Linguistics 42, 527–535. URL: https://doi.org/10.1162/

COLI_a_00257, doi:10.1162/COLI_a_00257.

[12] Bowman, S.R., Angeli, G., Potts, C., Manning, C.D., 2015. A large annotated corpus for learning natural language inference, in: Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y. (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, The Association for Computational Linguistics. pp. 632–642. URL: https://doi.org/10.18653/v1/d15-1075, doi:10.18653/v1/d15-1075.

[13] Bromberger, S., 1966. Why-questions, in: Colodny, R.G. (Ed.), Mind and Cosmos – Essays in Contemporary Science and Philosophy. University of Pittsburgh Press, pp. 86–111.

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

[15] Buçinca, Z., Lin, P., Gajos, K.Z., Glassman, E.L., 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, in: Paternò, F., Oliver, N., Conati, C., Spano, L.D., Tintarev, N. (Eds.), IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020, ACM. pp. 454–464. URL: https://doi.org/10.1145/3377325.3377498, doi:10.1145/3377325.3377498.

[16] Carnap, R., Schilpp, P.A., 1963. The Philosophy of Rudolf Carnap. Cambridge University Press Cambridge.

[17] Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M.B., Preece, A.D., Julier, S., Rao, R.M., Kelley, T.D., Braines, D., Sensoy, M., Willis, C.J., Gurram, P., 2017. 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 2017, San Francisco, CA, USA, August 4-8, 2017, IEEE. pp. 1–6. URL: https://doi.org/10.1109/UIC-ATC.2017.8397411, doi:10.1109/UIC-ATC.2017.8397411.

[18] Chen, D., Yih, W., 2020. Open-domain question answering, in: Savary, A., Zhang, Y. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, ACL 2020, Online, July 5, 2020, Association for Computational Linguistics. pp. 34–37. URL: https://doi.org/10.18653/v1/2020.acl-tutorials.8.

[19] Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (Eds.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, ACM. pp. 785–794. URL: https://doi.org/10.1145/2939672.2939785, doi:10.1145/2939672.2939785.

[20] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.J., Sandhu, S., Guppy, K.H., Lee, S., Froelicher, V., 1989. International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology 64, 304–310. URL: https://www.sciencedirect.com/science/article/pii/0002914989905249, doi:https://doi.org/10.1016/0002-9149(89)90524-9.

[21] Dhurandhar, A., Chen, P., Luss, R., Tu, C., Ting, P., Shanmugam, K., Das, P., 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives, in: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pp. 590–601. URL: https://proceedings.neurips.cc/paper/2018/hash/c5ff2543b53f4cc0ad3819a36752467b-Abstract.html.

[22] Dieber, J., Kirrane, S., 2022. A novel model usability evaluation framework (muse) for explainable artificial intelligence. Inf. Fusion 81, 143–153. URL: https://doi.org/10.1016/j.inffus.2021.11.017, doi:10.1016/j.inffus.2021.11.017.

[23] FitzGerald, N., Michael, J., He, L., Zettlemoyer, L., 2018. Large-scale QA-SRL parsing, in: Gurevych, I., Miyao, Y. (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics. pp. 2051–2060. URL: https://aclanthology.org/P18-1191/, doi:10.18653/v1/P18-1191.

[24] van Fraassen, B., Press, O.U., Van Fraassen, P., 1980. The Scientific Image. Clarendon Library of Logic and Philosophy, Clarendon Press. URL: https://books.google.it/books?id=VLz2F1zMr9QC.

[25] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M.A., Kagal, L., 2018. Explaining explanations: An overview of interpretability of machine learning, in: Bonchi, F., Provost, F.J., Eliassi-Rad, T., Wang, W., Cattuto, C., Ghani, R. (Eds.), 5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018, IEEE. pp. 80–89. URL: https://doi.org/10.1109/DSAA.2018.00018, doi:10.1109/DSAA.2018.00018.

[26] Guo, M., Yang, Y., Cer, D., Shen, Q., Constant, N., 2021. MultiReQA: A cross-domain evaluation forRetrieval question answering models, in: Proceedings of the Second Workshop on Domain Adaptation for NLP, Association for Computational Linguistics, Kyiv, Ukraine. pp. 94–104. URL: https://aclanthology.org/2021.adaptnlp-1.10.

[27] He, L., Lewis, M., Zettlemoyer, L., 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language, in: Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y. (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, The Association for Computational Linguistics. pp. 643–653. URL: https://doi.org/10.18653/v1/d15-1076, doi:10.18653/v1/d15-1076.

[28] Hempel, C.G., Oppenheim, P., 1948. Studies in the logic of explanation. Philosophy of Science 15, 135–175. doi:10.1086/286983.

[29] Hilton, D.J., 1996. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. Thinking & Reasoning 2, 273–308. URL: https://doi.org/10.1080/135467896394447, doi:10.1080/135467896394447, arXiv:https://doi.org/10.1080/135467896394447.

[30] Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J., 2018. Metrics for explainable AI: challenges and prospects. CoRR abs/1812.04608. URL: http://arxiv.org/abs/1812.04608, arXiv:1812.04608.

[31] Holland, J., Holyoak, K., Nisbett, R., Thagard, P., 1986. Induction: Processes of Inference, Learning, and Discovery. Bradford books, MIT Press. URL: https://books.google.it/books?id=Z6EFBaLApE8C.

[32] Holzinger, A., Carrington, A.M., Muller, H., 2020. Measuring the quality of explanations: The system causability scale (SCS). Kunstliche Intell. 34, 193–198. URL: https://doi.org/10.1007/s13218-020-00636-z, doi:10.1007/s13218-020-00636-z.

[33] Huang, Z., Xu, S., Hu, M., Wang, X., Qiu, J., Fu, Y., Zhao, Y., Peng, Y., Wang, C., 2020. Recent trends in deep learning based open-domain textual question answering systems. IEEE Access 8, 94341–94356. URL: https://doi.org/10.1109/ACCESS.2020.2988903, doi:10.1109/ACCESS.2020.2988903.

[34] Jansen, P., Balasubramanian, N., Surdeanu, M., Clark, P., 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams, in: Calzolari, N., Matsumoto, Y., Prasad, R. (Eds.), COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, ACL. pp. 2956–2965. URL: https://aclanthology.org/C16-1278/.

[35] Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L., 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, in: Barzilay, R., Kan, M. (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, Association for Computational Linguistics. pp. 1601–1611. URL: https://doi.org/10.18653/v1/P17-1147, doi:10.18653/v1/P17-1147.

[36] Karpukhin, V., Oguz, B., Min, S., Lewis, P.S.H., Wu, L., Edunov, S., Chen, D., Yih, W., 2020. Dense passage retrieval for open-domain

question answering, in: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics. pp. 6769–6781. URL: https://doi.org/10.18653/v1/2020.emnlp-main.550, doi:10.18653/v1/2020.emnlp-main.550.

[37] Keane, M.T., Kenny, E.M., Delaney, E., Smyth, B., 2021. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques, in: Zhou, Z. (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org. pp. 4466–4474. URL: https://doi.org/10.24963/ijcai.2021/609, doi:10.24963/ijcai.2021/609.

[38] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A.P., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S., 2019. Natural questions: a benchmark for question answering research. Trans. Assoc. Comput. Linguistics 7, 452–466. URL: https://doi.org/10.1162/tacl_a_00276, doi:10.1162/tacl_a_00276.

[39] Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J., 2017. Interpretable & explorable approximations of black box models. CoRR abs/1707.01154. URL: http://arxiv.org/abs/1707.01154, arXiv:1707.01154.

[40] Langkilde, I., Knight, K., 1998. Generation that exploits corpus-based statistical knowledge, in: Boitet, C., Whitelock, P. (Eds.), 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference, Morgan Kaufmann Publishers / ACL. pp. 704–710. URL: https://aclanthology.org/P98-1116/, doi:10.3115/980845.980963.

[41] Liao, Q.V., Gruen, D.M., Miller, S., 2020. Questioning the AI: informing design practices for explainable AI user experiences, in: Bernhaupt, R., Mueller, F.F., Verweij, D., Andres, J., McGrenere, J., Cockburn, A., Avellino, I., Goguey, A., Bjøn, P., Zhao, S., Samson, B.P., Kocielnik, R. (Eds.), CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020, ACM. pp. 1–15. URL: https://doi.org/10.1145/3313831.3376590, doi:10.1145/3313831.3376590.

[42] Lim, B.Y., Dey, A.K., Avrahami, D., 2009. *Why and why not* explanations improve the intelligibility of context-aware intelligent systems, in: Jr., D.R.O., Arthur, R.B., Hinckley, K., Morris, M.R., Hudson, S.E., Greenberg, S. (Eds.), Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009, ACM. pp. 2119–2128. URL: https://doi.org/10.1145/1518701.1519023, doi:10.1145/1518701.1519023.

[43] Lundberg, S.M., Erion, G.G., Chen, H., DeGrave, A.J., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67. URL: https://doi.org/10.1038/s42256-019-0138-9, doi:10.1038/s42256-019-0138-9.

[44] Madumal, P., Miller, T., Sonenberg, L., Vetere, F., 2019. A grounded interaction protocol for explainable artificial intelligence, in: Elkind, E., Veloso, M., Agmon, N., Taylor, M.E. (Eds.), Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019, International Foundation for Autonomous Agents and Multiagent Systems. pp. 1033–1041. URL: http://dl.acm.org/citation.cfm?id=3331801.

[45] Michael, J., Stanovsky, G., He, L., Dagan, I., Zettlemoyer, L., 2018. Crowdsourcing question-answer meaning representations, in: Walker, M.A., Ji, H., Stent, A. (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), Association for Computational Linguistics. pp. 560–568. URL: https://doi.org/10.18653/v1/n18-2089, doi:10.18653/v1/n18-2089.

[46] Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. 267, 1–38. URL: https://doi.org/10.1016/j.artint.2018.07.007, doi:10.1016/j.artint.2018.07.007.

[47] Miltsakaki, E., Prasad, R., Joshi, A.K., Webber, B.L., 2004. The penn discourse treebank, in: Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal, European Language Resources Association. URL: http://www.lrec-conf.org/proceedings/lrec2004/summaries/618.htm.

[48] Mohseni, S., Block, J.E., Ragan, E.D., 2021. Quantitative evaluation of machine learning explanations: A human-grounded benchmark, in: Hammond, T., Verbert, K., Parra, D., Knijnenburg, B.P., O'Donovan, J., Teale, P. (Eds.), IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021, ACM. pp. 22–31. URL: https://doi.org/10.1145/3397481.3450689, doi:10.1145/3397481.3450689.

[49] Nguyen, A., Martínez, M.R., 2020. On quantitative aspects of model interpretability. CoRR abs/2007.07584. URL: https://arxiv.org/abs/2007.07584, arXiv:2007.07584.

[50] Novaes, C.D., Reck, E.H., 2017. Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization. Synth. 194, 195–215. URL: https://doi.org/10.1007/s11229-015-0816-z, doi:10.1007/s11229-015-0816-z.

[51] Palan, S., Schitter, C., 2018. Prolific.ac—a subject pool for online experiments. Journal of Behavioral and Experimental Finance 17, 22–27. URL: https://www.sciencedirect.com/science/article/pii/S2214635017300989, doi:https://doi.org/10.1016/j.jbef.2017.12.004.

[52] Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W., Wallach, H.M., 2021. Manipulating and measuring model interpretability, in: Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T., Bjørn, P., Drucker, S.M. (Eds.), CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021, ACM. pp. 237:1–237:52. URL: https://doi.org/10.1145/3411764.3445315, doi:10.1145/3411764.3445315.

[53] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A.K., Webber, B.L., 2008. The penn discourse treebank 2.0, in: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco, European Language Resources Association. URL: http://www.lrec-conf.org/proceedings/lrec2008/summaries/754.html.

[54] Pyatkin, V., Klein, A., Tsarfaty, R., Dagan, I., 2020. Qadiscourse - discourse relations as QA pairs: Representation, crowdsourcing and baselines, in: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics. pp. 2804–2819. URL: https://doi.org/10.18653/v1/2020.emnlp-main.224, doi:10.18653/v1/2020.emnlp-main.224.

[55] Rebanal, J.C., Combitsis, J., Tang, Y., Chen, X.A., 2021. Xalgo: a design probe of explaining algorithms' internal states via question-answering, in: Hammond, T., Verbert, K., Parra, D., Knijnenburg, B.P., O'Donovan, J., Teale, P. (Eds.), IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021, ACM. pp. 329–339. URL: https://doi.org/10.1145/3397481.3450676, doi:10.1145/3397481.3450676.

[56] Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks, in: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics. pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410, doi:10.18653/v1/D19-1410.

[57] Ribera, M., Lapedriza, À., 2019. Can we do better explanations? A proposal of user-centered explainable AI, in: Trattner, C., Parra, D., Riche, N. (Eds.), Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI 2019), Los Angeles, USA, March 20,

2019, CEUR-WS.org. p. 38. URL: http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf.

[58] Rosenfeld, A., 2021. Better metrics for evaluating explainable artificial intelligence, in: Dignum, F., Lomuscio, A., Endriss, U., Nowé, A. (Eds.), AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021, ACM. pp. 45–50. URL: https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p45.pdf, doi:10.5555/3463952.3463962.

[59] Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215. URL: https://doi.org/10.1038/s42256-019-0048-x, doi:10.1038/s42256-019-0048-x.

[60] Salmon, W., 1984. Scientific Explanation and the Causal Structure of the World. Book collections on Project MUSE, Princeton University Press. URL: https://books.google.it/books?id=2ug9DwAAQBAJ.

[61] Sellars, W., 1963. Science, Perception and Reality. New York: Humanities Press.

[62] Sovrano, F., Palmirani, M., Vitali, F., 2020a. Legal knowledge extraction for knowledge graph based question-answering, in: Villata, S., Harasta, J., Kremen, P. (Eds.), Legal Knowledge and Information Systems - JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020, IOS Press. pp. 143–153. URL: https://doi.org/10.3233/FAIA200858, doi:10.3233/FAIA200858.

[63] Sovrano, F., Sapienza, S., Palmirani, M., Vitali, F., 2021. A survey on methods and metrics for the assessment of explainability under the proposed AI act, in: Erich, S. (Ed.), Legal Knowledge and Information Systems - JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021, IOS Press. pp. 235–242. URL: https://doi.org/10.3233/FAIA210342, doi:10.3233/FAIA210342.

[64] Sovrano, F., Vitali, F., 2021. From philosophy to interfaces: an explanatory method and a tool inspired by achinstein's theory of explanation, in: Hammond, T., Verbert, K., Parra, D., Knijnenburg, B.P., O'Donovan, J., Teale, P. (Eds.), IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021, ACM. pp. 81–91. URL: https://doi.org/10.1145/3397481.3450655, doi:10.1145/3397481.3450655.

[65] Sovrano, F., Vitali, F., 2022a. Explanatory artificial intelligence (yai): human-centered explanations of explainable ai and complex data. Data Mining and Knowledge Discovery URL: https://doi.org/10.1007/s10618-022-00872-x, doi:10.1007/s10618-022-00872-x.

[66] Sovrano, F., Vitali, F., 2022b. Generating user-centred explanations via illocutionary question answering: From philosophy to interfaces. ACM Trans. Interact. Intell. Syst. 12. URL: https://doi.org/10.1145/3519265, doi:10.1145/3519265.

[67] Sovrano, F., Vitali, F., 2022c. How to quantify the degree of explainability: Experiments and practical implications, in: 31th IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2022, Padova, July 18-23, 2022, IEEE. pp. 1–9.

[68] Sovrano, F., Vitali, F., Palmirani, M., 2020b. Modelling gdpr-compliant explanations for trustworthy AI, in: Ko, A., Francesconi, E., Kotsis, G., Tjoa, A.M., Khalil, I. (Eds.), Electronic Government and the Information Systems Perspective - 9th International Conference, EGOVIS 2020, Bratislava, Slovakia, September 14-17, 2020, Proceedings, Springer. pp. 219–233. URL: https://doi.org/10.1007/978-3-030-58957-8_16, doi:10.1007/978-3-030-58957-8_16.

[69] Stede, M., 2013. Discourse processing, in: Vanderwende, L., III, H.D., Kirchhoff, K. (Eds.), Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, The Association for Computational Linguistics. pp. 4–6. URL: https://aclanthology.org/N13-4002/.

[70] Szymanski, M., Millecamp, M., Verbert, K., 2021. Visual, textual or hybrid: the effect of user expertise on different explanations, in: Hammond, T., Verbert, K., Parra, D., Knijnenburg, B.P., O'Donovan, J., Teale, P. (Eds.), IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021, ACM. pp. 109–119. URL: https://doi.org/10.1145/3397481.3450662,

doi:10.1145/3397481.3450662.

[71] Vilone, G., Longo, L., 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion 76, 89–106. URL: https://doi.org/10.1016/j.inffus.2021.05.009, doi:10.1016/j.inffus.2021.05.009.

[72] Vilone, G., Rizzo, L., Longo, L., 2020. A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence, in: Longo, L., Rizzo, L., Hunter, E., Pakrashi, A. (Eds.), Proceedings of The 28th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Republic of Ireland, December 7-8, 2020, CEUR-WS.org. pp. 85–96. URL: http://ceur-ws.org/Vol-2771/AICS2020_paper_33.pdf.

[73] Wachter, S., Mittelstadt, B., Russell, C., 2018. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law and Technology 31. URL: http://dx.doi.org/10.2139/ssrn.3063289, doi:10.2139/ssrn.3063289.

[74] Wang, X., Yin, M., 2021. Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making, in: Hammond, T., Verbert, K., Parra, D., Knijnenburg, B.P., O'Donovan, J., Teale, P. (Eds.), IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021, ACM. pp. 318–328. URL: https://doi.org/10.1145/3397481.3450650, doi:10.1145/3397481.3450650.

[75] Webber, B., Prasad, R., Lee, A., Joshi, A., 2019. The penn discourse treebank 3.0 annotation manual. Philadelphia, University of Pennsylvania .

[76] Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Ábrego, G.H., Yuan, S., Tar, C., Sung, Y., Strope, B., Kurzweil, R., 2020. Multilingual universal sentence encoder for semantic retrieval, in: Celikyilmaz, A., Wen, T. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics. pp. 87–94. URL: https://doi.org/10.18653/v1/2020.acl-demos.12, doi:10.18653/v1/2020.acl-demos.12.

[77] Zufferey, S., Degand, L., 2017. Annotating the meaning of discourse connectives in multilingual corpora. Corpus Linguistics and Linguistic Theory 13, 399–422. URL: https://doi.org/10.1515/cllt-2013-0022.