

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

Francesco Sovrano^{a,*}, Fabio Vitali^a

^aDISI, University of Bologna, Mura Anteo Zamboni 7, Bologna, 40126, Italy

ARTICLE INFO

Keywords:

Degree of Explainability
Objective Explainability Metric
Explainable AI
Theory of Explanations

ABSTRACT

Explainable AI was born as a pathway to allow humans to explore and understand the inner working of complex systems. But establishing what *is* an explanation and *objectively* evaluating *explainability*, are not trivial tasks. With this paper, we present a new model-agnostic metric to measure the Degree of Explainability of information in an *objective* way, exploiting a specific theoretical model from Ordinary Language Philosophy called the *Achinstein's Theory of Explanations*, implemented with an algorithm relying on deep language models for knowledge graph extraction and information retrieval. In order to understand whether this metric is actually behaving as *explainability* is expected to, we devised a few experiments and user-studies involving more than 160 participants, evaluating two realistic AI-based systems for *healthcare* and *finance* using famous AI technology including Artificial Neural Networks and TreeSHAP. The results we obtained are statistically significant (with *P* values lower than 0.01), suggesting that our proposed metric for measuring the Degree of Explainability is robust on several scenarios and it aligns with concrete expectations.

1. Introduction

Recent advances in Artificial Intelligence (AI) are enabling computer science and engineering to create machines that can learn from rough data, hence automating tasks previously thought to be accessible only by biological intelligence. But these advances and results seem to come at a cost in terms of explainability, so that the most effective machine learning techniques are, so far, not easily interpretable in symbolic terms [15, 52].

The paradigms that address this explainability problem usually 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]. In fact, we are recently assisting to an increasing demand for explainability in AI applications, motivated by the growing awareness that transparency is pivotal for fairness and lawfulness.

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) [8]. As a result, the EU is indirectly posing an interesting challenge to the Explainable AI (XAI) community by calling for more transparent, user-centred 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. [8] have identified four main 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 letter we want to further expand the work of Bibal et al., trying to understand whether it is possible to objectively quantify how much of the information required by 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 that is mainly inspired by Ordinary Language Philosophy instead of Cognitive Science. Our approach is based on a specific theoretical model of explanation, called the *Achinstein's theory of explanations*, where explanations are the result of an *illocutionary* (i.e., broad yet pertinent and deliberate) act of pragmatically answering to a question. Accordingly, explanations are actually answers to many different basic questions (*archetypes*) each of which sheds a different light over the concepts being explained. As consequence, the more (archetypal) answers an *automated decision-making system* is able to give about the important aspects of its explanandum¹, 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 [58]. Thus, drawing also from Carnap's criteria of adequacy of an explication [44], 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)

¹The 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 DoX by means of pre-trained *deep language models* for general-purpose answer retrieval (e.g., [32, 10]) applied to a special 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, able to compute DoX scores. We also performed a few experiments and user-studies involving more than 160 participants, with the objective of showing that *explainability* changes in accordance with varying DoX scores. Importantly, the results of all our experiments and user-studies clearly and undoubtedly showed that, in fact, our hypothesis 1 holds.

This paper is structured as follows. In section 2 we give the necessary background information to properly introduce the theoretical models 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 by means of deep language models. In section 5 we describe our experiments, we then present the findings in section 6, and discuss them in section 7. Finally, in section 8 we discuss some possible limitations of DoX, pointing to future work and conclusions in section 9.

For guaranteeing the reproducibility of the experiments, we publish the source code³ of the algorithm for computing DoX scores, as well as the code of the XAI-based systems, the full details of our user-studies and the full set of data mentioned in this paper.

2. Background

In this section we provide some background to justify and support the rest of the paper. Hereby we briefly summarise a number of recent and less recent approaches to the theories of explanation, with a particular due focus on Achinstein's. After that, we discuss how Achinstein's theory of explaining as a question-answering process is compatible with existing XAI literature, highlighting how profound is the connection between answering questions and explaining in this field.

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

³<https://github.com/Francesco-Sovrano/DoXpy>

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 [14]. Even though Carnap studies the concept of *explication* rather than that of *explainable information*, we assert that they share a common ground making his criteria fitting in both cases. In fact, *explication* in Carnap's sense is the replacement of a somewhat unclear and inexact concept, the *explicandum*, by a new, clearer, and more exact concept, the *explicatum*⁴, and this is exactly what information does when made explainable.

Carnap's central criteria of explication adequacy[14] 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 seem to be transversal to all the identified definitions of explainability, possessing preliminary characteristics for any piece of information to be considered properly explainable. Interestingly, the property of *truthfulness* (being different from *exactness*) is not explicitly mentioned in Carnap's desiderata. That is to say that explainability and *truthfulness* are complementary, but different, as discussed also by [26]. In fact, 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 very poor 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" [25], 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 [11, 53]. Several more modern and competing theories of explanations were the result of this criticism.

Summarising our full analysis [56], the five most important theories of explanation in contemporary philosophy are: Causal Realism, Constructive Empiricism, Ordinary Language Philosophy, Cognitive Science, Naturalism and Scientific Realism. As consequence, there are at least five

⁴i.e., "what has been explained", in Latin.

⁵Carnap also discussed another desideratum, *simplicity*, but this criterion is presented as being subordinate to the others (especially exactness).

Table 1

Definitions of Explainability: 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 [53] | Descriptions of causality, expressed as chains of causes and effects. | What can fully describe causality. |
| Constructive Empiricism [21] | Contrastive information that answers <i>why</i> 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 <i>why</i> questions. |
| Ordinary Language Philosophy [1] | Pragmatically answering multiple types of questions (not just <i>why</i> ones), with the explicit intent of producing understanding in someone. | What can be used to pertinently answer questions about relevant aspects, in an illocutionary way. |
| Cognitive Science [28] | A process triggered as response to predictive failures and meant to provide information to fix failures in someone's mental model. | What can fix failures in mental models. |
| Naturalism and Scientific Realism [54] | An iterative process of confirmation of truths aimed at improving understanding. Explanations increase someone's understanding not simply by being the correct answer to a particular question, but by increasing the coherence of her/his entire belief system. | What can be used to increase understanding, i.e., by answering to particular questions. |

different definitions of “explanation”, one per theory. A summary of these definitions is shown in Table 1, highlighting that there is no full agreement between them on the nature of explanations.

Importantly, 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 Causal Realism [53], we have a definition of “explainable information” that slightly differs from that of “explanation”, as described in [56]. 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 are the result of an *illocutionary* act of pragmatically answering to a question. In particular, it means that there is a subtle and important difference between simply “answering to questions” and “explaining”, and this difference is *illocution*.

It appears that an *illocutionary* act results from a clear intent of achieving the goal of such act, as a promise being “what it is” just because of the intent of maintaining it. So that *illocution* in explaining makes an explanation as such just because it is the result of an underlying and proper intent of explaining.

Despite this definition, *illocution* seems to be too abstract to be implementable inside a real software application. Nonetheless, recent efforts towards the automated generation of explanations [57, 58], have shown that it may be possible to define *illocution* in a more “computer-friendly” way. Indeed, as stated in [57], *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 on 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”. In other terms, archetypal questions identify generic explanations about a specific aspect to explain (e.g., a topic, an argument, a concept), in a given informative context.*

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”, etc. Some archetypal questions in this case might be “What is an angina?” or “Why a stroke?”.

2.4. Explainable AI and Question Answering

If we assume that the interpretation of Achinstein's theory of explanations given by [57] is correct, then data or processes are said to be *explainable* when their informative content can adequately answer *archetypal questions*.

The idea of answering questions as explaining is not new to the field of XAI [36] and it is also quite compatible with our intuition of what constitutes an explanation. In fact, it is common to many works in the field [50, 37, 41, 22, 18, 64,

48, 30, 39] the use of generic (e.g., why, who, how, when) or more punctual questions to clearly define and describe the characteristics of explainability [36].

For example, Lundberg et al. [38] 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. [18] 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. [48] 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, and they highlight how deep is in this field the connection between answering questions and explaining.

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 [41, 27] instead. This is probably because Achinstein’s illocutionary theory of explanations is seemingly difficult to be implemented into software, by being utterly pragmatic. In fact, *user-orientedness* is challenging and sometimes not obviously connected to the main goal of XAI: “opening the black-box” (e.g., understanding how and why an opaque AI model works).

3. Related Work

Being able to measure 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 evidences and experiments backing up their own 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 suggested also by literature reviews (e.g., [62], and especially [56], 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, 51, 63, 43, 35, 33] can only be used with specific types of XAI approaches (e.g., prototype selection or feature attribution), while the metrics proposed in [27, 29, 19] rely on user-studies, as many other works [57, 42, 65, 61, 13, 46], based on classical usability metrics (i.e., effectiveness, efficiency, satisfaction).

Only one work among those examined, Hoffman et al. [27], claims that 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, [27] is mainly inspired by the interpretation of explanations given by Cognitive Science, requiring to measure: 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 [27] is non-deterministic and heavily relying on subjective measurements, despite being model-agnostic. The metric we propose here, DoX, is objective, deterministic and model-agnostic⁷ and 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 resulting explanations are perceived as satisfactory and good by the explainees.

Furthermore, only DoX and [35] appear to measure all the three main Carnap’s desiderata. More specifically, Lakkaraju et al. [35] evaluate Carnap’s criteria separately, while with DoX we propose a single metric that combines all of them.

Finally, as suggested in [56], all existing explainability metrics can be aligned to different interpretations of explainability coming from complementary theories of explanations. As shown in table 2, the vast majority of these metrics seems to be aligned to Causal Realism and Cognitive Science, while DoX is the very 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 objectively measure the degree of explainability with a fully automated software. With this paper we try to answer this question, by leveraging on an extension of Achinstein’s theory of explanations as proposed in [57] 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*. In fact, considering that *explainability* is fundamentally the *ability to explain*, it is clear that a proper definition of it requires a precise understanding of what is *explaining*.

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

4.1. Quantifying the Degree of Explainability

As discussed in section 2.4, the informative contents of state-of-the-art XAI is clearly polarised towards answering

⁷DoX 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.

⁶This table extends a similar one in [56].

Table 2

Comparison of Different Explainability Metrics⁶: 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 |
|------------------------------|------------------------------------|---|-----------------|--|--|
| [51] | Rule-based | Causal Realism | No | Exactness, Fruitfulness | Performance Difference, Number of Rules, Number of Features, Stability |
| [63] | Rule-based | Causal Realism | No | Similarity, Fruitfulness | Fidelity, Completeness |
| [43] | Feature Attribution | Causal Realism | No | Exactness, Fruitfulness | Monotonicity, Non-sensitivity, Effective Complexity |
| [35] | Rule-based | Causal Realism | No | Similarity, Exactness, Fruitfulness | Fidelity, Unambiguity, Interpretability, Interactivity |
| [29] | Any | Causal Realism, Cognitive Science, Naturalism & Co. | Yes | Exactness, Fruitfulness | System Causability Scale |
| [27] | Any | Cognitive Science, Naturalism & Co. | Yes | Exactness, Fruitfulness | Satisfaction, Trust, Mental Models, Curiosity, Performance |
| [19, 57, 42, 65, 61, 13, 46] | 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 |
| [33] | Prototype-based | Constructive Empiricism | No | Exactness | Proximity, Sparsity, Adequacy (Coverage) |
| [43] | 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 |

why, what if or how questions. Considering that why, what if and how are different questions pointing to different types of information, which type is the best one? We assert that the correct answer to this question is: “none”. In fact, depending on the needs of the explainee, its background knowledge, the context, and potentially many other factors, each archetype may be equally important.

In other words, depending on the characteristics of the explainee (e.g., background knowledge, objectives, context, etc.), 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 an XAI-based system 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 concept. For this reason, as discussed in section 2.3, in [57] we suggested that *illocution* (or, better, *explanatory illocution*) is in fact the process of answering multiple generic and primitive questions (e.g., why, how, what, etc.) called *archetypal questions*.

For example, if someone would be 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 in a quantitative way. 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 propose to estimate the degree of explainability of a piece of information by measuring the relevance with which it can answer a (pre-defined) set of archetypal questions.

Hence, our theoretical contribution, unfolded in the following sub-sections, consists in the precise and formal definition of: *cumulative pertinence*, *explanatory illocution*, *Degree of Explainability (DoX)*, and *Average DoX*.

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 and it 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, paragraphs, etc.) 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). *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 $\{\forall q \in Q \mid \langle q, P_{D_a, q_a} \rangle\}$ ⁸.*

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 $D = \{\forall a \in A, \forall d \in D_a \mid d\}$, then the DoX is the set $\{\forall q \in Q \mid \langle q, R_{D, q, A} \rangle\}$.*

4.1.2. Interpreting DoX in Terms of Carnap's Criteria

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 with respect to 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 formation⁹ [12], but rather that of being precise or pertinent enough as an answer to a given question.

In fact, the number of relevant aspects A covered by a given piece of information, and the number of details D that are pertinent about it, roughly say how much *similar*

that information is to the explanandum. More precisely, the formula used for computing the cumulative pertinences P_{D_a, q_a} sums the contribution of each single detail according to its pertinence to the aspects $a \in A$, telling us how much D_a is similar to a . Thus that if *pertinence* $p(d, q_a)$ is close to zero for all archetypes $q \in Q$, then a detail d has nothing to do with an aspect a . Furthermore, the average cumulative pertinence $R_{D, q, A}$ contains information about the *exactness* of multiple archetypal explanations, being an aggregation of pertinence scores. As result, by measuring $R_{D, q, A}$ for all the $q \in Q$ we obtain also an estimate of how D is *fruitful* for the formulation of many other different explanations intended as the result of an illocutionary act of answering questions.

For example, suppose the set of relevant aspects $A = \{\text{heart, stroke, vessel, disease, angina, symptom}\}$ and the sentence "I am happy that my article has been accepted in this prestigious journal" as Φ . In this case $D = \emptyset$ 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 fact, no detail of Φ is explaining anything about A , therefore the explainability of Φ with respect to A is 0.

On the other hand, we would not have a null DoX for A with the sentence "Angina happens when some part of your heart doesn't get enough oxygen" as Φ . In fact, this new Φ contains details about at least two relevant aspects $a \in A$ (i.e., "angina", "heart"). In particular, such details would probably score an higher average cumulative pertinence $R_{D, q, A}$ for q equal to why, because they are about causality. For instance, with the DoXpy algorithm presented in section 4.2, when using the FB pertinence estimator described in section 5.2, the archetypes with the best score are those related to causality (i.e., what effect scores 0.59, in what case, why and how score 0.57), while most of the others have a null score (i.e., who, when, etc.), and the Average DoX is 0.29.

This construction of DoX in terms of Carnap's central criteria of adequacy is the central theoretical contribution of this paper, because it allows us to move to the next step: implementing (in section 4.2) an algorithm to experimentally verify hypothesis 1.

4.1.3. The Average DoX

Despite all its good properties, DoX cannot help in judging whether some collections of information have higher degrees of explainability than others. This is due to the multidimensional nature of the estimated relevance of different archetypes.

This multidimensionality makes it hard to tell if a DoX score is greater than another. To overcome this issue, a mechanism is required for combining the set of pertinence scores composing DoX into a single score representing *explainability*. We do so by simply averaging pertinence scores. Hence, 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.

⁸The operator $\langle x, y \rangle$ is used here to represent tuples.

⁹Actually, 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.

Definition 5 (Average Degree of Explainability). *The Average DoX is the average of the pertinences 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, although this may not be necessarily true. In fact, as suggested in section 2.4 and in [36], 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 objectives of the stakeholders. Nonetheless, defining which (archetypal) question is the most relevant is clearly a challenging task and rather prone to subjectivity, therefore we believe that Average DoX is probably the only objective solution to this dispute.

4.2. The DoXpy algorithm

Given definition 4, we argue that it is possible to write an algorithm that can approximate a quantification of the Degree of Explainability of information representable with *natural language* (e.g., English).

Suppose we want to measure the DoX of a set of texts Φ called *explanandum support material*, containing correct content-giving textual information about an explanandum. For example, if the *explanandum* were “heart diseases”, there would be many aspects involved including “heart”, “stroke”, “vessel”, “diseases”, “angina”, “symptoms”, etc. Hence a reasonable *support material* for it would probably be a book describing all these aspects, or a set of web-pages (i.e., those published by the *U.S. Centers for Disease Control and Prevention*¹⁰), or any other kind of related corpus written in natural language.

As per 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 and the set D of details contained in Φ ;
- define a mechanism to identify D ;
- define the function p to compute the pertinence of an individual detail d to an archetypal question q_a .

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

¹⁰<https://www.cdc.gov>

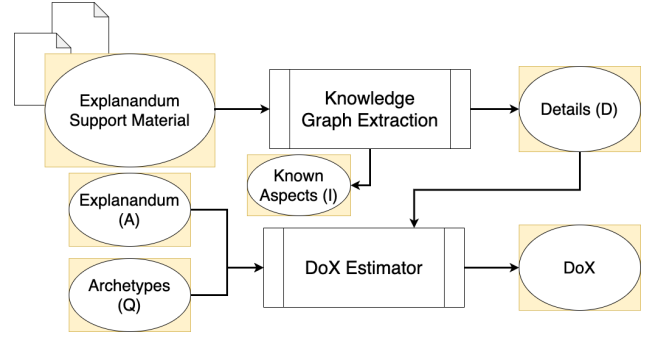


Figure 1: DoXpy Pipeline: The pipeline starts with the extraction of a graph from the *explanandum support material* Φ that is then converted into a set of details D . The set of details is then used in combination with the explanandum A and the set of archetypes Q to compute the DoX. To do so, we use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

Indeed, by leveraging on existing pre-trained deep language models [49, 66] capable of converting snippets of text (e.g., questions and answers) into numerical representations, in the following sections we show how to concretely implement DoXpy, an algorithm capable of estimating the DoX of any arbitrary piece of textual information with the pipeline shown in Figure 1. More precisely, this pipeline relies on a mechanism for the automatic extraction of a (knowledge) graph from any Φ , in order to identify the details needed to generate a score, as per definition 4.

For reproducibility purposes, we publish the DoXpy source code at: <https://github.com/Francesco-Sovrano/DoXpy>.

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 $a \in A$.

In particular, the set A of *explanandum aspects* is a collection of lemmatised words/syntagms. On the other hand, a detail d is a snippet of text with some specific characteristics. In fact, a detail d is an *information unit*, i.e., a relatively small sequence of words about one or more aspects (i.e., a sub-set of A) that is usually extracted from a more complex information bundle (i.e., a paragraph, a sentence) comprising several *information units*. In other terms, these details should carry enough information to describe different parts of an aspect a (possibly connected to many other aspects), so that we can use them to answer some (archetypal) questions about a and to correctly estimate a *level of detail*, as required by definition 4.

Considering the aforementioned characteristics of D and A , the most natural representation of them seems to be a (knowledge) graph. In fact, a graph is a set of nodes (i.e., A) connected by a set of edges (i.e., D). Therefore, we believe that the easiest way to identify the set of details D (and possibly also A) might pass through some mechanism for the extraction of a graph of *information units* from Φ .

Thus, an approach like the one used in [57, 55] for archetypal question answering, might be suitable to our ends, allowing for the identification of meaningful *information units* and suggesting also a mechanism for the estimation of *pertinence*. In particular, the algorithm proposed by [57, 55] consists in a pipeline of AI tools for the extraction from Φ of a graph of D and A designed for answer retrieval.

More specifically, this graph is extracted by detecting, with a dependency parser, all the phrases and sub-phrases within the *explanandum support material* that stand as an edge of the graph. In practice, these phrases are represented as special triplets of subjects, templates and objects called template-triplets. More specifically, the templates are composed by the ordered sequence of tokens connecting together 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 doesn’t 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 possible answer for retrieval by measuring the (cosine) similarity (or *pertinence* p) between its embedding (obtained through deep language models such as [23, 32]) and the embedding of a question q .

Importantly, as *information units*, Sovrano and Vitali [57, 55] use meaningful decompositions of grammatical dependency trees, to guarantee 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, so to avoid an over-estimation of the *level of detail*;
- an easy way to understand whether an answer is invalid when it is totally contained in the question, hence forcing its *pertinence* to be zero.

All these properties meet the requirements that a good detail $d \in D$ should possess for the generation of a DoX score. Therefore, this motivates our decision to use inside our pipeline the algorithm for answer retrieval from [57, 55]. In these papers you can also find additional technical details about how this algorithm works.

4.2.2. Selection of Archetypes

According to definition 1, an archetypal question is a generic question characterised by one or more interrogative formulas. Literature is full of different examples of such archetypal questions, and many of them are used to classify semantic relations (see for instance [24, 20, 40, 47], etc.).

Interestingly, it is possible to identify a sort of hierarchy or taxonomy of these archetypes, ordered by their intrinsic level of specificity. For example, the simplest interrogative formulas (made only of an interrogative particle: what, why, when, who, etc.) can be seen as the most generic archetypes. While the more complex and composite is the formula (e.g., what for, what cause), the more specific is the question. Hence, we decided to consider as set Q of main *archetypes* the most generic interrogative formulas used by literature [24, 20, 40, 47] to classify semantic relations within discourse.

In particular, the main *archetypes* coming from Abstract Meaning Representation theory [40] are: what, who, how, where, when, which, whose, why. We refer to these archetypes as the *primary* ones because they consist only of interrogative particles.

On the other hand, the main *archetypes* coming from PDTB-style discourse theory [47] (also called *secondary archetypes* because they make use of the *primary archetypes*) are: in what manner, what is the reason, what is the result, what is an example, after what, while what, in what case, despite what, what is contrasted with, before what, since when, what is similar, until when, instead of what, what is an alternative, except when, unless what.

In addition to the fact that 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 for any other question¹¹, whereas more specific questions can be always framed by using the interrogative particles we considered (e.g., why, what). In fact, *primary archetypes* can be used to represent any fact and abstract meaning [9], while the *secondary archetypes* can cover all the discourse relations between them (at least according to the PDTB theory).

5. Experiments

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

- a Heart Disease Predictor based on XGBoost [16] and TreeSHAP [38];

¹¹For concrete examples of how all these questions (especially the primary ones) are related to XAI algorithms, we point the reader to this recent survey by IBM Research [36] or to section 2.4 of this paper.

- a Credit Approval System based on a simple Artificial Neural Network and on CEM [18].

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 surely when a simple XAI-based system is considered. In fact, in a standard XAI-based system, the amount of *explainability* is (by design) clearly and explicitly dependent on the output of the underlying XAI, for the black-box not being explainable by nature. 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 the reason 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.

In fact, if hypothesis 1 is correct, the lower is the DoX score, the fewer explanations can be extracted, the less effective (as per ISO 9241-210) an explaineer 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”, and in our case it is measured through multiple-choice domain-specific quizzes.

Hence, once all the components that may affect effectiveness were fixed, including the explanandum and the presentation logic (i.e., the mechanism for re-elaborating explainable information into explanations), we expect that an increase in DoX always corresponds to an increase in effectiveness, at least on those tasks covered by the information provided by the increment of DoX. To show this, we borrowed and extended the results of two independent user-studies [57, 58], observing how DoX correlates with the effectiveness scores measured by these studies. Therefore, in the following sub-sections we present:

- the two XAI-based systems object of these experiments;
- the pertinence functions p and threshold t that we considered for computing the DoX scores and why;
- the sets A of *explanandum aspects* that were identified in each experiment.

5.1. XAI-Based Systems

The XAI-based systems we considered for the two experiments of this paper are a Credit Approval System and a

Heart Disease Predictor, respectively on finance and health-care topics. 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 fixed explanation for all users, together with the output of the wrapped AI, a few extra details to ensure the readability of the results, and a minimum of context.

5.1.1. Finance: Credit Approval System

The Credit Approval System is the same used also by [57, 58] and it has been designed by IBM to showcase AIX360¹². In particular, the explanandum of the Credit Approval System is about finance and the system is used by a bank. The bank deploys an Artificial Neural Network to decide whether to approve a loan request, and it uses CEM [18] to create post-hoc contrastive explanatory information. This information is meant to help the customers, showing them which minimal set of factors is to be manipulated for changing the outcome of the system from denial to approval (or vice versa).

The Artificial Neural Network behind the Credit Approval System was trained on the “FICO HELOC” dataset¹³, containing anonymized information about loan applications made by real homeowners. Importantly, the Artificial Neural Network is trained to answer the following question: “What is the decision on the loan request of applicant X?”.

Given the specific characteristics of the Credit Approval System, it is possible to assume that the main goal of its users is to understand what are the causes behind a loan rejection and what to do for getting a loan accepted. This is why CEM is deployed to answer the following questions:

- What are the easiest factors to consider in order to change the result of the application of applicant X?
- How factor F should be modified in order 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 to answer before a user of the system can be satisfied, reaching its goals. These questions include: “How to perform those minimal actions?”, “Why are these actions so important?”, etc.

Finally, to summarise, the output of the Credit Approval System is composed by:

- Context: a titled heading section kindly introducing Mary (the user) to the system.
- AI Output: the decision of the Artificial Neural Network for the loan application. This decision normally can be “denied” or “accepted”. For Mary it is: “denied”.

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

¹³<https://fico.force.com/FICOCommunity/s/explainable-machine-learning-challenge?tabset=3158a=a4c37>

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

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

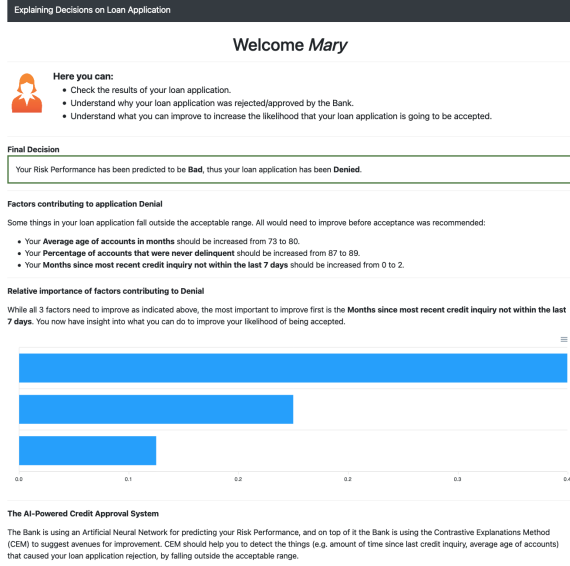


Figure 2: Screenshot of the Credit Approval System

5.1.2. Health: Heart Disease Predictor

Similarly to the Credit Approval System, also the Heart Disease Predictor comes from [58]. 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 right physician in the eventuality of a reasonable risk of heart disease. First level responders get basic questions from callers, they are not doctors but they have to decide on the fly whether the caller should speak to a real doctor or not. So, they quickly use the Heart Disease Predictor to figure out what to answer to the callers and what are the next actions to suggest. In other words, this system is used directly by the responder, and indirectly by the caller through the responder. These two types of users have different but overlapping goals and objectives. It is reasonable to assume that the goal of the responders is to answer in the most efficient and effective way the questions of a caller.

The Heart Disease Predictor uses XGBoost [16] to predict the likelihood of a patient having a heart disease given its demographics (gender and age), health (diastolic blood pressure, maximum heart rate, serum cholesterol, presence of chest-pain, etc.) and the electrocardiographic (ECG) results. This likelihood is classified into 3 different risk areas: low (probability p of heart disease below 0.25), medium

($0.25 < p < 0.75$) or high. Therefore, XGBoost is used to answer the following questions:

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

More specifically, the dataset used to train XGBoost is the “UCI Heart Disease Data” [17, 2].

On top of XGBoost, the Heart Disease Predictor uses TreeSHAP [38], a famous XAI algorithm specialised on tree ensemble models (i.e., XGBoost) for post-hoc explanations. In particular, TreeSHAP is used to understand what is the contribution of each feature to the output of XGBoost. Therefore, TreeSHAP is used to answer the following questions:

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

Nonetheless, many other important questions should probably be answered, including: “What is the easiest thing that the patient could actually do to change his heart disease risk from medium to low?”, “How could the patient avoid raising one of the factors, preventing his heart disease risk to raise?”, etc.

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 inserting the patient’s parameters.
- **AI Outputs:** a section displaying the likelihood of heart disease estimated by XGBoost and a few generic suggestions about the next actions to suggest.
- **XAI Outputs:** a section showing the contribution (positive or negative) of each parameter to the likelihood of heart disease, generated by TreeSHAP.

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

5.2. Pertinence Functions and Thresholds

In order to compute DoX, according to definition 4, we need to define a pertinence function p and pick a threshold t . As discussed in section 4.2, we are interested in using as pertinence function p a deep language model for answer retrieval. Though, the point is that many different deep

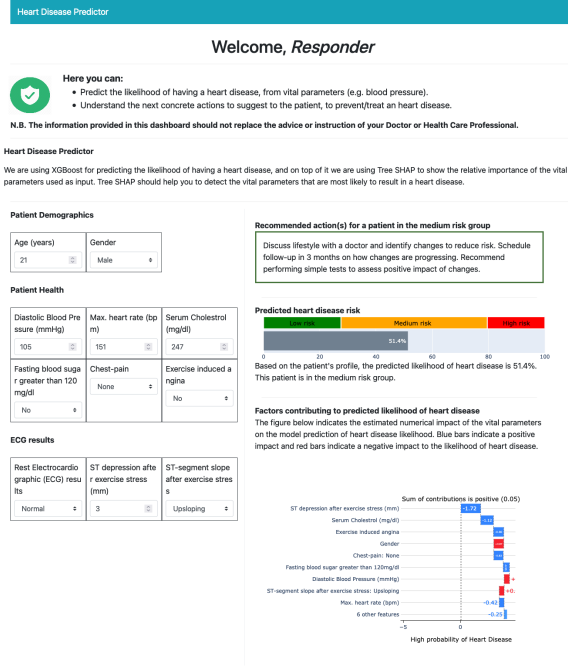


Figure 3: Screenshot of the Heart Disease Predictor

language models exist for this task, i.e. [23, 55, 32], 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, during our experiments we decided to study the behaviour of more than one deep language model as pertinence function p . Hence the models we considered are:

- FB: published by [32] and [49]. FB has been trained on the combination of the following datasets: Natural Questions [34], TriviaQA [31], WebQuestions [6], and CuratedTREC [5].
- TF: or Multilingual Universal Sentence Encoder [66]. TF has been trained on the Stanford Natural Language Inference corpus [10].

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

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

The 1st experiment is meant to shed more light on how a few changes to the explainability of a system affect the estimated DoX. Specifically, XAI-based systems are considered for this experiment, instead of other AI-based systems, because their amount of *explainability* is by design, clearly and explicitly dependent on the output of the underlying XAI. So that, by masking the output of the XAI, the overall

system can be forced to be less explainable. Hence, this characteristic can be exploited to (at least partially) verify hypothesis 1, in a very simple but effective way.

In other words, a XAI-based system is composed by 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 result, we expect the (average) DoX of the XAI-based explainer to be clearly higher than that of its wrapped AI-based system.

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 we obtain the *AI-based explainers* we need.

In order to compute the (average) DoX of these systems, we take as 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 [16] and TreeSHAP [38] in the Heart Disease Predictor are 5:

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

While the main *explanandum aspects* A targeted by the Artificial Neural Network and CEM [18] in the Credit Approval System are 4:

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

After properly converting the images produced by the *XAI-based explainers* to textual explanations, the resulting *explanandum aspects coverage*¹⁴ of both the Heart Disease Predictor and the Credit Approval System is 100%, while that of their *AI-based explainers* is respectively 60% and 50%.

In particular, for estimating the DoX we used the pipeline described in section 4.2, extracting different sets of details D from the textual representations of both the XAI-based and *AI-based explainers*. Thus, we were able to compute the DoX scores in accordance with definition 4 by using:

- the set of archetypes Q described in section 4.2.2,

¹⁴The *explanandum aspects coverage* is the percentage of *explanandum aspects* that are covered by explanations.

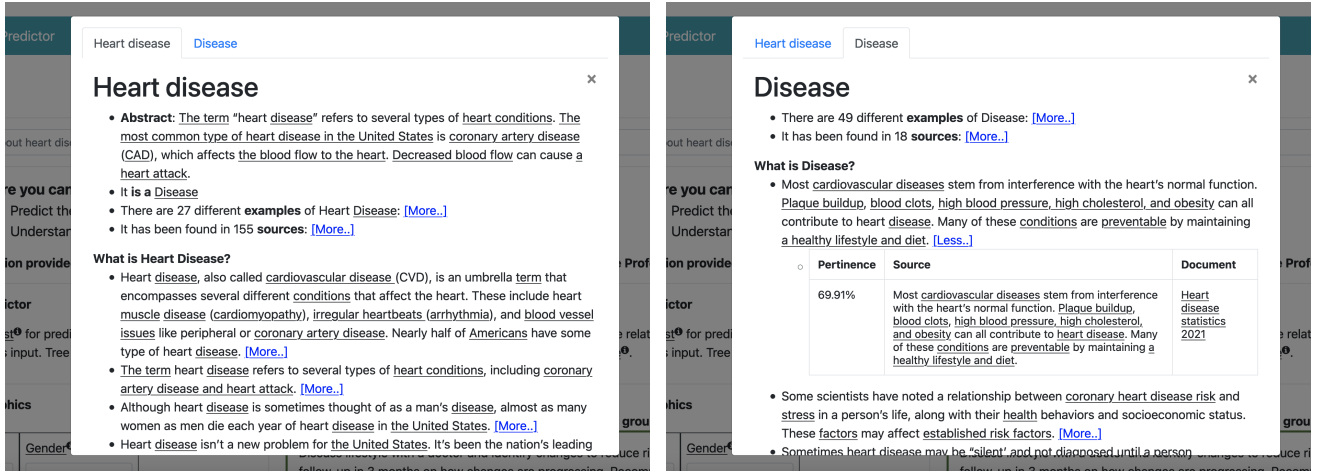


Figure 4: Example of Overview: this figure shows an example of interactive *overview* displaying relevant information about important concepts for the Heart Disease Predictor. Clicking on any underlined word would open a new *overview* in a new tab, as shown. Furthermore, every given answer is linked to its source document.

- the pertinence functions p and the thresholds t presented in section 5.2,
- the aforementioned set of details D and *explanandum aspects* A .

Results are presented and discussed in section 6.

5.4. 2nd Experiment: Explainability vs Effectiveness

Unlike the first, this second experiment aims to understand if there is a correlation between DoX and the effects of explainability on the explainees. In fact, we have that more explainability implies a greater ability to explain, therefore more explanations. In short, the lower the DoX, the less explanations can be produced, the less effective is likely to be an explainee on the tasks related to the explanandum.

So, if hypothesis 1 would be correct, an increase in the DoX of the (explanatory) system should 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 and extended the user-studies published by [57, 58] and involving more than 160 human subjects.

Importantly, these user-studies consider the same XAI-based systems used during the first experiment and described throughout section 5.1. In particular, each user-study analyses the effectiveness of the explanations given by the XAI-systems, when changing the *explanandum support material* and the way it is presented to the explainee. In the following sub-sections we will present the aforementioned user-studies more in detail.

5.4.1. 1st User-Study: Finance

The first user-study comes from [57]. [57] present a novel mechanism, that we call *overview-based explainer*, to explain large collections of heterogeneous documents (i.e., more than 50 web-pages) about the Credit Approval

System, in a user-centred and interactive way. This is done by organising knowledge as a graph of abstract aspects whose related explanations are ordered by relevance and simplicity, according to a set of pre-defined archetypal questions (i.e., what, how, when, why, etc.). In particular, *overviewing* can be performed iteratively from an initial explanation by clicking on annotated words for which an explanation (in the form of a cluster of questions and answers) is needed. An example of *overview* is shown in figure 4.

To show that the *overview-based explainer* generates more effective explanations, the authors compare the effectiveness scores of the Credit Approval System with a version of it enhanced by the *overview-based explainer*. In particular, the effectiveness scores are generated by the users interacting with the system and answering a quiz (shown in table 3) on the Credit Approval System comprising 7 different questions.

For this user-study 103 different participants were recruited (57 males, 44 females, 2 unknowns; ages 18-55) on the online platform Prolific [45]. 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%¹⁵. Participants were randomly allocated to use either the Credit Approval System or its *overview-based explainer*, in a between-subjects test. In the end, 51 participants evaluated the normal *XAI-based explainer* and 52 evaluated the *overview-based explainer*¹⁶.

Importantly, the results of the user-study showed that the *overview-based explainer* produces significantly better effectiveness statistics than its counterpart, especially on those questions (questions number 2, 3, 4, 5 and 7 in table 3) not covered by the *explanandum support material* of the normal *XAI-based explainer*, as shown in figure 5.

¹⁵Mainly because they are unlikely to answer poorly/randomly to questions.

¹⁶For more details about the evaluation, please read [57].

Table 3

Quiz - Credit Approval System: in this table are shown the questions used for the quiz on the Credit Approval System and their archetypes. The “Archetype” column indicates which interrogative particle is representative of the question. *Steps* is the minimum number of steps (in terms of links to click, overviews to open and/or questions to pose) required by each explanatory tool. “NXE” stands for the normal *XAI-based explainer*, while “Others” stands for the *overview-based explainer* or its extensions described in section 5.4.2. Negative *steps* mean that the correct answer cannot be found, while 0 *steps* means that the answer is immediately available without clicking on any link.

| Question | Archetype | Steps | |
|--|-----------|-------|--------|
| | | NXE | Others |
| 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 |

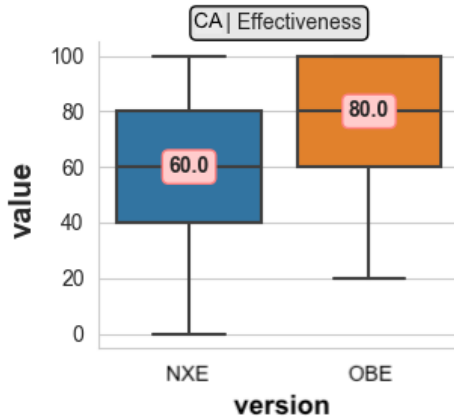


Figure 5: 2nd Experiment - 1st User-Study: Comparison of the median effectiveness scores obtained on the Credit Approval System (CA) with the normal *XAI-based explainer* (NXE; the blue one) and the *overview-based explainer* (OBE; the orange one) on those questions whose answer is not contained in the *explanandum support material* of NXE. Results are shown in the form of box plots (25th, 50th, 75th percentile, and whiskers covering all data and outliers). The numerical value of medians is shown inside pink boxes. Differently from [57], here effectiveness scores are normalised in [0, 100].

Indeed, the difference between a normal *XAI-based explainer* and an *overview-based explainer* is twofold. First of all, the explanations produce by the *overview-based explainer* are interactive and more user-centred, while those

of the normal *XAI-based system* are not. Secondly, the normal *XAI-based explainer* considers a smaller amount of explainable information. In fact, the *overview-based explainer* builds its explanations using more than 50 extra web-pages that its counterpart does not see.

This last difference is exactly what allowed us to exploit the user-study, to verify hypothesis 1. In fact, the amount of information handled by the normal *XAI-based explainer* is roughly $\frac{1}{100}$ of its *overview-based explainer* and this is enough to say that its explainability is not none and it is probably the lowest amongst the two explainers.

In order to use this first user-study, to show that an increment in DoX causes a consequent increment in the effectiveness of the explanations, we have to compute the DoX scores of a normal *XAI-based explainer* and its *overview-based explainer* as in the 1st experiment. To do so, we identified the set of *explanandum aspects* A from the quiz used to generate the effectiveness scores, applying on it the same technique for extracting the set of details D mentioned in section 4.2. In fact, the quiz defines exactly what the users should know in order to be effective, indirectly defining also what is important for the system to explain: the *explanandum aspects*. The resulting DoX scores are given in section 6.

5.4.2. 2nd User-Study: Health

The second user-study comes from [58]. Unlike the first though, this user-study is on the Heart Disease Predictor and it analyses an extension of the *overview-based explainer* called YAI4Hu, together with other two explainers: a *two-layered explainer* and a *how-why explainer*.

The *two-layered explainer* is static, as the normal *XAI-based explainer*, in fact it is made of the output a normal *XAI-based explainer* directly connected to a 2nd (non-expandable) layer of information consisting in an exhaustive and verbose set of autonomous static explanatory resources. The *two-layered 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, differently from that, it uses only the archetypes *why* and *how* for generating explanations. Furthermore, also YAI4Hu is an extension the *overview-based explainer* that instead adds a mechanism (called *open questioning*) for users to ask their own questions to the system. More specifically, *open questioning* can be performed asking questions in English through a search box that uses the same graph-based answer retrieval mechanism described in section 4.2. Importantly, YAI4Hu, the *how-why explainer* and the *two-layered explainer* share the same *explanandum support material* of the *overview-based explainer* (see section 5.4.1 for more details).

For the user-study, [58] recruited 64 different participants amongst the university students of the following

Table 4

Quiz - Heart Disease Predictor: in this table are shown the questions used for the quiz on the Heart Disease Predictor and their archetypes. NXE stands for the normal *XAI-based explainer*, HWN for the *how-why explainer*, and 2EC for the *two-layered explainer*. Furthermore, “no OQ” means that *open questioning* cannot be used for correctly answering the question. For more details about how to read this table, see the caption of table 3.

| Question | Archetype | Steps | | | |
|--|----------------|-------|-----|-----|-----------|
| | | NXE | 2EC | HWN | YAI4Hu |
| What are the most important factors leading that patient to a medium risk of heart disease? | what, why | 0 | 0 | 0 | 0 (no OQ) |
| What is the easiest thing that the patient could actually do to change his heart disease risk from medium to low? | what, how | 0 | 0 | 0 | 0 (no OQ) |
| According to the predictor, what level of serum cholesterol 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 to shift from medium to high? | how | -1 | 1 | 2 | 2 |
| What kind of 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 to 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 |

courses of study¹⁷: bachelor degree in computer science; bachelor degree in management for informatics; master degree in digital humanities; master degree in artificial intelligence. The 64 participants were randomly allocated to test only one of the three types of explainers. In other words, similarly to the first user-study, also this second user-study followed a between-subjects design. In the end, there were approximatively 20 participants per explainer.

Each participant evaluated the effectiveness of the XAI-based system by answering a quiz shown in table 4. For more details about the evaluation, please read [58].

The results of the user-study published by [58] show that YAI4Hu produces significantly greater effectiveness scores than the *how-why explainer*, and that the *how-why explainer* is better than the *two-layered explainer* as well. Though, considering that YAI4Hu, the *how-why explainer* and the *two-layered explainer* share the same *explanandum support material* (of the *overview-based explainer*), these results tell

¹⁷All the courses of study were of an Italian university, and only the master degrees were international, i.e., with English teachings and students from countries other than Italy.

us only that changing the explainer might change the quality of the explanations.

In other words, differently from the first user-study, these results do not show any improvement in effectiveness due to changes in the explainability of the *explanandum support material*. Considering that, we decided to extend the user-study recruiting 19 more participants¹⁸ from the same pool of users, asking them to answer the same quiz but with a normal *XAI-based explainer*. Indeed, the normal *XAI-based explainer* has a smaller *explanandum support material* than the others.

After this modification, also the second user-study can be used to check whether there is a correlation between the DoX scores and the perceived effectiveness score. That is because we can compare the scores of the normal *XAI-based explainer* against the others, in a situation where it is possible to study the effects on effectiveness of different *explanandum support materials*. Hence, to do this comparison, similarly to the first user-study, we extracted the set of *explanandum aspects A* needed for computing the DoX scores from the evaluation quizzes. As result we were able to identify 82 *explanandum aspects* for the Heart Disease Predictor.

6. Results

In this section we present and discuss the results of the experiments defined in section 5.

6.1. 1st Experiment

Computing the DoX scores for the first experiment (described in section 5.3), we got the results displayed in table 5. As expected, on both the XAI-based systems, the results of the first experiment neatly show that the (average) DoX achieved by the normal *XAI-based explainer* is way greater than its *AI-based explainer*, regardless the adopted *deep language model*. Although, as discussed in section 7, we can see that TF and FB (the two adopted *language models*) produce numerically different DoX scores, suggesting that the choice of the pertinence function p can sensibly impact on the value of DoX.

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

¹⁸We made sure that none of our 19 extra participants was involved in the user-study published by [58].

¹⁹This table is different from that shown in [59] because we used DoXpy v2.1 instead of DoXpy v1.0.

Table 5

1st Experiment - Degree of Explainability¹⁹: in this table DoX and Average DoX are shown for the Credit Approval System (CA) and the Heart Disease Predictor (HD). As columns we have the different explanatory mechanisms used for experiment 1: the normal *AI-based explainer* (NAE) and the normal *XAI-based explainer* (NXE). As rows we have different explainability estimates using different deep language models for computing pertinence: FB and TF. For simplicity, with DoX we show only the *primary archetypes*.

| | | CA | | HD | |
|---------|----|-------------|-------------|-------------|-------------|
| | | NAE | NXE | NAE | NXE |
| Avg DoX | FB | 0.607 | 0.659 | 0.596 | 0.609 |
| | TF | 0.264 | 0.344 | 0.214 | 0.227 |
| DoX | FB | which: 0.63 | which: 0.66 | which: 0.6 | which: 0.63 |
| | | whose: 0.63 | whose: 0.66 | what: 0.6 | what: 0.62 |
| | | how: 0.6 | how: 0.66 | whose: 0.59 | whose: 0.62 |
| | | when: 0.6 | where: 0.66 | why: 0.59 | how: 0.62 |
| | | what: 0.6 | what: 0.66 | how: 0.59 | why: 0.61 |
| | | why: 0.6 | who: 0.66 | where: 0.58 | where: 0.61 |
| | | where: 0.59 | when: 0.65 | who: 0.57 | when: 0.59 |
| | | who: 0.58 | why: 0.65 | when: 0.56 | who: 0.58 |
| | TF | what: 0.33 | what: 0.4 | why: 0.31 | what: 0.28 |
| | | when: 0.31 | which: 0.37 | what: 0.28 | why: 0.28 |
| | | which: 0.26 | where: 0.37 | when: 0.24 | which: 0.25 |
| | | why: 0.25 | how: 0.36 | whose: 0.22 | when: 0.24 |
| | | whose: 0.24 | why: 0.36 | how: 0.21 | how: 0.22 |
| | | where: 0.23 | when: 0.35 | who: 0.21 | who: 0.21 |
| | | who: 0.23 | who: 0.33 | where: 0.2 | whose: 0.21 |
| | | how: 0.2 | whose: 0.31 | which: 0.2 | where: 0.2 |

effectiveness of the resulting explanations on humans, as shown in section 6.2.

6.2. 2nd Experiment

With the second experiment we studied if there is a correlation between the effectiveness of the explainers described in section 5.4 and their DoX scores. We do it by considering two different user-studies and by comparing the effectiveness scores of different explainers. The 1st user-study comes from [57], while the 2nd is an extension to the user-study presented by [58]. The approach we followed to extend this latter study is described in section 5.4.2.

In particular, the results of the 1st user-study (summarised in figure 5) show that the *overview-based explainer* is statistically more effective than the normal *XAI-based explainer*, on the Credit Approval System. In fact, according to a one-sided Mann-Whitney U-Test (a non-parametric version of the t-test for independent samples) there is enough statistical evidence²⁰ to claim that the *overview-based explainer* is better ($U = 849.5$, $P = 0.007$) than the *XAI-based explainer* on the questions covered by the increment in DoX of the *overview-based explainer* (questions number 2, 3, 4, 5 and 7 in table 3). This is true at least on the considered pool of subjects. For more details about these results, please read [57].

On the other hand, the results we got by extending the 2nd user-study are summarised in the box-plots of figure 6. In particular, these box-plots show the effectiveness score of the considered explainers on the questions that the normal

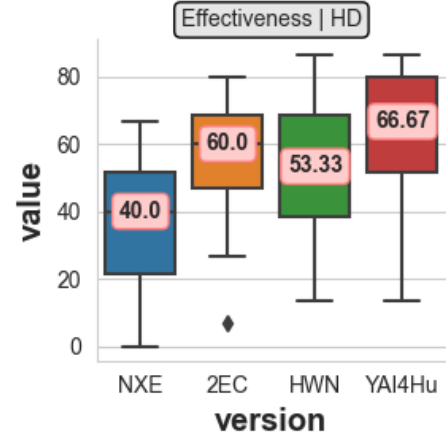


Figure 6: 2nd Experiment - 2nd User-Study - Effectiveness Scores on Questions *not covered* by the XAI-based explainer: Comparison of the results achieved on the Heart Disease Predictor (HD) with the normal *XAI-based explainers* (NXE) and the other explainers, only on those questions whose aspects are *not covered* by the information presented by NXE. In particular, the other explainers are: the *two-layered explainer* (2EC), the *how-why explainer* (HWN) and YAI4Hu. Effectiveness scores are normalised in [0, 100].

XAI-based explainer cannot answer (i.e., the questions with negative steps in the NXE column of table 4).

As expected, also in this case we see the median effectiveness score of the normal *XAI-based explainers* being significantly lower than the other explainer. More precisely, according to a one-sided Mann-Whitney U-Test ($U = 40.0$, $P = 0.0002$) there is enough statistical evidence to claim that YAI4Hu is better than the *XAI-based explainer* on the questions covered by the increment in DoX of the *overview-based explainer*. Furthermore, there is also enough statistical evidence to claim that the *two-layered explainer* ($U = 48.0$, $P = 0.003$) and the *how-why explainer* ($U = 65.5$, $P = 0.02$) are significantly more effective than the normal *XAI-based explainer*.

Now, if hypothesis 1 is true, we would expect that the higher is DoX, the higher is the effectiveness of an explainer. Importantly, the opposite is not necessarily true, in fact, two explainers (with different presentation logics; e.g., the *two-layered explainer* and YAI4Hu) might have different effectiveness scores despite having the same DoX.

Computing the DoX scores for the second experiment we got the results shown in Table 6. These results confirm our expectations for them. In fact, they show that the *two-layered explainer*, the *overview-based explainer*, the *how-why explainer* and YAI4Hu have higher DoX scores than the normal *XAI-based explainer*.

7. Discussion

The results presented in section 6 clearly show that DoX behaves as expected, roughly confirming hypothesis 1. Indeed, we see that DoX grows whenever a black-box AI is

²⁰A $P < 0.05$ is normally considered to be a significant statistical evidence.

Table 6

2nd Experiment - Degree of Explainability: these scores are different from those of experiment 1 (table 5) because a different explanandum is considered for experiment 2. As columns we have the different explanatory mechanisms used for experiment 2: the normal *XAI-based explainer* (NXE) and the “others”. For more details about how to read this table, refer to the caption of table 5.

| Avg DoX | | CA | | HD | |
|---------|----|-------------|-------------|-------------|-------------|
| | | NXE | Others | NXE | Others |
| DoX | FB | 0.43 | 2.066 | 0.936 | 2.465 |
| | TF | 0.163 | 0.636 | 0.504 | 1.708 |
| | FB | how: 0.58 | which: 2.42 | which: 1.18 | what: 3.95 |
| | | why: 0.58 | whose: 2.37 | how: 1.17 | how: 3.42 |
| | | what: 0.57 | how: 2.36 | what: 1.16 | which: 2.95 |
| | | whose: 0.56 | why: 2.35 | whose: 1.15 | why: 2.52 |
| | | which: 0.29 | what: 2.34 | why: 0.9 | whose: 2.26 |
| | | where: 0.28 | where: 2.24 | when: 0.89 | when: 2.23 |
| | | when: 0.27 | when: 1.71 | where: 0.89 | who: 1.95 |
| | | who: 0.27 | who: 1.71 | who: 0.88 | where: 1.91 |
| | TF | what: 0.29 | when: 1.04 | why: 0.57 | what: 2.19 |
| | | why: 0.23 | what: 1.02 | what: 0.56 | how: 1.77 |
| | | when: 0.22 | why: 0.97 | which: 0.49 | why: 1.76 |
| | | who: 0.21 | which: 0.93 | how: 0.48 | when: 1.76 |
| | | whose: 0.2 | how: 0.72 | when: 0.48 | which: 1.62 |
| | | how: 0.18 | who: 0.65 | where: 0.48 | whose: 1.59 |
| | | which: 0.13 | whose: 0.65 | who: 0.48 | who: 1.58 |
| | | where: 0.09 | where: 0.62 | whose: 0.47 | where: 1.25 |

wrapped within a XAI (as it should be) and that an increment of DoX corresponds to a statistically meaningful increase in the effectiveness of the explanatory system.

In particular, the results of the 1st 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 2nd experiment. Furthermore, the results of the 2nd 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-layered explainer*, despite the fact that it is not interactive and it does not re-organise information to make it simpler and easier to access, dumping on the user hundreds of pages of content.

Our user-studies involved more than 160 participants and were consistent across two fairly different and broad user pools, producing statistically significant results (hence with P values lower than 0.05). Therefore, considering that *explainability* is fundamentally the *ability to explain*, the two experiments combined together 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 proxy for measuring the explainability of an explanatory system, as long as a set of *explanandum aspects* can be defined. In fact, DoX is deterministic and fully 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 on all applications where it is important to objectively evaluate explainability. Indeed, the main benefit of DoX is the fact that it

works with any set of *explanandum aspects A*, and 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 as identified in [8].

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 simply 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 [60].
- *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.

In this sense, the benefits of using DoX over a normal user-study are manifold, in fact:

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

In other terms, 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 happened to be 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” [13].

8. Limitations and Future Work

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 2nd 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, so as to understand how many questions can be properly answered: the higher DoX, the greater the chances to properly explain to a variety of users.

Secondly, the numerical differences between the DoX estimates shown in table 5 and 6 suggest that our algorithm for computing DoX scores may be sensitive to the choice of deep language model for pertinence estimation (see section 5.2). 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 TF to FB. Nonetheless, we also see that in all the considered experiments the DoX scores increase as expected, with both TF and FB, suggesting that the alignment of DoX to *explainability* is independent from the chosen deep language model. This intuition is supported by the fact that both TF and FB, in 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 in similar ways on standard benchmarks may produce proportionally similar Average DoX scores. This does not exclude the fact that there might be deep language models that are better than others for computing the DoX score, or that multiple standardised 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 normalised 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 is able to measure the similarity of the *explanandum support material* to the explanandum. Unfortunately, it is not possible to know in advance the total number of details of any possible *explanandum support material* and therefore it is not possible to normalize the score by dividing the *cumulative pertinence* by such number. It is worth noting that such sum is necessary. Indeed, if the *cumulative pertinence* were a mean instead of a sum, then the resulting score for an *explanandum support material* could not be compared to that of any larger (in terms of number of details) *explanandum support material*, making pointless the use of DoX in the first place.

Furthermore, it is important to mention that DoX, alone, is not sufficient for a thorough quantification of how much of the information is explained by an AI. In fact, our definition of DoX does not take into consideration the correctness of

information of the *explanandum support material*, assuming that truth is given and that it is a different thing from explainability. In other terms, DoX should always be used in combination with other metrics that describe how correct the available information is.

Finally, as discussed in section 7, although DoX can be used to verify many of the requirements defined by [8], it is still unclear how to apply DoX to verify also Government-to-Citizen legal requirements. In fact, being able to select a reasonable threshold of DoX scores for law-compliance is certainly one of the next challenges we envisage for a proper standardisation of *explainability* in the industrial context. We also leave these analyses for future work.

9. Conclusions

In this paper, we proposed a new metric for explainability called DoX that could be used to 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. [8].

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 over the concepts being explained, so that the more (archetypal) answers a corpus is able to give about important aspects of an explanandum, the more that corpus is explainable. Thus DoX is the first explainability metric based on Ordinary Language Philosophy and 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 actually behaving as expected, we designed a few experiments on two realistic AI-based systems for heart disease prediction and credit approval, involving state of the art AI technologies such as Artificial Neural Networks, TreeSHAP [38], XGBoost [16] and CEM [18]. The results we obtained show that the DoX is aligned with our expectations, and that it 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 XAIs or any other explainable information (i.e., documentation, papers, books, etc.) to understand how that information can be used to explain. In this sense, DoX is the most useful when used to evaluate large collections of explainable information (i.e., the output of an ensemble of XAIs).

Another context of application of DoX could be education. Not surprisingly, many would argue that explanations are one of the main artefacts through which humans understand reality and learn to solve complex problems [7]. Therefore, *explaining* is not only central to XAI but also to education, and these are two contexts where our technology and our 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., Bénéto, 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] 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.
- [6] 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/>.
- [7] Berland, L.K., Reiser, B.J., 2009. Making sense of argumentation and explanation. *Science Education* 93, 26–55.
- [8] 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.
- [9] 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.
- [10] Bowman, S.R., Angeli, G., Potts, C., Manning, C.D., 2015. A large annotated corpus for learning natural language inference, in: Márquez, 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.
- [11] 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.
- [12] Brun, G., 2016. Explication as a method of conceptual re-engineering. *Erkenntnis* 81, 1211–1241. doi:10.1007/s10670-015-9791-5.
- [13] Bućinca, 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.
- [14] Carnap, R., Schilpp, P.A., 1963. The Philosophy of Rudolf Carnap. Cambridge University Press Cambridge.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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>.
- [19] 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.
- [20] 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.
- [21] 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>.
- [22] 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.
- [23] Guo, M., Yang, Y., Cer, D., Shen, Q., Constant, N., 2021. MultiReQA: A cross-domain evaluation for Retrieval 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>.
- [24] He, L., Lewis, M., Zettlemoyer, L., 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language, in: Márquez, 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.
- [25] Hempel, C.G., Oppenheim, P., 1948. Studies in the logic of explanation. *Philosophy of Science* 15, 135–175. doi:10.1086/286983.
- [26] 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>.
- [27] 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.
- [28] 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=Z6EFBaApE8C>.
- [29] Holzinger, A., Carrington, A.M., Müller, H., 2020. Measuring the quality of explanations: The system causability scale (SCS). *Künstliche Intell.* 34, 193–198. URL: <https://doi.org/10.1007/s13218-020-00636-z>, doi:10.1007/s13218-020-00636-z.
- [30] 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/>.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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/tac1_a_00276, doi:10.1162/tac1_a_00276.
- [35] 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.
- [36] 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., Goguy, A., Bjørn, 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.
- [37] 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.
- [38] 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.
- [39] Madumal, P., Miller, T., Sonenberg, L., Vetere, F., 2020. Explainable reinforcement learning through a causal lens, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press. pp. 2493–2500. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/5631>.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] 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.
- [45] 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.
- [46] 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.

- [47] 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.
- [48] 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.
- [49] 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.
- [50] Ribera, M., Lapedriza, A., 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>.
- [51] 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.
- [52] 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.
- [53] 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>.
- [54] Sellars, W., 1963. Science, Perception and Reality. New York: Humanities Press.
- [55] 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.
- [56] Sovrano, F., Sapienza, S., Palmirani, M., Vitali, F., 2022. Metrics, explainability and the european ai act proposal. J 5, 126–138. URL: <https://www.mdpi.com/2571-8800/5/1/10>, doi:10.3390/j5010010.
- [57] 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.
- [58] Sovrano, F., Vitali, F., 2022a. Generating user-centred explanations via illocutionary question answering: From philosophy to interfaces. ACM Trans. Interact. Intell. Syst. URL: <https://doi.org/10.1145/3519265>, doi:10.1145/3519265. just Accepted.
- [59] Sovrano, F., Vitali, F., 2022b. How to quantify the degree of explainability: Experiments and practical implications, in: 31th IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2022, Padova, July 18-23, 2022, IEEE. pp. 1–9.
- [60] 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.
- [61] Szymanski, M., Millicamp, 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.
- [62] 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.
- [63] 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.
- [64] 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.
- [65] 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.
- [66] Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Ábrego, G.H., Yuan, S., Tar, C., Sung, Y., Strophe, 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.