How to Quantify the Degree of Explainability: Experiments and Practical Implications

Francesco Sovrano
DISI
University of Bologna
Bologna, Italy
0000-0002-6285-1041

Fabio Vitali
DISI
University of Bologna
Bologna, Italy
0000-0002-7562-5203

***Abstract*—Explainable AI was born as a pathway to allow humans to explore and understand the inner working of complex systems. Though, establishing what *is* an explanation and *objectively* evaluating *explainability*, are not trivial tasks. With this paper, we present a new model-agnostic metric to measure the Degree of Explainability of (correct) information in an *objective* way, exploiting a specific theoretical model from Ordinary Language Philosophy called the *Achinstein's Theory of Explanations*, implemented with an algorithm relying on deep language models for knowledge graph extraction and information retrieval. In order to understand whether this metric is actually behaving as *explainability* is expected to, we have devised an experiment on two realistic Explainable AI-based systems for *healthcare* and *finance*, using famous AI technology including Artificial Neural Networks and TreeSHAP. The results we obtained suggest that our proposed metric for measuring the Degree of Explainability is robust on several scenarios.**

***Index Terms*—Objective Explainability Metric, XAI, Degree of Explainability,**

I. Introduction

The ability of humans to understand and appreciate the outputs and behaviours of autonomous systems is complicated by the increasing complexity and unpredictability of the most recent and advanced systems. In fact, Automated Decision-Making systems are changing our society, so people and governments (e.g., EU's, California, etc.) have begun to be concerned about the impact that they may have on our lives. This concern gave birth to the so-called *Right to Explanation*, which was introduced in the EU legislation within the General Data Protection Regulation (GDPR), and further explored by the High-Level Expert Group on Artificial Intelligence [1], established in 2018 by the EU Commission.

As a result, the EU indirectly posed an interesting challenge to the Explainable AI (XAI) community, by demanding more transparent, user-centred, and accountable approaches to Automated Decision-Making systems that guarantee explainability of their working. More precisely, the GDPR art. 35 requires data controllers to prepare a Data Protection Impact Assessment for operations that are "likely to result in a high risk to the rights and freedoms of natural persons". To this end, Algorithmic Impact Assessment can be intended as an instrument for ensuring certain minimal criteria [2] of explainability in Automated Decision-Making systems, serving as an important "suitable safeguard" (Article 22) of individual rights.

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

In this paper, we propose a new model-agnostic approach and metric to *objectively* evaluate explainability, through knowledge graph extraction, in a manner that is mainly inspired by Ordinary Language Philosophy instead of Cognitive Science. Our approach is based on a specific theoretical model of explanation, called the *Achinstein's theory of explanations*, where explanations are the result of an *illocutionary* (i.e., broad yet pertinent and deliberate) act of pragmatically answering 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 explanandum, the more it is explainable.

Therefore, we assert that it is possible to quantify the degree of explainability of a set of texts by applying the Achinstein-based definition of explanation proposed in [3]. Thus, drawing also from Carnap's criteria of adequacy of an explication [4], we frame the Degree of Explainability (DoX) as the average *Explanatory Illocution* of information on the *Explanandum Aspects*[2]. More precisely, we hereby present an algorithm for measuring DoX by means of pre-trained *language models* for general-purpose question answer retrieval, as [5, 6], applied to a special knowledge graph of triplets automatically extracted from text to facilitate this type of information retrieval.

As proof of concept, we performed an experiment with the objective of showing that *explainability* changes in accordance with DoX. To conduct the experiment we considered two different XAI-based systems, respectively for the healthcare and finance domains:

- A *Heart Disease Predictor* based on XGBoost [7] and TreeSHAP [8].

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

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

- A *Credit Approval System* based on a simple Artificial Neural Network and on CEM [9].

Our experiment follows a *direct* approach, comparing the DoX of the XAI-based systems with their non-explainable counterpart. This approach is said to be *direct*, because the amount of *explainability* of a XAI-based system is, by design, clearly and explicitly dependent on the output of the underlying XAI. Therefore, by filtering away the XAI's output, the overall system can be forced to be not explainable enough, by construction. For guaranteeing the reproducibility of the experiments, we published [3] the source code of DoX, as well as the code of the XAI-based systems, the user-study questionnaires and the remaining data mentioned within this paper.

II. Related Work

Being able to measure the quality of XAI tools is pivotal for claiming technological advancements, understanding existing limitations, developing better solutions and delivering XAI that can go into production. Not surprisingly, every good paper proposing a new XAI algorithm comes with some evidence or experiments to back up the underlying claims, usually relying on *ad hoc* or subjective mechanisms to measure the quality of explainability. In other words, it is very common to encounter explainability metrics that can work only with specific XAI models or that require to collect opinions/results generated by human subjects interacting with the system . For example, the metrics proposed by [10, 11, 12, 13] can only be used with specific types of XAI (i.e. prototype selection, feature attribution, etc.). While the metrics proposed by [14, 15] rely on usability tests and user-studies.

Interestingly, only [14] claim that their work is generic enough to be used to evaluate any XAI, proposing to measure explainability indirectly, by estimating the effects that the resulting explanations have on the subjects. More precisely, [14] 's metric is mainly inspired by the interpretation of explanations given by Cognitive Science, requiring to measure: [i)] the subjective goodness of explanations, whether users are satisfied by explanations, how well users understand the AI systems, how curiosity motivates the search for explanations, whether the user's trust and reliance on the AI are appropriate, how the human-XAI work system performs. In other terms, the metric presented by [14] is heavily relying on subjective measurements. Differently, our DoX is the very first example of a fully objective metric that can be used to evaluate the explainability of any textual information, to understand whether the amount of explainability is objectively poor even if the resulting explanations are perceived as satisfactory and good by the explainees.

III. Background

Being able to automatically generate explanations has attracted the interest of the scientific community for long. This interest has increased together with the importance of AI in our society and the growing need to explicate the complexity of modern software systems. Understanding what constitutes an explanation is a long-standing problem, with a complex history of debates and philosophical traditions, often rooted in Aristotle's works and those of other philosophers. In the present letter we will focus on Achinstein's theory of explanations [16].

A. Archetypal Questions

In 1983, Achinstein was one of the first scholars to analyse the process of generating explanations as a whole, introducing his philosophical model of a *pragmatic* explanatory process. According to Achinstein's theory, an explanation can be summarized as a correct content-giving answer to questions of various kinds, not necessarily linked to causality. More precisely, explanations are the result of an *illocutionary* act of pragmatically answering a question. In other terms, this means that there is a subtle and important difference between simply "answering questions" and "explaining", and that is *illocution*. Meaning that explaining is an act coming from an explicit intent of producing new understandings in an explainee by providing a correct content-giving answer to an open question.

Anyway, notwithstanding this definition, *illocution* seems to be too abstract to be implementable into a concrete software. Nonetheless, recent efforts towards the automated generation of explanations [3, 17], have shown that it may be possible to define *illocution* in a more "computer-friendly" way. As stated by [17], illocution in explaining involves informed and *pertinent* answers not just to the main question, but also to other (archetypal) questions of various kinds, even unrelated to causality, that are relevant to the explanations.

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

In other words, archetypal questions provide generic explanations on a specific aspect of the explanandum, in a given informative context, with a local or a global slant, 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?".

B. Carnap's Criteria of Adequacy

In philosophy, the most important work about the central criteria of adequacy of *explainable information* is likely to be Carnap's [18]. Even though Carnap studies the concept of *explication* rather than that of *explainable information*, we assert that they share a common ground making his criteria fitting in both cases. In fact, *explication* in Carnap's sense is the replacement of a somewhat unclear and inexact concept (the explicandum) by a new, clearer, and more exact concept called explicatum, and that is exactly what information does when made explainable.

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

Carnap's central criteria of explication adequacy are [18]: *similarity*, *exactness* and *fruitfulness*[4]. *Similarity* means that the explicatum should be similar to the explicandum, in the sense that at least many of its intended uses, brought out in the clarification step, are preserved in the explicatum. On the other hand, *Exactness* means that the explication should, where possible, be embedded in some sufficiently clear and exact linguistic framework. While *Fruitfulness* means that the explicatum should be used in a high number of other *good* explanations (the more, the better).

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 [19]. In fact an explanation is such regardless its truth (wrong but high-quality explanations exist, especially in science). Vice-versa, highly correct information can be very poorly explainable.

IV. Proposed Solution

In Section II 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 the present letter we try to answer this question, by leveraging on an extension of Achinstein's theory of explanations proposed by [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 thus also an *explanation*. In fact, considering that *explainability* is fundamentally the *ability to explain*, it is clear that a proper definition of it requires a precise understanding of what is *explaining*. So, in this Section we discuss both the new theory behind our proposed solution for computing the DoX and a concrete implementation we devised to measure the DoX in practice.

A. Explanatory Illocution and Degree of Explainability

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

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

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

Consequently we have that:

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

Importantly, the DoX, as we defined it, is akin to Carnap's *central* criteria of adequacy of explanation (introduced in Section III-B). Although, differently from Carnap, our understanding of *exactness* is not that of adherence to standards of formal concept formation[5], but rather that of being precise or pertinent enough as an answer to a given question.

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

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

The Averaged DoX represents a naive approach to quantify explainability with one single score, because it implies that all the archetypal questions and aspects have the same weight, although this may not be always true.

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

Given definition 4, we argue that it is possible to write an algorithm to quantify the DoX of information representable with a *natural language*, i.e. English. Let's suppose we want to measure the DoX of a set of texts called *explanandum support material*, containing correct textual information (in English) about a given explanandum. For example, if the *explanandum* were "heart diseases", there would be many aspects involved including "heart", "stroke", "vessel", "diseases", "angina", "symptoms", etc. Hence a reasonable *support material* for it would probably be a book describing all these aspects and more (if deemed relevant by the author), or a set of web-pages (i.e. those published by the *U.S. Centers for Disease Control and Prevention*[6]), or any other kind of corpus written in natural language.

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

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

[6] https://www.cdc.gov

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

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

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

Indeed, by leveraging on existing pre-trained deep language models, i.e. [20, 21], capable of converting snippets of text (e.g. questions and answers) into numerical representations, in the following sections we show how to concretely implement an algorithm capable of estimating the DoX score of any arbitrary piece of textual information with the pipeline shown in Figure 1.

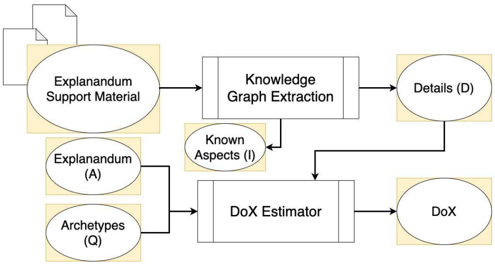

Fig. 1: **The DoX Pipeline**: The pipeline starts with the extraction of a knowledge graph from the *explanandum support material* that is then converted into a set of details D.

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

Considering the aforementioned characteristics of D and A, we believe that the most natural representation of them might be a knowledge graph. Indeed, knowledge graphs are sets of triplets connecting two different nodes (i.e. a sub-set of aspects in A) with some kind of relation or edge (i.e. a detail $d \in D$), hence any such (knowledge) graph representation of D and A would automatically give us a mechanism to identify a $D_a \subseteq D$ for every $a \in A$. Therefore, we believe that the easiest way to identify the set of details D (and possibly also A) might pass through some mechanism for the extraction of a (knowledge) graph of *information units* from the explanandum support material. Thus, an approach like the one used by [17, 22], for archetypal question answering, might be suitable to our ends, allowing for the identification of meaningful *information units* and (importantly) suggesting also a mechanism for the estimation of *pertinence*. In fact, the algorithm proposed by [22] relies on an automated mechanism that is capable of decomposing sentences into dependency trees, converting them into a special knowledge graph of Subject-Template-Object triplets (or template-triplets in short) specifically designed to facilitate (archetypal) question answering through state-of-the-art algorithms for Question Answer Retrieval [5, 22, 23]. Specifically, these algorithms can estimate the pertinence p of a detail $d \in D$ to a question q by generating a numerical embedding of both the question and the answer, so that the inner product (or other similarity metrics, i.e. the cosine similarity) between these embeddings is a measure of the pertinence of the latter to the first.

Therefore, template-triplets are used instead of normal Subject-Verb-Object triplets (or verb-triplets in short) to cope with the limitations of modern Question-Answer Retrieval. In fact, existing algorithms for Question Answer Retrieval [5, 23] have, usually, constraints on the size of their inputs and outputs; are trained on natural language snippets of text and not verb-triplets. More in detail, these template-triplets are a sort of function, where the template is the body of the function and the object and the subject are the parameters, so that obtaining a natural language representation of these template-triples is straightforward by design, by replacing the instances of the parameters in the body. Differently from the verb in the verb-triplets, the template can be any snippet of text, possibly containing multiple verbs or referring to external concepts that are not the subject or the object. Also considering that serialising natural language into verb-triples is a challenging open-problem, template-triplets have the potential to fully harness the expressive power of existing deep language models. An example of template-triple (in the form subject, predicate, object) is: “the applicable law”, “Surprisingly {subj} is considered to be clearly more related to {obj} rather than to something else.”, “that Member State”.

Importantly, as *information units*, to form the aforementioned template-triplets, [17] use meaningful decompositions of grammatical dependency trees, so to empower the units with the smallest granularity of information. As consequence, using such sub-trees as *information units* guarantees:

- a disentanglement of complex information bundles, into the most simple units, so to be able to correctly estimate the *level of detail* covered by the information pieces, as required by Definition 4.
- a better identification of duplicated units scattered

throughout the information pieces, so to avoid an overestimation of the *level of detail*.
- an easy way to understand whether an answer is invalid, as being totally contained in the question, hence forcing its *pertinence* to be zero.

All these properties meet the requirements that a good detail $d \in D$ should possess to be used for the generation of a DoX score, supporting our decision to re-use inside our pipeline the technology adopted by [17].

2) Archetypes Selection: According to Definition 1, an archetypal question is a very generic question characterised by one or more interrogative formulas. Literature is full of different examples of such archetypal questions, and many of them are used to classify both semantic and discourse relations [24, 25, 26]. Interestingly, it is possible to identify a sort of hierarchy or taxonomy of such archetypes, ordered by their intrinsic level of specificity. For example, the simplest interrogative formulas (made only of an interrogative particle, e.g. what) 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, 25, 26] to classify semantic relations within discourse. The main *archetypes* coming from Abstract Meaning Representation theory [25] are: What? Who? How? Where? When? Which? Whose? Why?
We refer to these archetypes as the *primary* ones because they consist only of interrogative particles. While the main *archetypes* coming from discourse theory [26] (also called *secondary archetypes* because they make use of the *primary archetypes*) are: In what manner? What is the reason? What is the result of it? What is an example of it? After what? While what? In what case? Despite what? What is contrasted with it? Before what? Since when? What is similar to it? Until when? Instead of what? What is an alternative to it? Except when? Unless what?

Despite the fact that many other archetypes may be identified (i.e. "Where to?" or "Who by?"), we believe that the list of questions we provided is rich enough to be generally representative for any other question, whereas more specific questions can be always framed by using the interrogative particles (i.e. why, what, etc.) we considered. In fact, *primary archetypes* can be used to represent any fact and abstract meaning [27], while the *secondary archetypes* can cover all the discourse relations between them. For concrete examples of how all these questions (especially the primary ones) are related to XAI algorithms, we point the reader to this recent survey by IBM Research [28].

V. Experiment

In Section IV we argued that the degree of explainability of any collection of text (i.e. the output of a XAI-based system) can be measured in terms of DoX on a set of chosen *Explanandum Aspects*. To this end, we devised a proof of concept using some XAI-based systems:

- a Heart Disease Predictor based on XGBoost [7] and TreeSHAP [8];
- a Credit Approval System based on a simple Artificial Neural Network and on CEM [9].

In a standard XAI-based system, the amount of *explainability* is by design, clearly and explicitly dependent on the output of the underlying XAI, for the black-box not being explainable by nature. So that, by masking the output of the XAI, the overall system can be forced to be not explainable enough. This characteristic can be exploited to show that an increment of explainability implies an increment of DoX.

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

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

A. XAI-Based Systems

The XAI-Based Systems we considered are a Credit Approval System and a Heart Disease Predictor, respectively on finance and healthcare topics. Both these two systems are an example of XAI 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.

1) Finance: Credit Approval System: The Credit Approval System is the same used also in [3, 17], designed by IBM to showcase AIX360 [29]. This explanandum is about finance and the system is used by a bank. The bank deploys an Artificial Neural Network to decide whether to approve a loan request, and it uses the CEM [9] algorithm to create post-hoc contrastive explanatory information. This information is meant to help the customers, showing them what minimal set of factors is to be manipulated for changing the outcome of the system from denial to approval (or vice-versa).

The Artificial Neural Network was trained on the "FICO HELOC" dataset [30]. The FICO HELOC dataset contains anonymized information about Home Equity Line Of Credit (HELOC) applications made by real homeowners. Importantly, the Artificial Neural Network is trained to properly answer the following question: "What is the decision on the loan request of applicant X?".

Given the specific characteristics of this system, it is possible to assume that the main goal of its users is about understanding what are the causes behind a loan rejection and what to do to get the loan accepted. This is why the output of CEM is designed to answer the questions:

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

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

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

- Context: a titled heading section kindly introducing Mary (the user) to the system.
- AI Output: the decision of the Artificial Neural Network for the loan application. This decision normally can be “denied” or “accepted”. For Mary it is: “denied”.
- XAI Output: a section showing the output of CEM. This output consists in a minimal ordered list of factors that are the most important to change for the outcome of the AI to switch.

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

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

The dataset used to train XGBoost is the “UCI Heart Disease Data” [31, 32]. TreeSHAP [8], a famous XAI algorithm specialised on tree ensemble models (i.e. XGBoost) for post-hoc explanations, is used to understand what is the contribution of each feature to the output of the model (that is XGBoost). TreeSHAP can be used to answer the following questions:

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

The first level responder is responsible for handling the patient’s requests for assistance, forwarding them to the right physician in the eventuality of a reasonable risk of heart disease. First level responders get basic questions from callers, they are not doctors but they have to decide on the fly whether the caller should speak to a real doctor or not. So, they quickly use the XAI system to figure out what to answer to the callers and what are the next actions to suggest. In other words, this system is used directly by the responder, and indirectly by the caller through the responder. These two types of users have different but overlapping goals and objectives. It is reasonable to assume that the goal of the responders is to answer in the most efficient and effective way the questions of a caller. To this end, the questions answered by TreeSHAP are quite useful, but many other important questions should probably be answered, including: “What is the easiest thing that the patient could actually do to change his heart disease risk from medium to low?”, “How could the patient avoid raising one of the factors, preventing his heart disease risk to raise?”, etc.

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

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

B. Pertinence Functions and Thresholds

According to Definition 4, we need to define a pertinence function p and pick a threshold t in order to compute the DoX. As discussed in Section IV-B, we are going to use as pertinence function p a deep language model for Question-Answer Retrieval. The point is that many different deep language models exist for this task, i.e. [5, 22, 23], 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. In fact, assuming that these models get good results on state-of-the-art benchmarks for *pertinence estimation*, i.e. [5, 21], we believe that the results of the computation of DoX should be consistent across them. Hence the models we considered are:

- FB: published by [5] and [20], and trained on the combination of the following datasets: Natural Questions [33], TriviaQA [34], WebQuestions [35], and CuratedTREC [36].
- TF: or Multilingual Universal Sentence Encoder [21] and trained on the Stanford Natural Language Inference corpus [6].

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

C. Direct Evaluation on Normal XAI-generated Explanations

With this experiment we compare the DoX of the output of a XAI Explainer with the DoX of the same explainer but with no XAI, also called non-XAI Explainer. We expect the (averaged) DoX of the XAI Explainer to be clearly higher than its non-XAI Explainer.

For this experiment, we used the XAI-based systems defined in Section V-A. In fact, both the Credit Approval System and the Heart Disease Predictor are examples of XAI Explainer. Therefore, by simply removing the output of the XAI (respectively CEM and TreeSHAP) from these systems we obtain a non-XAI Explainer. In order to compare the (averaged) DoX of a XAI Explainer to that of its non-XAI Explainer, as set of *Explanandum Aspects* we take those targeted by the

XAI Explainer and the non-XAI Explainer. More precisely, the main *Explanandum Aspects* targeted by XGBoost [7] and TreeSHAP [8] 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* targeted by the Artificial Neural Network and CEM [9] 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

For computing the DoX, we used the pipeline described in Section IV to extract different knowledge graphs of details D. After properly converting the images produced by the XAI Explainer to textual explanations, the resulting *Explanandum Aspects Coverage* of the XAI Explainer for both the Heart Disease Predictor and the Credit Approval System is 100%, while that of the non-XAI Explainer is 60% for the Heart Disease Predictor and 50% for the Credit Approval System.

After extracting the knowledge graphs, we used the set of *Explanandum Aspects* A to select from them all the details D_a. We did it for each $a \in A$, checking every a against the nodes of the graph, exploiting the properties of the template-triplets to identify every detail $d \in D_a$. More precisely, we were able to understand whether a template-triplet is likely to be related to an $a \in A$ by using the algorithm[7] described in [37] and used also by [22]. Finally, we were able to compute the DoX scores in accordance with Definition 4 by using:

- the set of archetypes Q described in Section IV-B2,
- the pertinence functions p and the thresholds t presented in Section V-B,
- the aforementioned set of details D_a of each $a \in A$.

Computing the DoX we got the results displayed in Table I. As expected, on both the Heart Disease Predictor and the Credit Approval System, the results of the experiment neatly show that the averaged DoX of the XAI Explainer is way higher than the non-XAI Explainer, regardless the adopted *deep language model*.

VI. Discussion

Applying DoX to any XAI technique is straightforward as soon as it is possible to encode the output of such XAI into a textual (English) representation. The results of our experiment tell us that whenever new information about different aspects to be explained is added to the explanandum support material (see Section IV for a definition of what that is), the DoX scores increase. Importantly, no inconsistencies were found across the considered pertinence functions (TF and FB; see Section V-B). This suggests that the alignment of DoX with explainability may be independent from the chosen deep language model, at least in the considered environments.

TABLE I: **Experiment - Degree of Explainability**: in this table DoX and Averaged DoX are shown for the Credit Approval System (CA) and the Heart Disease Predictor (HD). As columns we have the non-XAI Explainer (NAE) and the XAI Explainer (NXE). As rows we have different explainability estimates using FB and TF. For simplicity, with DoX we show only the *primary archetypes*.

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

We believe that this is happening because both TF and FB, in average, perform reasonably well on the same benchmarks for evaluating Question-Answer Retrieval. In other terms, it could be that if the Averaged DoX aggregates enough archetypes and the number of considered aspects and details is also enough, then different pertinence functions performing in similar ways on some good benchmarks may produce similar Averaged DoX scores despite their differences (i.e. the archetype with the best explanatory illocution in TF is "what", while in FB is "which"). Anyway, this does not exclude that there might be a deep language model that is better than others for computing the DoX, or that multiple standardised deep language models should be adopted for a thorough estimate of the DoX. Furthermore, it is also important to see whether there is an alignment between the explainability perceived by subjects and DoX. We leave these analyses for future work.

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

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

In other terms, DoX could be used to understand whether a piece of information is enough to explain something. Indeed, our DoX is a fully objective metric that can evaluate the explainability of any textual information and understand

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

whether the amount of explainability is objectively poor, even if the resulting explanations are perceived as satisfactory and good by the explainees. We deem that this characteristic of DoX is very important, in fact if explanations are built over explainable information, a poor degree of explainability objectively implies poor explanations, no matter how good the adopted explanatory process is (perceived): "Users also do not necessarily perform better with systems that they prefer and trust more. To draw correct conclusions from empirical studies, explainable AI researchers should be wary of evaluation pitfalls, such as proxy tasks and subjective measures" [38].

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

On the other hand, identifying a proper A is not enough, for estimating the DoX also a set of archetypes Q is needed. Considering the impressive and possibly infinite amount of archetypal questions that our language can conceive, it would appear that also the choice of Q might be a source of subjectivity. But questions and archetypes have been studied for a very long time in linguistics, resulting in many theories capable of organising our understanding of what constitutes a discourse and a representation of knowledge. This is why we assert that instead of relying on subjective choices of Q, we can exploit the plethora of (what we call) archetypal questions, identified by linguistic theories as those discussed in Section IV-B2, ensuring an objective DoX.

VII. Conclusions and Future Work

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

In order to understand whether the DoX is actually behaving as *explainability* is expected to, we designed an experiment on two realistic AI-based systems for heart disease prediction and credit approval, involving famous AI technology as Artificial Neural Networks, TreeSHAP [8], XGBoost [7] and CEM [9]. The results we obtained show that the DoX is aligned to our expectations, and it is possible to actually quantify *explainability* in natural language information.

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

References

[1] A. HLEG, "Ethics guidelines for trustworthy ai," 2019.

[2] M. E. Kaminski and G. Malgieri, "Algorithmic impact assessments under the gdpr: producing multi-layered explanations," *U of Colorado Law Legal Studies Research Paper*, no. 19-28, 2019.

[3] F. Sovrano and F. Vitali, "Generating user-centred explanations via illocutionary question answering: From philosophy to interfaces," *ACM Trans. Interact. Intell. Syst.*, feb 2022. [Online]. Available: https://doi.org/10.1145/3519265

[4] C. D. Novaes and E. Reck, "Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization," *Synthese*, vol. 194, no. 1, pp. 195–215, 2017.

[5] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," 2020.

[6] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," 2015.

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

[8] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

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

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

[11] A.-p. Nguyen and M. R. Martínez, "On quantitative aspects of model interpretability," *arXiv preprint arXiv:2007.07584*, 2020.

[12] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Interpretable & explorable approximations of black box models," *arXiv preprint arXiv:1707.01154*, 2017.

[13] M. T. Keane, E. M. Kenny, E. Delaney, and B. Smyth, "If only we had better counterfactual explanations: Five

key deficits to rectify in the evaluation of counterfactual xai techniques," *arXiv preprint arXiv:2103.01035*, 2021.

[14] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," 2018. [Online]. Available: https://arxiv.org/abs/1812.04608

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

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

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

[18] H. Leitgeb and A. Carus, "Rudolf carnap," 2021. [Online]. Available: https://plato.stanford.edu/archives/sum2021/entries/carnap/

[19] D. J. Hilton, "Mental models and causal explanation: Judgements of probable cause and explanatory relevance," *Thinking & Reasoning*, vol. 2, no. 4, pp. 273–308, 1996.

[20] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019.

[21] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung *et al.*, "Multilingual universal sentence encoder for semantic retrieval," 2019.

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

[23] M. Guo, Y. Yang, D. Cer, Q. Shen, and N. Constant, "Multireqa: A cross-domain evaluation for retrieval question answering models," *arXiv preprint arXiv:2005.02507*, 2020.

[24] L. He, M. Lewis, and L. Zettlemoyer, "Question-answer driven semantic role labeling: Using natural language to annotate natural language," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 643–653.

[25] J. Michael, G. Stanovsky, L. He, I. Dagan, and L. Zettlemoyer, "Crowdsourcing question-answer meaning representations," 2017.

[26] V. Pyatkin, A. Klein, R. Tsarfaty, and I. Dagan, "Qadiscourse–discourse relations as qa pairs: Representation, crowdsourcing and baselines," 2020.

[27] J. Bos, "Expressive power of abstract meaning representations," *Computational Linguistics*, vol. 42, no. 3, pp. 527–535, 2016.

[28] Q. V. Liao, D. Gruen, and S. Miller, *Questioning the AI: Informing Design Practices for Explainable AI User Experiences*. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–15. [Online]. Available: https://doi.org/10.1145/3313831.3376590

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

[30] S. Holter, O. Gomez, and E. Bertini, "Fico explainable machine learning challenge," 2019. [Online]. Available: https://fico.force.com/FICOCommunity/s/explainable-machine-learning-challenge?tabset-3158a=a4c37

[31] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American journal of cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[32] R. Alizadehsani, M. Roshanzamir, M. Abdar, A. Beykikhoshk, A. Khosravi, M. Panahiazar, A. Koohestani, F. Khozeimeh, S. Nahavandi, and N. Sarrafzadegan, "A database for using machine learning and data mining techniques for coronary artery disease diagnosis," *Scientific data*, vol. 6, no. 1, pp. 1–13, 2019.

[33] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee *et al.*, "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 2019.

[34] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," 2017.

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

[36] P. Baudiš and J. Šedivỳ, "Modeling of the question answering task in the yodaqa system," in *International Conference of the cross-language evaluation Forum for European languages*. Springer, 2015, pp. 222–228.

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

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