

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

From Philosophy to Interfaces: an Explanatory Method and a Tool Inspired by Achinstein's Theory of Explanation

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Sovrano, F., Vitali, F. (2021). From Philosophy to Interfaces: an Explanatory Method and a Tool Inspired by Achinstein's Theory of Explanation. New York : Association for Computing Machinery [10.1145/3397481.3450655].

Availability:

This version is available at: <https://hdl.handle.net/11585/819532> since: 2021-04-29

Published:

DOI: <http://doi.org/10.1145/3397481.3450655>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Francesco Sovrano and Fabio Vitali. 2021. From Philosophy to Interfaces: an Explanatory Method and a Tool Inspired by Achinstein's Theory of Explanation. In 26th International Conference on Intelligent User Interfaces (IUI '21). Association for Computing Machinery, New York, NY, USA, 81–91.

The final published version is available online at: <https://doi.org/10.1145/3397481.3450655>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

From Philosophy to Interfaces: an Explanatory Method and a Tool Inspired by Achinstein's Theory of Explanation

FRANCESCO SOVRANO and FABIO VITALI

We propose a new method for explanations in Artificial Intelligence (AI) and a tool to test its expressive power within a user interface. In order to bridge the gap between philosophy and human-computer interfaces, we show a new approach for the generation of interactive explanations based on a sophisticated pipeline of AI algorithms for structuring natural language documents into knowledge graphs, answering questions effectively and satisfactorily. Among the mainstream philosophical theories of explanation we identified one that in our view is more easily applicable as a practical model for user-centric tools: Achinstein's *Theory of Explanation*. With this work we aim to prove that the theory proposed by Achinstein can be actually adapted for being implemented into a concrete software application, as an interactive process answering questions. To this end we found a way to handle the generic (*archetypal*) questions that implicitly characterise an explanatory processes as preliminary overviews rather than as answers to explicit questions, as commonly understood. To show the expressive power of this approach we designed and implemented a pipeline of AI algorithms for the generation of interactive explanations under the form of overviews, focusing on this aspect of explanations rather than on existing interfaces and presentation logic layers for question answering. Accordingly, through the identification of a minimal set of archetypal questions it is possible to create a generator of explanatory overviews that is generic enough to significantly ease the acquisition of knowledge by humans, regardless of the specificities of the users outside of a minimum set of very broad requirements (e.g. people able to read and understand English and capable of performing basic common-sense reasoning). We tested our hypothesis on a well-known XAI-powered credit approval system by IBM, comparing CEM, a static explanatory tool for *post-hoc* explanations, with an extension we developed adding interactive explanations based on our model. The results of the user study, involving more than 100 participants, showed that our proposed solution produced a *statistically relevant* improvement on effectiveness ($U=931.0$, $p=0.036$) over the baseline, thus giving evidence in favour of our theory.

CCS Concepts: • **Human-centered computing** → **HCI theory, concepts and models**; **Empirical studies in HCI**.

Additional Key Words and Phrases: Methods for explanations, Education and learning-related technologies, ExplanatorY Artificial Intelligence (YAI)

ACM Reference Format:

Francesco Sovrano and Fabio Vitali. 2021. From Philosophy to Interfaces: an Explanatory Method and a Tool Inspired by Achinstein's Theory of Explanation. In *26th International Conference on Intelligent User Interfaces (IUI '21)*, April 14–17, 2021, College Station, TX, USA. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3397481.3450655>

1 INTRODUCTION

The complexity of modern software and the increasing discomfort of humans towards the correctness and fairness of the output of such complex systems has caused the birth and growth of a new discipline to reduce the distance between individuals, society, and machines: eXplainable AI (XAI). Governments have also started to act towards the establishment of ground rules of behaviour from complex systems, for instance through the enactment of the European

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.
Manuscript submitted to ACM

General Data Protection Regulation (GDPR) (2016¹), which identifies *fairness*, *lawfulness*, and in particular *transparency* as basic principles for every data processing tools handling personal data; even creating a new *Right to Explanation* for individuals whose legal status is affected by a solely-automated decision. The High-Level Expert Group on Artificial Intelligence (AI-HLEG) was established in 2018 with the explicit purpose of applying the principles of the GDPR specifically to AI software, and produced a list of fundamental ethical principles for *Trustworthy AI* tools that include *fairness* and *explicability*. XAI, therefore, is understood as a way to answer these fundamental principles in the field of AI. Among the characteristics that the AI-HLEG has identified, there is the concept of *user-centrality*. In order to address such concept, explanations have to cater to the individuals in terms of their background, context, needs and purposes: a single explanation output (i.e., a *One-Size-Fits-All* explanation) regardless of all these backgrounds and needs would become massive and unwieldy as soon as the data and processes to be explained become averagely complex, and the readers approaching such explanations would be immediately overwhelmed by the sheer size of the data. Consequently, One-Size-Fits-All approaches to explanations cannot be considered appropriate to this end as soon as the complexity of these systems surpasses a fairly trivial threshold, and a more sophisticated approach needs to be identified. Despite these considerations, it appears that XAI techniques are mainly focused on pursuing One-Size-Fits-All explanations, justified by convenient definitions framing an explanation as the product of an act of making things explainable rather than a pragmatic (user-centric) act of explaining based on explainability.

In this paper, we propose a new approach to explanations in Artificial Intelligence. We are interested in modelling an explanatory process for producing user-centric explanatory software and quantifying the difference it bears in terms of effectiveness with respect to non-pragmatic approaches. More precisely, we want to understand how to structure information in order to facilitate the production of explanations of complex decision-making processes (such as explainable artificial intelligence processes). We acknowledge that we are not the first to try to model an explanatory process. In literature there were various efforts in this direction and a long history of debates and philosophical traditions, often rooted in Aristotle's works and those of other philosophers. Among the many models proposed over the last few centuries some are now considered fallacious, albeit historically useful (e.g. Hempel et al.'s one [13]), but we have identified the one that, we believe, is mostly convertible into a practical model for user-centric explanatory software: Achinstein's *illocutionary* theory of explanations, as answers to questions. Despite its criticalities, Achinstein's theory seemed to us the most suitable among all, for our purposes, for allowing the assessment of the quality of an explanation/answer on the base of its *pragmatic* relevance to a question. A task that may seem too onerous and subjective, nonetheless recent developments in modern artificial intelligence have shown there might exist tools [24, 31] to objectively estimate the pertinence of an answer thus allowing for the automation of a question answering process.

Though, modelling an explanatory process as a standard Question Answering (QA) process gave us the first impression of being a little bit unrealistic. Think of the following example of the "university lectures": students (the explainees) follow the lessons to acquire (initially obscure) information provided by the professor (the explainer). A lesson can normally include the intervention of students in the form of observations and/or questions, but these interventions are, in practice, always after an initial phase of information acquisition. In other terms, the initial overview given by the professor may not be the answer to a preliminary question (especially if the students know absolutely nothing about what the professor is supposed to say). Regardless this apparent lack of a question, we might all agree that the professor could actually explain something good to the students. At this point it would seem that Achinstein's theory, being heavily based on question-answering, fails to capture the need for preliminary overviews during an explanatory process, as in

¹Regulation (EU) 2016/679.

the “university lectures” example. Despite this first impression, we think that overviews can be generated as answers as well, therefore partially confirming Achinstein’s original theory. In fact, for the generation of an overview it is necessary to select and group information appropriately, so as to facilitate the production of different explanatory paths for different users. As analogy, we might see the space of all explanations (or explanatory space [28]) as a sort of manifold where every single overview is not user-centred locally, but globally (as an element of a sequence of information, chosen by a user). What we theorize is that it is possible to generate such overviews through the identification of a set of archetypal questions (e.g. Why? What for? How? When? etc..) that allow contextualised information to be grouped and filtered according to its relative pertinence with respect to the archetypes. If our hypothesis is correct, through the identification of a minimal set of archetypal questions, it is possible to obtain a generator of explanatory overviews generic enough to be able to significantly ease the acquisition of knowledge, regardless of the specific user but depending instead on a category of selected users (e.g. people able to read and understand English, hence capable of performing common-sense reasoning), thus resulting in a user-centred explanatory tool that is more effective than its non-pragmatic counterpart based on the same informative contents.

In order to test our hypothesis, we had to: i) invent a new pipeline of AI algorithms for the generation of overviews, ii) design a simple user interface for presenting the overviews, iii) and run a user study on it. The pipeline is meant to organize the information contained in non-structured documents written in natural language (web pages, pdf, etc..), allowing efficient information clustering, according to a set of archetypal questions, aiming to build a sufficiently rich and effectively explorable Explanatory Space (ES) for the automated generation of user-centred explanations.

This paper is structured as follows. In Section 2 we provide a brief introduction to the contemporary philosophical developments in the theory of explanation, focusing on Achinstein’s one. In Section 3 we describe our proposed solution going through the details of the pipeline, thus comparing our work with existing literature in Section 4. In Section 5 we present a proof of concept evaluation in the form of an user study on a XAI-powered credit approval system augmented with a simple extension for the navigation of the ES. In Section 6 we show and discuss the obtained results, drawing the conclusions in Section 7.

2 BACKGROUND: CONTEMPORARY DEVELOPMENTS IN THE THEORY OF EXPLANATION

In philosophy the terms “truth” and “explanation” have divergent interpretations [20]. On one hand, according to the realist interpretation “the truth and explanatory power of a theory are matters of the correspondence of language with an external reality”, so it is more about depicting reality with words and intents. On the other hand, the epistemic interpretation defines explanation as a mere re-ordering of phenomena and experience to a greater degree, focusing on its power to explain observable phenomena rather than its literal truth. The failure to distinguish these senses of “explanation” can and does foster disagreements that are purely semantic in nature.

According to Mayes [20], explanation in philosophy has been conceived within the following five traditions:

- **Causal Realism** [25]: explanation as the articulation of the fundamental causal mechanisms of a phenomenon.
- **Constructive Empiricism** [29]: epistemic theory of explanation that draws on the logic of why-questions and on a Bayesian interpretation of probability.
- **Ordinary Language Philosophy** [1]: the act of explanation as the attempt to produce understanding in another by answering a certain kind of question in a certain kind of way.
- **Cognitive Science** [14]: explaining as a process of belief revision, etc..

- **Naturalism and Scientific Realism** [26]: rejects any kind of explanation of natural phenomena that makes essential reference to unnatural phenomena. Explanation is not something that occurs on the basis of pre-confirmed truths. Rather, successful explanation is actually part of the process of confirmation itself.

2.1 Achinstein's Theory of Explanations

In 1983, Achinstein [1] was one of the first to analyse the whole process of generating explanations, introducing his philosophical model of a *pragmatic* explanatory process. According to his model, explaining is an illocutionary act coming from a clear intention of producing new understandings in an explainee by providing a correct content-giving answer to an open question. Therefore, according to this view, answering by “filling the blank” of a pre-defined template answer (as most of One-Size-Fits-All approaches do) prevents the act of answering from being explanatory, by lacking illocution. These conclusions are quite clear and explicit in Achinstein's last works [2], consolidated after a few decades of public debates. More precisely, according to Achinstein's theory, an explanation can be summarized as a pragmatically correct content-giving answer to one or more questions of various kinds, not necessarily linked to causality. As consequence we can see a deliberate absence of a taxonomy of questions to refer to (helpful to categorize and better understand the nature of human explanations). This results in a refusal to define a quantitative way to measure how pertinent an answer is to a question, justified by the important assertion that explanations have a pragmatic character, so that what exactly has to be done to make something understandable to someone may (in the most generic case) depend on the interests and background knowledge of the person seeking understanding [8]. In this sense, the strong connection of Achinstein's theory to natural language and natural users is quite evident, for example in the Achinsteinian concept of *elliptical understandings* as “understandings of what significance or importance X has (in the present context)” [2] or in the concept of *u-restrictions* where an utterance/explanation can be said to express a proposition if and only if it can appear in (many) contexts reasonably known by the explainee. But, despite this, Achinstein does not reject at all the utility of formalisms, hence suggesting the importance of following *instructions* (protocols, rules, algorithms) for correctly explaining some specific things within specific contexts.

3 PROPOSED SOLUTION

1983's illocutionary theory of explanation is pragmatic, user-centric, but it seems very hard to be concretised into a software. The point is: what is *illocution*, precisely? Although also the dictionary gives its own definition, this seems to be too abstract to be implementable into a real user-centric explanatory software, also considering that in Achinstein's theory there is a deliberate absence of a taxonomy of questions, that might be needed for a more objective evaluation of the quality of an explanation. Despite its well-known criticalities, Achinstein's model seemed to us the most suitable among all for building a user-centric explanatory software, by suggesting that explaining is akin to Question Answering (QA). On this aspect, one of the main technological limitations of state-of-the-art automated QA [24, 31] is that it tends to lose its effectiveness when the questions are too broad. This results in several issues when the user has no knowledge of the domain, thus forcing him/her to resort to generic questions in order to acquire enough information to be able to approach more specific questions. Seemingly, this technological issue may have its own counterpart in Achinstein's theory as well, as shown in the “university lectures” example presented in Section 1. In fact, it would seem that Achinstein's theory, being so heavily based on question-answering, fails to capture the need for preliminary overviews during an explanatory process. Despite this first impression, we think that overviews can be generated as answers as well, therefore partially confirming Achinstein's original theory. Indeed, we believe that illocution in explanations is equivalent to the act of *pertinently answering archetypal questions* (e.g. Why? What

for? How? When? etc..), and that is different from the Achinstein’s concept of *instructions* but in some way akin to the Achinsteinian concept of *elliptical understandings* briefly introduced in Section 2. In fact, explaining is not just answering a given question in a punctual way (that would simply be answering) but it is also answering all the other implicit (archetypal) questions defined by: the explainee’s background knowledge, the objectives of the explanatory process and the given context. It is, in some sense, attempting to anticipate the (conceivably mostly unknown) explainee’s needs for an explanation by providing, as an archetypal answer, possibly expandable summaries of (more punctual) pertinent information. In other terms, the more archetypal answers about the explanandum’s aspects are covered by the act of explaining, the more likely the resulting explanation is going to meet the explainee’s objectives, the better is the explanatory process. The archetypal questions prevent by design any “filling the blank” answer, thus meeting the tricky but reasonable assumption of illocution given by Achinstein for his pragmatic theory of explanations. What is of most interest is that, perhaps, the assumption of explaining as *pertinently answering* also to *archetypal questions* is simpler, stronger and more precise than the assumption of illocution, removing from the “equation” (therefore simplifying it) the need for a model of human intention. Nonetheless, one main technical issue with this theory of explanations is with respect to user-centrality, requiring an explainer (e.g. an explanatory system) to actually anticipate the explainee’s (non-explicit) needs; how could one estimate pertinence without exactly knowing with respect to what? In a context in which only very minimal assumptions can be made on the explainee, we propose one more strong difference with Achinstein’s model. In this context, we define *pertinently answering* (to archetypal questions) as the process of giving (archetypal) answers that are likely to be pertinent for a given (archetypal) question. The likelihood can be quantitatively estimated on strong-enough statistical evidence collected from large corpora and built in language models. This statistical definition of pertinence is also compatible with the definition of *u-restrictions* given by Achinstein and it does not preclude a pragmatic (user-centred) explanatory process, as we are going to show, that is locally non-pragmatic but globally pragmatic. In fact, for the generation of an overview it is necessary to select and group information appropriately, so as to facilitate the production of different explanatory paths for different users. As analogy, we might see the space of explanations (or Explanatory Space [28]) as a sort of manifold space where every single overview is not user-centred locally, but globally (as an element of a sequence of information, choosable by a user).

The proposed solution builds over the extraction and structuration of an Explanatory Space (ES), intended (as in [28]) as the set of all possible explanations (about an explanandum) reachable by a user, through an explanatory process, starting from an initial explanans, via a pre-defined set of actions. According to the model of Sovrano et al., we might see the ES as a graph of interconnected bits of explanation, and an explanation as nothing more than a path within the ES. The aforementioned interconnected bits of explanation can be seen as overviews about different aspects of the explanandum. We can see an explanandum aspect as an information cluster. Assuming that the explanandum is a set of documents written in a natural language (e.g. English), the core aspects to explain might be (for example) the different concepts/entities within the corpus, so that to each concept it is possible to associate an overview; e.g. in the sentence “the customer opened a new bank account” different entities are “customer”, “bank”, “bank account”. The choice of an initial explanans is generally dependent on the nature of the explanandum and the objectives associated with the category of users involved in the explanatory process. A good choice of initial explanans could be an overview of the whole explanandum or of the explanatory process. Therefore, in the case of XAI, a proper initial explanans might be the static explanation provided by the XAI algorithm (e.g. by compiling a template or generating text through a formal language). An explanatory process can be defined by the choice of appropriate heuristics for exploring the ES and structuring information clusters. Considering that an ES is a particular type of graph, and considering that many

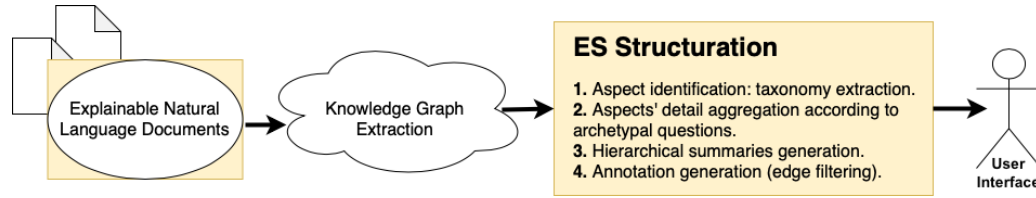


Fig. 1. **The Pipeline:** A simple diagram summarising the pipeline of our user-centric explanatory software.

instances of NP-difficult problems on graphs can be efficiently solved via tree decomposition, the heuristics in question shall provide a policy for at least: i) organising the ES's nodes or aspects, ii) structuring the information internal to the ES's nodes, iii) ordering/filtering the ES's edges in a way that would effectively decompose the graph into a tree.

The heuristics we adopted are respectively: abstraction (for picking the explanandum's aspects, or ES's nodes), relevance (for organising the information internal to the ES's nodes) and simplicity (for filtering the information internal to ES's nodes and also for selecting the viable ES's edges). In order to implement the three aforementioned heuristics, for structuring and exploring of the ES we need an algorithm that (for example) would, as shown in figure 1:

- (1) Identify and extract out of the explanandum all the different aspects (concepts/entities) and their related information.
- (2) Build a knowledge graph so that concepts/entities are linked together.
- (3) Extract a taxonomy from the knowledge graph.
- (4) Build one or more information clusters for every aspect, according to the identified archetypal questions.
- (5) Present information clusters through a hierarchy of expandable summaries.
- (6) Filter the external edges of the ES, favouring shorter and simpler paths/explanations, thus reasonably reducing the amount of redundant information for a human.

3.1 Knowledge Graph Extraction

Knowledge Graph (KG) extraction is the extraction of concepts and their relations, from natural language text, in the form of a graph where concepts are nodes and relations are edges. We are looking for a way to extract Knowledge Graphs (KGs) that somehow preserve the original natural language, preferring them over classical Resource Description Framework (RDF) graphs. This way we can easily make them inter-operate with deep-learning based QA algorithms and existing language models. More in detail, as in [27], we perform KG extraction by:

- (1) Analysing the grammatical dependencies of the tokens extracted by Spacy's Dependency Parser, thus identifying the (target) concepts and entities in the form of syntagms.
- (2) Using the dependency tree to extract all the tokens connecting two different target concepts in a sentence, thus building a textual template formed by the ordered sequence of the identified tokens and the target concepts replaced with the placeholders "{subj}" and "{obj}" (in accordance with their grammatical dependencies).
- (3) Creating a graph of subject-predicate-object triples where the target concepts are the subject and the object and the textual template is the predicate.

The resulting triples are a sort of function, where the predicate is the body of the function and the object/subject are its parameters. Obtaining a natural language representation of these template-triples is straightforward by design, by replacing the instances of the parameters in the body. An example of such a 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”. Therefore, to increase the interoperability of the extracted KG with external resources we performed the following extra steps: i) We automatically assigned a URI and a RDFS label to every node of the graph. The URI is obtained by lemmatising the label. ii) We automatically added special triples to keep track of the snippets of text (the sources) from which the concepts and the relations are extracted. iii) We automatically added sub-class relations between composite concepts (syntagms) and the simpler concepts composing the syntagm.

Because of the adopted extraction procedure, the resulting KG is not perfect, containing mistakes caused by wrong dependency assignments or similar issues. Despite this, due to the fact that the original natural language is practically preserved thanks to the textual templates, this will not impact significantly on QA.

3.2 Taxonomy Construction: Nodes Clustering

In order to efficiently use, query and explore the extracted KG, we need to structure it in a proper way. We believe that effective abstract querying can be possible by structuring the KG as a light ontology, giving it a solid backbone in the form of a taxonomy. In fact, being able to identify the types/classes of a concept would allow to perform queries with a reasonable level of abstraction, making possible to refer to all the sub-types (or to some super-types) of a concept without explicitly mentioning them.

The taxonomy construction phase consists in building one or more taxonomies, via Formal Concept Analysis (FCA) [12]. In order to build a taxonomy via FCA one approach consists in exploiting, as FCA’s properties, the hypernym relations of the concepts in the KG. We found that the simplest way to extract such relations is through the alignment of the KG to WordNet², through a Word-Sense Disambiguation algorithm. The application of FCA on the aligned WordNet concepts (and their respective hypernyms) produces as result a forest of taxonomies. Every taxonomy in the forest is a cluster of concepts rooted into very abstract WordNet concepts that we can use as label for the respective taxonomies.

3.3 Overview Generation via Question Answering: Information Clustering and Summarisation

As mentioned in the previous sections, we can generate an overview by clustering and ordering information with respect to its pertinence to a set of archetypal questions. The archetypal questions we considered are: why (standing for causal or justificatory explanations), what for (teleological), how (expository), when (temporal), where (spatial), what and who (descriptive). The essential idea is to generate a concept overview by performing KG-based question answering, retrieving the most similar concept’s triples for each archetype. KG-based question answering consists in answering natural language questions about information contained in the KG. More in detail, let Q be an archetypal question and C a concept, we perform information clustering by:

- (1) **Extracting** all the template-triples related to C , including those of C ’s sub-classes.
- (2) **Selecting**, among the natural language representations of both the retrieved triples and their respective subjects/objects, the snippets of natural language that are sufficiently likely to be an answer to Q .
- (3) **Returning** as set of answers the contexts (the source paragraphs) of the selected triples, ordered by pertinence.

More in detail, the *selecting* phase is performed by computing the pertinence of an answer as the inner product between the embeddings of the contextualised snippets of text and the embedding of Q . The aforementioned embedding is obtained by means of a specialised language model such as the Universal Sentence Encoder (USE) for QA [31], while

²We are aware that WordNet is not omni-comprehensive, but at this stage of the work we are only interested in extracting a reasonable taxonomy.

the context is the source paragraph from which a snippet of text is extracted from the original document. If a snippet of text has a similarity above a given threshold, then it is said to be sufficiently likely an answer to Q , therefore pertinent.

Considering that an answer could be reasonably associated to more than one archetypal question, we decided to apply an heuristic filtering strategy in the attempt to minimise redundant information, thus following the simplicity heuristic. To do so, we had to attempt a sort of hierarchical organisation of the archetypes, defining some questions as more generic than others, thus prioritising the less generic ones. In fact, in some cases an answer to the question “What?” could also be a valid answer to “What for?”. This is because “What for?” is intuitively more specific than “What?”. Hence, to reduce redundancy, we can force answers to be exclusive to a single archetypal question, assigning first the answers to the most specific archetypes. A descending ordering of specificity, that we found meaningful for the identified archetypal questions, is: why, what for, how, who, where, when, and what. Such ordering seemed to be proper for the purposes of the proof of concept presented in Section 5, but it is likely that a different ordering is required for different purposes.

Finally, after the identification of a set of answers for a question Q , we can build an expandable summary by recursively concatenating together few answers and by summarising them (thus recursively building a tree of summaries) through one of the state-of-the-art deep learning algorithms for extractive or abstractive summarisation provided by Wolf et al.[30]. At the end of the process we have that an overview is defined by a (sometimes empty) expandable summary for every archetypal question, plus the list of super-classes, sub-classes, sub-types (if any) and eventually few other external resources considered to be of any use (e.g. a short abstract of few words). Therefore, we have that the additional taxonomical information is used to guarantee the abstraction policy, while the rest of the information is meant to guarantee both the relevance and the simplicity policies.

3.4 Overview Annotation: Edge Filtering

Every sentence in the overview is annotated. Annotations consist in linking a concept’s embodiment to its corresponding overview (so that clicking on the link would open the overview). The edge filtering algorithm has to decide which syntagms to annotate, in order to avoid annotating every possible concept expressed in a sentence, including redundant or useless ones. More precisely, the edge filtering algorithm would remove:

- Those concepts that can be assumed of scarce relevance for a common user, as those likely to be already known by someone with a basic understanding of English (examples are: day, time, space, November, etc..). These concepts are associable to generic world-knowledge and they can be heuristically identified by analysing the words frequency in the Brown corpus [10] or similar corpora.
- The concepts with a betweenness centrality equal to 0. In fact, filtering these concepts would reduce the average length of an explanation (intended as a path over the ES) without preventing the user from reaching the information it needs.

4 RELATED WORK

In literature we found many works on Question Answering (QA), and only few of them [5, 32, 33] were about (RDF) Knowledge Graphs (KGs). As comparison to our work, we point to the many state-of-the-art deep-learning based QA algorithms implemented by Wolf et al. [30]. With these algorithms, using the whole explanandum as input context would require an impractical amount of time (try it³) for every posed question, in order to obtain very short (e.g. 2-3

³<https://huggingface.co/models?filter=question-answering>

words) answers which quality heavily depends on the selected linguistic model. The practical advantage of our approach is that it is capable of selecting the most relevant text fragments in the context, limiting the search for an answer to very few paragraphs rather than the entire corpus, thus improving efficiency by orders of magnitude. Furthermore our approach can effectively handle very generic questions such as the archetypal ones.

In literature we found many works on explanatory tools, and most of them are focused on one-size-fits-all solutions. For example, Cai et al. [4] propose to use different types of figurative explanations to understand which one is the best, assuming there is always one type of explanation that is the best (or most specific) in average. This attempt to build a sort of hierarchy of explanations is close to what we attempted in Section 3.3, but it clearly lacks any explicit connection to Achinstein’s model. In [18] there is a more evident attempt to combine different types of explanations, implicitly considering them useful as answers to different questions a user might pose to itself. Despite this, we see there is a wanted lack of depth in the ES generated by Kouki et al.’s tool, thus explanations are shallow and cannot be further expanded. Differently from [7], we do not re-organise questions on the go, thus answers to the same (archetypal) questions are the same for every user. Despite this, our system is still capable of taking into account the inevitable interest drift pointed by the user during knowledge acquisition. In this sense our approach is a special kind of organisation-based explanation shown to be highly effective by Pu and Chen in [23].

5 PROOF OF CONCEPT EVALUATION

In order to investigate the usefulness of the proposed solution, we designed a user study on a XAI-based credit approval system. Being interested in modelling an explanatory process for producing user-centric explanatory software and in quantifying the difference it bears in terms of effectiveness with respect to non-pragmatic approaches, our solution draws from state-of-the-art philosophical theories of explanation. What we show is that an abstract philosophical theory of explanations such as Achinstein’s can be beneficially implemented into a concrete software, as a question answering process. We do it by theorising that illocution in explaining involves the act of pertinently answering archetypal questions. We then show how estimations of the pertinence of non-pragmatic answers can result in a globally pragmatic (user-centred) explanatory process. If our theory is correct, through the identification of a minimal set of archetypal questions, it is possible to obtain a generator of explanatory overviews generic enough to be able to significantly ease the acquisition of knowledge, regardless of the specific user but depending instead on a fairly broad category of selected users (e.g. people able to read and understand English, hence capable of performing common-sense reasoning), thus resulting in a user-centred explanatory tool that is more effective than its non-pragmatic counterpart on the same explanandum. In other terms, our hypothesis is that users, with an explanatory tool based on our model, can understand relevant information more effectively than users without it. Despite the specificity of our hypothesis we also want to measure the other aspects of usability (efficiency and satisfaction), in order to have a deeper understanding of the explanatory tool.

In short, we adopt the definition of usability as the combination of *effectiveness*, *efficiency*, and *satisfaction*, as per ISO 9241-210, that defines usability as the “extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” [9]. Effectiveness (“accuracy and completeness with which users achieve specified goals”) and efficiency (“resources used in relation to the results achieved. [...] Typical resources include time, human effort, costs and materials.”) can be assessed through objective measures (in our case, pass vs. fail at domain-specific questions and time to complete tasks, respectively). Satisfaction, defined as “the extent to which the user’s physical, cognitive and emotional responses that result from the

use of a system, product or service meet the user’s needs and expectations”, is a subjective component and it needs a direct confrontation with the user (in our case through the use of the System Usability Scale (SUS) questionnaire [3]).

In order to verify our hypothesis, we test our pipeline on a XAI-powered credit approval system [17] that is using the “FICO Explainable Machine Learning Challenge” dataset [15] and probe into it from the perspective of different users. More in detail, we compare a static explanatory tool for *post-hoc* explanations named Contrastive Explanations Method (CEM) [6], with an interactive version built upon our model and that is based on the same informative contents easily reachable by any user of the static explanatory tool (e.g. searching on the web, reading the output of the tool, etc.). The aforementioned interactive explanatory tool is specifically designed to be an extension of its static version, so that comparing the usability of those two tools would indirectly allow us to measure the usability of the overviews generator presented throughout the whole paper.

5.1 The Credit Approval System

As the context for our evaluation, we took a credit approval system for bank customers, designed by IBM, that is using a static explanatory tool for *post-hoc* explanations named Contrastive Explanations Method (CEM) [6]. We created a web-page mimicking, with high fidelity, the original IBM’s system [17], thanks to the available documentation. In Figure 2 we see an example of static explanation (the initial explanans) produced by the (baseline) credit approval system. A scenario was introduced to the study’s participants. In this scenario the customer of a bank (John) wants to get an explanation about its loan application, in order to understand why it has been denied by the automated credit approval system. The bank uses an artificial neural network, as the automatic process, to decide whether to approve the loan request or not, and it uses the CEM algorithm to create a contrastive explanatory information filled into a textual template provided as explanation. This explanation aims at helping customers understand whether they have been treated fairly, providing insights into ways to improve their qualifications so as the likelihood of a future acceptance can be increased.

As a comparison with the baseline explanatory tool, we created an interactive version of it that is using our pipeline for generating explanatory overviews⁴. This interactive version is obtained by means of a javascript module that automatically annotates the static explanation generated by the baseline tool, making it interactive. As result, the user can click on annotations thus opening an overview modal containing annotations clickable as well, as shown in Figure 3. The content of the overview modal is obtained by the system by interrogating a python server exposing the necessary APIs to interact with the pipeline described in Section 3. The overall extension is designed to be as generic as possible, in other terms it would be possible to use it on any explanatory system providing textual explanations and rich enough documentation (as IBM’s), because the aforementioned annotation process is fully automated, as described in Section 3. The documentation we used for building the ES was taken mainly from “myfico.com” (50 web-pages) and IBM’s website (10 web-pages), plus a few extra web-pages from “forbes.com”, “bankrate.com” and Wikipedia, that overall resulted in roughly 57.000 different triplets.

5.2 User Study

We recruited 103 different participants (57 males, 44 females, 2 unknowns, ages 18-55) on the online platform Prolific [22]. All the participants were recruited among those who: (1) are resident in UK, US or Ireland; (2) have a Prolific’s

⁴Source code and additional material is available at <https://github.com/Francesco-Sovrano/From-Philosophy-to-Interfaces-an-Explanatory-Method-and-a-Tool-Inspired-by-Achinstein-s-Theory-of-E>, for reproducibility purposes.

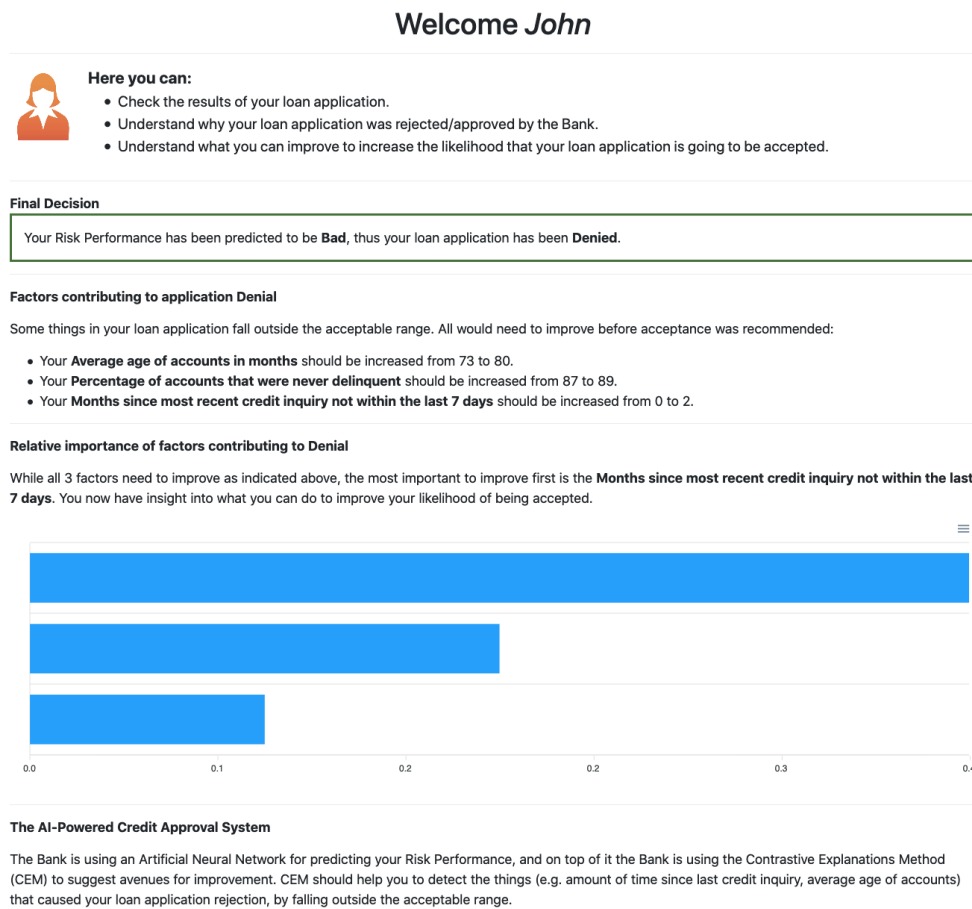


Fig. 2. **Baseline:** A screenshot of the credit approval system for bank customers used as baseline for the proof of concept.

acceptance rate greater or equal to 75%.⁵ Participants were randomly allocated to test only one of the two versions of the credit approval system: either with (Group 2) or without (Group 1) our extension. 51 participants ended in Group 1 and 52 in Group 2. Participants were invited to answer two sets of questions, in English. Participants were told that answering those questions would have taken an average time of 4 minutes. The first set of questions (S1) was domain-specific, so as to measure the *effectiveness* and *efficiency* of the explanations. Three participants were discarded, by answering (more or less) randomly/nonsensically to most of S1. At the end of the attention-check process, 97 participants were kept. The questions in S1 are: (1) What did the Credit Approval System decide for John? (2) What is an inquiry (in this context)? (3) What type of inquiries can affect John's score, the hard or the soft ones? (4) Provide an example of hard inquiry. (5) How can an account become delinquent? (6) Which specific process was used by the Bank to automatically decide whether to assign the loan? (7) What are the known issues of the specific technology used by the Bank (to automatically predict John's risk performance and to suggest avenues for improvement)?

⁵Mainly because they are unlikely to answer poorly/randomly to questions.

Explaining Decisions on Loan

Months since most recent credit inquiry not within the last 7 days Inquiry

Inquiry

- Abstract:** An inquiry is an item on a consumer's credit report that shows that someone with a permissible purpose has previously requested a copy of the consumer's report.
- There are 23 different examples of Inquiry: [\[More..\]](#)
- It has been found in 27 sources: [\[More..\]](#)

Why?

- The number of credit inquiries in last 6 months excluding the last 7 days is used to compute the FICO Score and to decide whether to assign the loan. Excluding the last seven days removes inquiries that are likely due to price comparison shopping. [\[More..\]](#)

What for?

- An inquiry is when a lender makes a request for your credit report or score. FICO Scores have been carefully designed to count only those inquiries that truly impact credit risk. Not all inquiries are related to credit risk, FICO's research shows. [\[Less..\]](#)
 - That said, there are definitely a few things to be aware of depending on the type of credit you are applying for. When you apply for credit, a credit check or "inquiry" can be requested to check your credit standing. Let's take a look at the common inquiries you might find on your credit report. [\[Less..\]](#)

Pertinence	Source	Document
68.92%	That said, there are definitely a few things to be aware of depending on the type of credit you are applying for. When you apply for credit, a credit check or "inquiry" can be requested to check your credit standing. Let's take a look at the common inquiries you might find on your credit report.	MyFICO - minimizing the effects of credit shopping

- An inquiry is when a lender makes a request for your credit report or score. Although FICO Scores only consider inquiries from the last 12 months, inquiries

Fig. 3. **Proof of Concept:** Example of overview displaying relevant information about a concept that is directly involved in the initial explanans.

The answers to these questions were scored by a human evaluator as correct (score 1) or not (score 0). For example, given the 1st question, a common correct answer was “Denial” and a common wrong one was “The AI-Powered Credit Approval System”. In order to ensure that both the explanatory systems (the user-centred one and the baseline) were on an explanandum based on the same informative contents, the participants of Group 1 were explicitly allowed to search the web for correct answers, and many of them did, as shown by the good results on effectiveness of Group 1. Furthermore, as discussed also in Section 6, questions 2, 3, 4, 5 and 7 have been designed so that providing the correct answers would require the exploration of at least 2 or 3 different overviews.

The second set of questions (S2) is a modified System Usability Scale (SUS) questionnaire [3] used to measure the participants' *Satisfaction* with the explanatory systems. The original SUS questionnaire has been modified, in order to specifically target the usability of the explanatory systems, by changing every occurrences of the word “system” with the word “explanatory system”. Participants were asked to answer S2 by considering only the effort they had for answering S1. Of the 97 selected participants:

- 48 (27 males, 20 females, 1 unknown, ages 18-55) tested the baseline system (V1);
- 49 (24 males, 24 females, 1 unknown, ages 18-53) tested the system extended with our pipeline (V2).

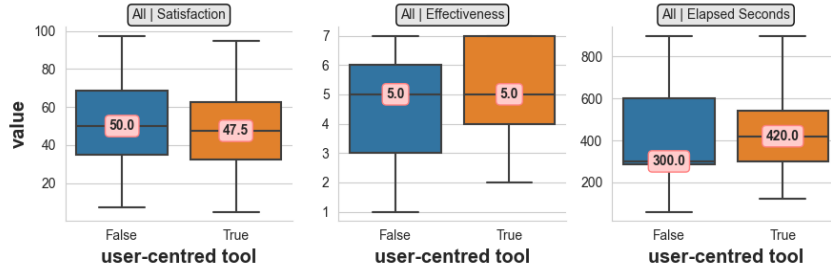


Fig. 4. Results are shown in the form of box plots (25th, 50th, 75th percentile, and whiskers covering all data and outliers). The numerical value of medians is shown inside pink boxes. Results for V2 are in orange, while results for V1 are in blue.

6 RESULTS

Due to the limited number of samples, we choose to not make assumptions of parametrisation in the data⁶. Every sample is grouped according to the adopted version of the credit approval system: V2 is the system using our user-centred explanatory tool; V1 is the baseline. We defined our results measures as: **Satisfaction**, **Effectiveness**, and **Elapsed Seconds** (that are inversely proportional to Efficiency). In Figure 4 we show the resulting box-plots for every given measure. According to these box-plots, results seem to indicate that V2 is likely to be more effective than V1, but less satisfying. We then performed a one-sided Mann-Whitney U-Test (MW) on the other scores, under the alternative hypothesis that the elapsed seconds and the effectiveness of V2 is greater than V1 and the satisfaction of V2 is lower than V1. The results show that the global (between-subjects):

- **Effectiveness score** of V2 is statistically greater than V1. According to MW ($U=931.0$, $p=0.036$) there is enough statistical evidence⁷.
- **Satisfaction score** in V2 seems lower than V1, but there is not enough statistical evidence, according to MW ($U=1272.0$, $p=0.245$).
- **Elapsed Seconds** in V2 seems greater than V1, but there is not enough statistical evidence, according to MW ($U=1017.5$, $p=0.125$).

The obtained results highlight a poor correlation between objective (effectiveness) and subjective (satisfaction) metrics, an interesting phenomenon that has been object of study in many works: [11, 16, 21]. We believe that a lower satisfaction is due to the learning process required during the phase of understanding an explanation. In fact, learning can require high cognitive resources. The median amount of time spent in learning could be estimated by computing the difference between the medians of the E elapsed seconds of V2 and V1, and that is roughly 120 seconds (definitely something). Filtering information and analysing it can be a mentally onerous process. Given the evident intrinsic link between explaining and teaching (and therefore learning) it cannot be excluded that, for a user not strongly interested in achieving the objectives defined for the explanatory process (and therefore a user who is paid to carry out a usability test, like ours), a sufficiently rich explanatory tool is perceived as less satisfactory (because the less time they take to finish the test, the better is for them) than an explanatory tool that follows a one-size-fits-all approach (thus hiding too much complexity, reducing the effectiveness of the system). This intuition is supported by a statistically significant difference

⁶Anonymised data is available at <https://github.com/Francesco-Sovrano/From-Philosophy-to-Interfaces-an-Explanatory-Method-and-a-Tool-Inspired-by-Achinstein-s-Theory-of-E>, for reproducibility purposes.

⁷Assuming $p < 0.05$ is enough.

between the answers to the question number 8 of the SUS, where the hypothesis that the information provided by our explanatory tool is more cumbersome than the baseline is confirmed by a MW test ($U=822.0$, $p=0.004$). In other words, our tool provides a more effective exploration of the ES, but it fails to satisfactorily hide all its complexity, according to the participants. In fact, the archetypal questions used to structure the ES are very generic and the user shall choose those most relevant to him. This phase of choosing questions (and therefore explanations) is mentally expensive (even if working with summaries) and it requires having to read snippets of explanation perhaps not useful for the task. Said that, the exploratory system we proposed could be suitably designed to facilitate an efficient retrieval of answers to pre-defined questions (such as those of the effectiveness questionnaire) thus minimizing the number of clicks and consequently increasing the overall satisfaction. But to properly evaluate our hypothesis, we decided to design the evaluation questionnaires in such a way that providing the correct answers⁸ would have required a minimum of ES exploration, i.e. the exploration of at least 2 or 3 different overviews. This, obviously, affected the average satisfaction and efficiency but allowed us a better comparison with the baseline and to avoid studying the optimal case (which, by definition, would have given optimal results, but useless for the scientific purposes of the paper).

7 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a new method for explanations in AI and, consequently, a tool to test its expressive power within a user interface. Being interested in modelling an explanatory process for producing user-centric explanatory software, and in quantifying the difference it bears in terms of effectiveness with respect to non-pragmatic approaches, our solution drawn from state-of-the-art philosophical theories of explanation. Among the few philosophical theories of explanation, we identified the one that, we believe, is mostly convertible into a practical model for user-centric explanatory software: Achinstein's. But Achinstein's is an abstract illocutionary theory of explanation, therefore we proposed a way to concretely implement illocution as the act of *pertinently answering* archetypal questions (e.g. Why? What for? How? When? etc.), removing from the "equation" the need for a model of human intention. What we showed is that an abstract philosophical theory of explanations can be beneficially implemented into a concrete software, as a question answering process. In fact, through the identification of a minimal set of archetypal questions, it is possible to obtain a generator of explanatory overviews generic enough to be able to significantly ease the acquisition of knowledge, regardless of the specific user but depending instead on a fairly broad category of selected users, thus resulting in a user-centred explanatory tool that is more effective than its non-pragmatic counterpart on the same explanandum. In other terms, our hypothesis was that users, with an explanatory tool based on our model, can understand relevant information more effectively than users without it. In order to test our hypothesis, we had to invent a new pipeline of AI algorithms (briefly summarised in figure 1) and run a user study on it. This pipeline is able to organize the information contained in non-structured documents written in natural language (e.g. web pages, pdf, etc.), allowing efficient information clustering, according to a set of archetypal questions, aiming to build a sufficiently rich and effectively explorable Explanatory Space (ES) for the automated generation of user-centred explanations. We tested our hypothesis on a XAI-powered credit approval system [17], comparing a static explanatory tool for *post-hoc* explanations named Contrastive Explanations Method (CEM) [6], with its interactive version based on our model. The results of the user study, involving more than 100 participants, showed that our proposed solution produced a statistically relevant improvement on effectiveness ($U=931.0$, $p=0.036$) over the baseline, thus giving evidence in favour of our hypothesis.

⁸To questions 2,3,4,5, and 7 of the effectiveness questionnaire.

As a future development, we therefore plan to test our work against other baselines, and to integrate a more classical QA system into the system presented here. If the user could ask more specific questions (also obtaining the appropriate answers), then he could save himself part of the effort to search for information, through a more natural mechanism. This intuition is also supported by the findings of Madumal et al. [19]. In general, however, it is not always possible to evaluate the effectiveness of a more sophisticated QA system through the use of a questionnaire, because the first thing a user would do would be asking the questions in the questionnaire.

REFERENCES

- [1] Peter Achinstein. 1983. *The nature of explanation*. Oxford University Press on Demand.
- [2] Peter Achinstein. 2010. *Evidence, explanation, and realism: Essays in philosophy of science*. Oxford University Press.
- [3] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.
- [4] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 258–262.
- [5] Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. 2019. KBQA: learning question answering over QA corpora and knowledge bases. *arXiv preprint arXiv:1903.02419* (2019).
- [6] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in neural information processing systems*. 592–603.
- [7] Cecilia Di Sciascio, Vedran Sabol, and Eduardo E Veas. 2016. Rank as you go: User-driven exploration of search results. In *Proceedings of the 21st international conference on intelligent user interfaces*. 118–129.
- [8] Igor Douven. 2012. Peter Achinstein: Evidence, Explanation, and Realism: Essays in Philosophy of Science.
- [9] International Organization for Standardization. 2010. *Ergonomics of human-system interaction: Part 210: Human-centred design for interactive systems*. ISO.
- [10] W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Letters to the Editor* 5, 2 (1979), 7.
- [11] Erik Frøkjær, Morten Hertzum, and Kasper Hornbæk. 2000. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 345–352.
- [12] Bernhard Ganter and Rudolf Wille. 2012. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media.
- [13] Carl G Hempel et al. 1965. Aspects of scientific explanation. (1965).
- [14] John H Holland, Keith J Holyoak, Richard E Nisbett, and Paul R Thagard. 1989. *Induction: Processes of inference, learning, and discovery*. MIT press.
- [15] Steffen Holter, Oscar Gomez, and Enrico Bertini. [n.d.]. FICO Explainable Machine Learning Challenge. ([n. d.]).
- [16] Kasper Hornbæk. 2006. Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies* 64, 2 (2006), 79–102.
- [17] IBM. 2019. AI Explainability 360 - Demo. https://aix360.mybluemix.net/explanation_cust. Online; accessed 29-Mar-2020.
- [18] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 379–390.
- [19] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. *arXiv preprint arXiv:1903.02409* (2019).
- [20] GR Mayes. 2005. Theories of Explanation. The Internet Encyclopedia of Philosophy.
- [21] Jakob Nielsen. 2012. User satisfaction vs. performance metrics. *Nielsen Norman Group* (2012).
- [22] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.
- [23] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces*. 93–100.
- [24] Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LARQA: Language-agnostic answer retrieval from a multilingual pool. *arXiv preprint arXiv:2004.05484* (2020).
- [25] Wesley C Salmon. 1984. *Scientific explanation and the causal structure of the world*. Princeton University Press.
- [26] Wilfrid Sellars. 1963. Philosophy and the scientific image of man. *Science, perception and reality* 2 (1963), 35–78.
- [27] Francesco Sovrano, Monica Palmirani, and Fabio Vitali. 2020. 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, 143–153.
- [28] Francesco Sovrano, Fabio Vitali, and Monica Palmirani. 2020. Modelling GDPR-Compliant Explanations for Trustworthy AI. In *International Conference on Electronic Government and the Information Systems Perspective*. Springer, 219–233.
- [29] Bas C Van Fraassen et al. 1980. *The scientific image*. Oxford University Press.
- [30] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

- [31] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307* (2019).
- [32] Weiguo Zheng, Hong Cheng, Jeffrey Xu Yu, Lei Zou, and Kangfei Zhao. 2019. Interactive natural language question answering over knowledge graphs. *Information Sciences* 481 (2019), 141–159.
- [33] Lei Zou, Ruizhe Huang, Haixun Wang, Jeffrey Xu Yu, Wenqiang He, and Dongyan Zhao. 2014. Natural language question answering over RDF: a graph data driven approach. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 313–324.