



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

From Philosophy to Interfaces:

*an **Explanatory Method** and a **Tool**
Inspired by **Achinstein's Theory of
Explanation***

F. Sovrano and F. Vitali

Department of Computer Science (DISI)
University of Bologna

Background

- Nowadays **Algorithmic Decision Making** is changing industry. This is why people and Countries (EU, California, etc..) begun to be concerned about **the impact** that may have on everyone's lives.
- This is one of the reasons that gave birth to the so-called **Right to Explanation** “recently” introduced in the EU legislation, and further explored by the **High-Level Expert Group on Artificial Intelligence** (AI-HLEG) established in 2018 by the **EU Commission**.
- Some of the given guidelines (by the AI-HLEG) point to the need for **user-centrality** in **explanations**.
- The fact that **explanations** are **better** conceived when they are **pragmatic** is also **shared** by all the most recent **philosophical theories** of explanation.



Explanation in Philosophy

- Explanation in **philosophy** has been conceived within the following 5 traditions:
 - Causal Realism
 - Constructive Empiricism
 - Ordinary Language Philosophy
 - Cognitive Science
 - Naturalism and Scientific Realism
- The first one is not pragmatic, the others are pragmatic.



XAI Literature

- Apparently a lot of literature on user-centred XAI is now steering towards the **interpretation** of explanations given by **Cognitive Science**.
- In our paper, we propose a **new approach** to explanations in AI that is mainly inspired by **Ordinary Language Philosophy** instead of Cognitive Science.



Achinstein's Illocutionary Theory of Explanations

- Achinstein's model is probably the most prominent example of Theory of Explanations within the scope of **Ordinary Language Philosophy**.
- According to this 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 a **question**.
- **But, what is illocution, precisely?**



How to Model Illocution

- Our hypothesis is that **illocution** (in explanations) is equivalent to the act of **pertinently answering to archetypal questions** (e.g. Why? What for? How? When? etc..).
- 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 explaine's background **knowledge**, the **objectives** of the explanatory process and the given **context**.
- 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 explaine's objectives.



How to Achieve User-Centrality

- So, how can we measure the pertinence of an answer? The likelihood of pertinence to an answer can be **quantitatively estimated** with strong-enough **statistical evidence**; for example collected in deep language models, as suggested by existing literature.
- But we are defining pertinence by means of statistical tools, where is user-centrality? A statistical definition of pertinence does not preclude a pragmatic **iterative** explanatory process that is **locally non-pragmatic** but **globally pragmatic**. In other words: an explanans at step t is **non-pragmatically** built by organising knowledge according to a statistical pertinence, but the sequence of explanans is **pragmatically** chosen by the users.
- So, we can design an explanatory process as an **iterative** process where user-centrality is in the way answers are visited, explored, expanded (and eventually asked).



From Theory to Practice

- So far we discussed about theory, how can we build this iterative process in practice?
To do so we need an algorithm for the extraction and organisation of Explanatory Spaces (or ES) out of explainable documents.
- An ES is defined as the set of all possible explanations (on an explainable explanandum) reachable by a user, through an explanatory process, starting from an initial explanans, via a pre-defined set of actions.



The ES as a Graph

- Therefore an ES is a **graph** of interconnected bits of explanation that are about **different aspects** of the **explanandum**.
- To every aspect is associated an **information cluster** explaining that aspect.
- An **explanatory process** over an ES can be defined by the **choice** of appropriate **heuristics** for **visiting** and **organising** the space.

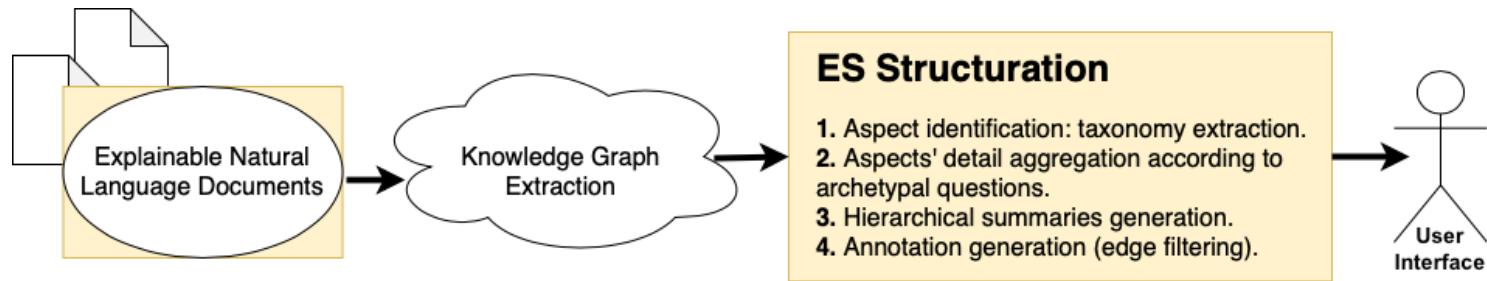


The ARS Heuristics

- To this end, we identified the **ARS heuristics**, that are: Abstraction, Relevance and Simplicity.
- **Abstraction** is for selecting and organising the different aspects of the explanandum (that are the external nodes of the ES). For instance, the core aspects to explain might be the different concepts or entities mentioned in the explanandum.
- **Relevance** is for organising the information about the aspects (the internal nodes of the ES), according to the archetypal questions.
- **Simplicity** is for filtering or grouping the edges of the graph. For instance, some pieces-of-information can be hidden, because redundant, or some others may be grouped together to simplify the visualisation.



The Pipeline



In order to implement these ARS heuristics, we created a pipeline for:

1. Extracting a **knowledge graph of aspects** out of natural language documents.
2. **Building information clusters** for every aspect, according to the identified **archetypal questions**.
3. **Presenting** information clusters through a simplified **hierarchy of expandable summaries**.

The Experiment - Proof of Concept Evaluation

- We put this pipeline at work.
- If our **hypothesis** is correct, through the identification of a minimal set of **archetypal questions**, it is possible to obtain a generator of explanations that is good enough to significantly **ease the acquisition of knowledge**, thus resulting in a user-centred explanatory tool that is **more effective** than its non-pragmatic counterpart (on the same explanandum).
- To verify this hypothesis, we designed a **user study on a baseline credit approval system**.



The Baseline

Welcome *John*



Here you can:

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

Final Decision

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

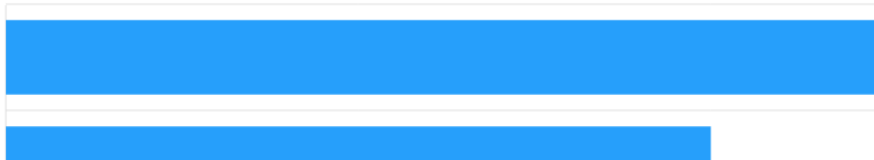
Factors contributing to application Denial

Some things in your loan application fall outside the acceptable range. All would need to improve be

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

Relative importance of factors contributing to Denial

While all 3 factors need to improve as indicated above, the most important to improve first is the **Mo**
7 days. You now have insight into what you can do to improve your likelihood of being accepted.



- The baseline **credit approval system** we are considering has been designed by IBM and it consists in a XAI used to generate a **static** explanation about the system's decision.
- This explanation is meant to explain to a customer what to change in order to get a loan application accepted.
- On the left you may see a screenshot of such explanation.




Our System

- For the sake of the experiment we created an **interactive version of that baseline** that is using our pipeline for generating explanatory overviews of the explanandum's aspects.
- This interactive version is obtained by means of a javascript module that **automatically** annotates the **static** explanations, by using the extracted knowledge graph.
- The **documentation** we used for building the knowledge graph was taken from roughly **60 online web pages**.



Our System

Explaining Decisions on Loans



Here you can:

- Check the reason for the decision
- Understand the factors contributing to the decision
- Understand the relative importance of the factors

Final Decision

Your Risk Performance[®] has been classified as Low.

Factors contributing to applying for a loan

Some things in your loan application that we took into account:

- Your Average age of accounts
- Your Percentage of accounts in good standing
- Your Months since most recent credit inquiry

Relative importance of factors

While all 3 factors need to improve your Risk Performance[®], the Months since most recent credit inquiry is the most important factor. You now have insight into why this is the case.

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 <u>aware</u> of depending on <u>the type of credit</u> you are <u>applying</u> for. When you <u>apply</u> for <u>credit</u> , a <u>credit check</u> or "inquiry" can be requested to <u>check your credit standing</u> . Let's take a look at the common inquiries you might find on <u>your credit report</u> .	<u>MyFICO - minimizing the effects of credit shopping</u>

- 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

- As result, the user can **interactively explore** the ES by clicking on annotations (the underlined syntagms), opening an overview containing further annotations.
- As we can see, on the image on the left, information is organised according to the **ARS heuristics**.
- Aspects are explained through expandable archetypal answers summarising more punctual information.

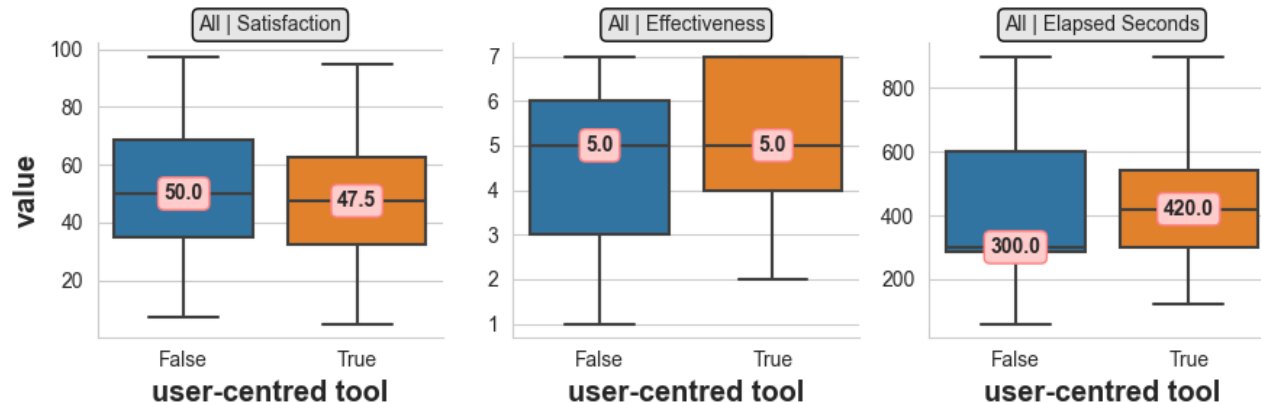


The User Study

- For the user study we recruited more than **100** different participants on the online platform **Prolific**.
- The first half (Group 1) of the participants tested the baseline system, while the second one (Group 2) tested our extension.
- Participants were invited to answer **two sets of questions**: a SUS questionnaire and some domain-specific questions.
- The set of **domain-specific** questions was used to measure the **effectiveness** and **efficiency** of the explanatory tools. **Importantly**: More than 70% of these questions were designed so that the correct answers were not immediately available in the initial explanans, forcing the users to visit multiple nodes of the ES.
- The participants of Group 1 were **explicitly allowed to search the web** for the correct answers (that were **available online** on the websites used for the extraction of the knowledge graph).



Experimental Results



- We chose to **not** make **assumptions** on the **distribution** of the collected **data**. This choice excluded a lot of statistical tests. So we performed a few one-sided **Mann-Whitney U-Tests**. The results showed that:
 - **Effectiveness** (that is the number of correct answers) was **significantly greater** than the baseline with a p-value lower than 0.05 ($p=0.036$).
- There was **not enough statistical evidence** for claiming anything on **Satisfaction** and **Efficiency**.



Conclusions

- We proposed an original and **generic paradigmatic** explanatory AI inspired to **Achinstein's illocutionary theory of explanations**, where most of literature usually prefers to focus on **Cognitive Science**.
- We proposed a way to concretely **implement illocution** as the act of **pertinently answering archetypal questions** (e.g. Why? What for? How? When? etc..).
- We tested our theory inventing a **new pipeline** of AI algorithms and running a **user study** on a realistic Credit Approval System, showing a **statistically significant improvement in effectiveness** over the non-paradigmatic explanatory AI provided by IBM.





ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Francesco Sovrano

Dipartimento di Informatica – Scienze e Ingegneria
Alma mater – Università di Bologna

francesco.sovrano2@unibo.it

www.unibo.it