ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Usability, Explanatory Artificial Intelligence and Legal Informatics

Fabio Vitali [1], Francesco Sovrano [2]

[1] *Dept of Computer Science – University of Bologna*

[2] *Faculty of Informatics – Università della Svizzera Italiana*

Summary

The legislative framework of explanations

What are explanations?

Some principles from UUXD

Some recommendations

Legislative framework

GDPR (2016)

The GDPR is technology-neutral, so it does not directly refer to AI. However, several provisions are highly relevant to the use of AI (or any other software) for automated decision-making processes. For instance, the most important of these provisions are:

- Article 5(1) point (a), that requires personal data processing to be fair, lawful, transparent, necessary and proportional.
- Article 12, which defines the obligations for transparent communication and the modalities for data subjects to exercise their rights.
- Articles 13, 14 and 15, that give individuals the right to be informed of solely automated decision-making, meaningful information about the logic involved, and the significance and envisaged consequences for the individual.
- Article 22, that gives individuals the right not to be subject to a solely automated decision producing legal or similarly significant effects.
- Article 22(3), that obliges organizations to adopt suitable measures to safeguard individuals when using solely automated decisions, including the right to obtain human intervention, to express his or her view, and to contest the decision.

Article 27 of the Digital Services Act (2022)

Recommender system transparency

1. Providers of online platforms that use recommender systems shall set out in their terms and conditions, in plain and intelligible language, the main parameters used in their recommender systems, as well as any options for the recipients of the service to modify or influence those main parameters.
2. The main parameters referred to in paragraph 1 shall explain why certain information is suggested to the recipient of the service. They shall include, at least:
   a) the criteria which are most significant in determining the information suggested to the recipient of the service;
   b) the reasons for the relative importance of those parameters.
3. Where several options are available pursuant to paragraph 1 for recommender systems that determine the relative order of information presented to recipients of the service, providers of online platforms shall also make available a functionality that allows the recipient of the service to select and to modify at any time their preferred option. That functionality shall be directly and easily accessible from the specific section of the online platform's online interface where the information is being prioritised.

Recital 27 of the AI Act (2024)

27. While the risk-based approach is the basis for a proportionate and effective set of binding rules, it is important to recall the 2019 Ethics guidelines for trustworthy AI developed by the independent AI HLEG appointed by the Commission. In those guidelines, the AI HLEG developed seven non-binding ethical principles for AI which are intended to help ensure that AI is trustworthy and ethically sound. The seven principles include human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being and accountability. Without prejudice to the legally binding requirements of this Regulation and any other applicable Union law, those guidelines contribute to the design of coherent, trustworthy and human-centric AI, in line with the Charter and with the values on which the Union is founded. According to the guidelines of the AI HLEG, [...] transparency means that AI systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights. [...]

Article 14 of the AI Act (2024)

Human Oversight

1. High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use.

[...]

4. For the purpose of implementing paragraphs 1, 2 and 3, the high-risk AI system shall be provided to the deployer in such a way that natural persons to whom human oversight is assigned are enabled, as appropriate and proportionate:
   a) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance;
   b) to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;
   c) to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available;
   d) to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system;
   e) to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state.

[...]

Article 86 of the AI Act (2024)

Right to explanation of individual decision-making

1. Any affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system listed in Annex III, with the exception of systems listed under point 2 thereof, and which produces legal effects or similarly significantly affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken.
2. Paragraph 1 shall not apply to the use of AI systems for which exceptions from, or restrictions to, the obligation under that paragraph follow from Union or national law in compliance with Union law.
3. This Article shall apply only to the extent that the right referred to in paragraph 1 is not otherwise provided for under Union law.

What are explanations?

Introduction

Transparency (def)

– [An AI tool] is transparent if the processes that extract model parameters from training data and generate labels from testing data can be described and motivated by the system's designer.

Interpretability (def)

– Interpretability is presenting some of the properties of [an AI tool] in understandable terms to a human.

Explainability (def)

– [An AI tool is explainable if it provides a] collection of features of the interpretable domain, that, for a given example, contributed to produce a decision.

– [in explainable ML] these definitions assume implicitly that the concepts expressed in the understandable terms composing an explanation are self-contained and do not need further explanations.

from R. Roscher, B. Bohn, M. F. Duarte and J. Garcke, "Explainable Machine Learning for Scientific Insights and Discoveries," in IEEE Access, vol. 8, pp. 42200-42216, 2020

Explainability is a feature of the AI tool, not of the interaction between the tool and its user.

But if we agree on this idea of explainability, then we need to consider a further step: *actual explanations.*

What do people mean with explanations

Theory	Explanations
Causal Realism	Descriptions of causality, expressed as chains of causes and effects.
Constructive Empiricism	Contrastive information that answers why questions, allowing one to calculate the probability of a particular event relative to a set of (possibly subjective) background assumptions.
Ordinary Language Philosophy	Answers to questions (not just *why* ones) given with the explicit intent of producing understanding in someone, i.e., the result of an illocutionary act.
Cognitive Science	Mental representations resulting from a cognitive activity. They are information which fixes failures in someone's mental model.
Naturalism and Scientific Realism	Information which increases the coherence of someone's belief system, resulting from an iterative process of confirmation of truths aimed at improving understanding.

What do we mean with explanations

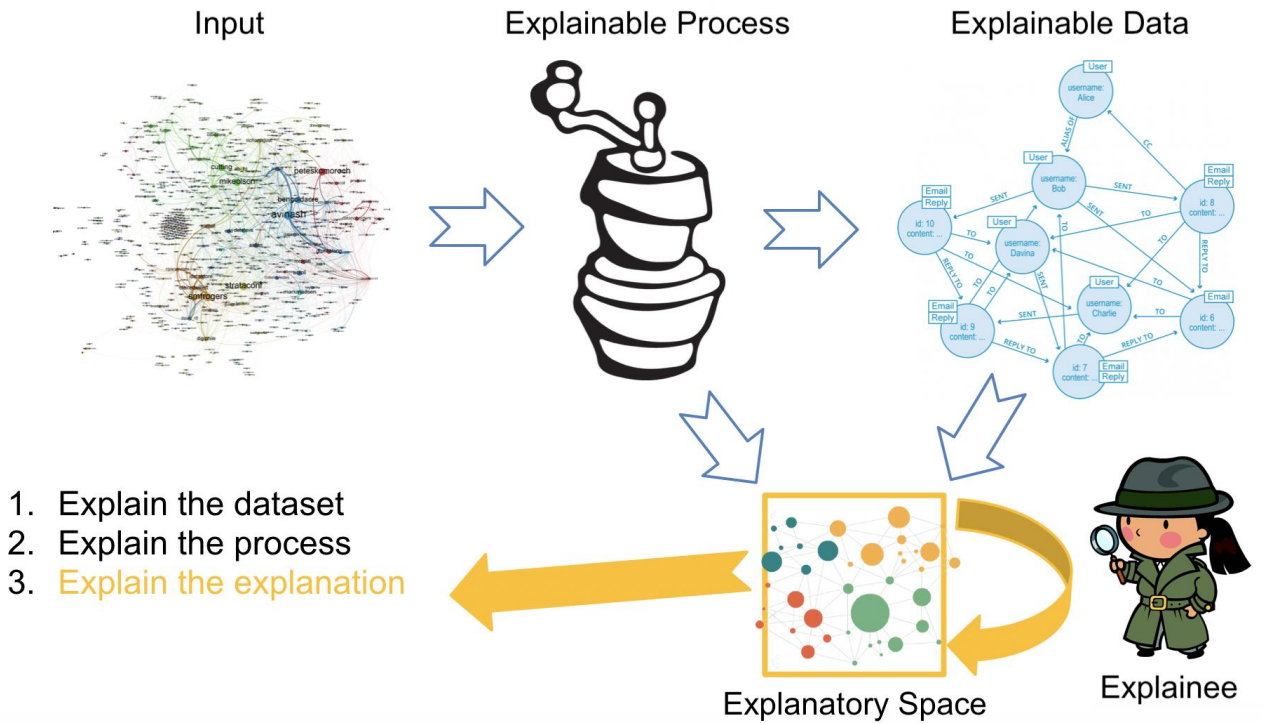
Explanations as question answering

Archetypal questions: *why, how, what, who, when, where*, etc., and derivations: *why not, what for, what if, how much*, etc.

The role of the user: user-centred, goal-oriented; the user decides what is useful.

When is an explanation adequate?

- When the user is subjectively satisfied by the quality of the explanation (*satisfaction*):
- When the user can answer to questions correctly (*efficacy*);
- When the user can carry out correctly tasks (*effectiveness*).



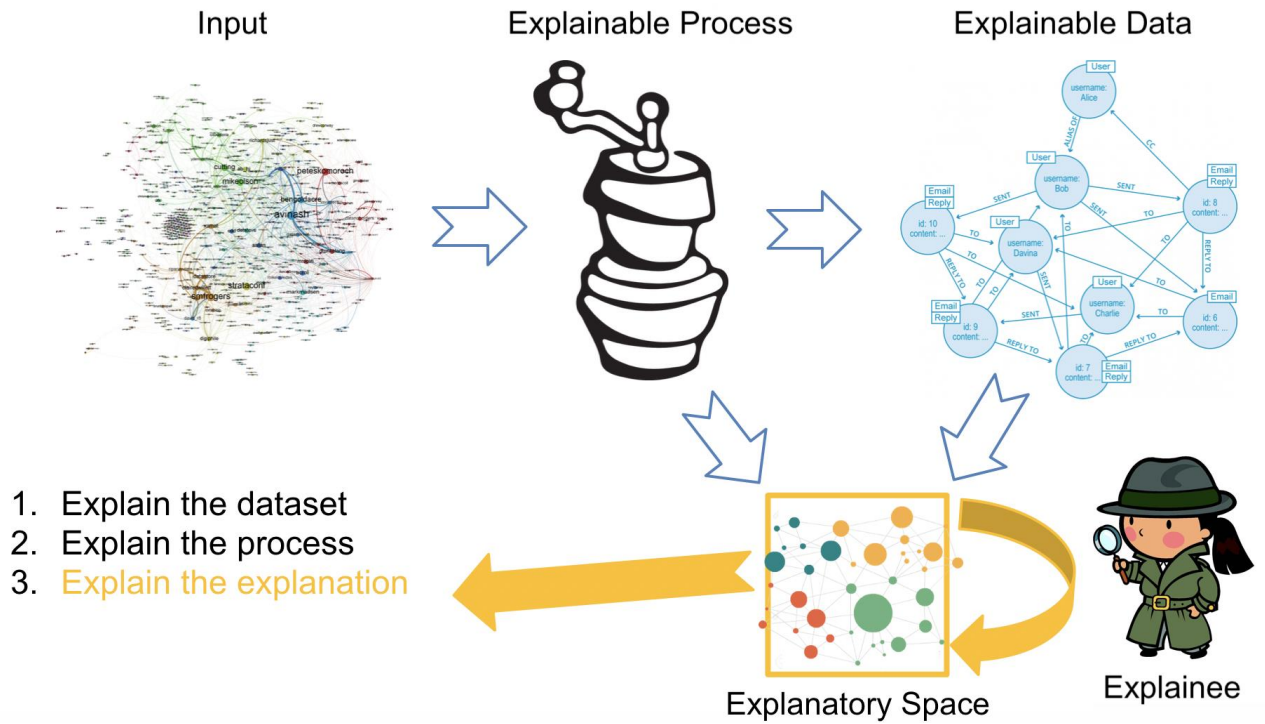
What do we mean with explanations

Normal XAI explanations: one-size-fits-all answers.

- ***Normal XAI***: answers one or a few questions
- ***Selected narratives***: answers all questions of one type (e.g., how or why)
- ***Exhaustive Explanatory Closures***: answers all possible questions about the answer (1st level), about how it came to the answer (2nd level), about how it created the previous explanations (3rd level).

EEC has all the answers, but it creates a huge ***explanatory space*** that is overwhelming to humans needing to find efficiently and effectively an answer to a specific problem they have.

An explanation is an itinerary within the explanatory space that fits the user's needs.



Some concepts from UUXD

Goal-orientedness: what is the explanation being used for?

Consider a bank's CCTV system as a tool for the police to investigate robberies.

The CCTV system is composed of many high-quality cameras and records up to 96 hours of video.

We can assume that we will find good images of anything happening in or around the bank.

The problems is:

- we do not need 96 hours of video from 27 cameras - we only need four minutes from the three cameras that were actually recording the action of the robbery.
- we do not need to see the faces of the robbers: we know that they had masks on. We need to know how many robbers there were, if they wore something conspicuous or had some physical features that could let us identify them: tattoos, hair, height, etc.
- Maybe we need to know the brand and the plate of the getaway car, or the direction they left towards
- Maybe there was no robbery at the bank, but at the liquor store nearby, and the bank's CCTV needs to be used for tasks unrelated to the bank at all.

These needs are known to the investigators, not to the CCTV system. The same system can be used for different goals which the system, and its designers, are completely unaware of.

Goal-orientedness

Goal-orientedness is therefore the ability of a tool to accommodate a variety of uses that satisfy the actual goals of its intended user, and not simply a collection of (easy to activate) functions completely independent of its intended uses.

In other words, understanding your users is as important as understanding the internal model of your AI tool and provide explainable, white-box access to its workings.

Goal-oriented design divides goals in three main categories:

- ***Experience goals***: how I expect to feel like while I use the tool. Pre-conscious, connected to our visceral processing level - feeling smart, stimulated, have fun, be in control, interested, not bored. Very short-term relevance.
- ***End goals***: how I want to use the tool. The motivations of the user when carrying out the tasks at hand with the tool. Short-term relevance.
- ***Life goals***: how I want to think about myself ***[also]*** thanks to the tools. Connected to the aspirations and ambitions we have. They are independent of the artefacts, but they can be influenced (positively or negatively) by it. Long-term relevance.

Be sure to not mix them up with non-user goals:

- ***Client / Buyer Goals:*** *Organizational*: make our internal processes more effective, limiting costs, etc., *individual* (eg. parents): be educative, help socialization, foster physical and mental development, etc.
- **Goals of the organization owning the product:** *Commercial* (e.g., increase profit), *technical (e.g.,* data integrity, performance), *organizational:* provide a service, optimize the use of resources, etc.

Techniques for goal-oriented design

User segmentation

- ***Demographic approach***: Segmenting by observable traits (e.g., age, education, locale, profession) to anticipate needs and constraints.
- ***Psychological approach***: Segmenting by cognitive styles, motivations, risk attitudes, and mental models that shape explanation preferences.

User research

- ***Market research***: Studying markets, segments, and competitors to understand demand and prioritize features and messaging.
- ***Contextual inquiry***: In-situ observation and interviewing while users perform real tasks to capture workflows and constraints.
- ***Task analysis***: Decomposing tasks into goals, steps, decisions, and information needs to design efficient explanation flows.

Personas: Evidence-based synthetic users (*archetypes*) representing key segments, used to align design decisions and communication.

Some recommendations

The SAGE-ARS Model

Crafting the explanation process to allow for sense-making (helping the user "understand"), articulating (aggregating information into an explanatory narrative) and evaluating (verifying whether the goals have been reached), we propose the SAGE-ARS model:

Explanatory spaces must be (SAGE)

- ***Sourced:*** faithful to the concept being explained.
- ***Adaptable:*** shaped by the narrative associated to user goals & tasks.
- ***Grounded:*** true representation of (a fragment of) the explanatory space.
- ***Expandable:*** forming an explorable network for extending the narrative according to user's actions.

Navigating the explanatory space should support (ARS):

- ***Abstraction:*** aggregating content via a navigable hierarchy of concepts***.***
- ***Relevance:*** ordering material according to user's goals***.***
- ***Simplicity:*** filtering/prioritizing structures in order to reduce SPINdle

Primitive Actions & Commands

Explanatory interfaces should support:

- ***Open Question Answering***: the user controls the narrative by providing questions and receiving appropriate and focussed answers.
- ***Aspect Overviewing***: The user selects which aspect of the initial answer calls for further exploration, probably according to one or more archetypal questions, with an increasing level of details.
- ***Argumentation***: the user examines critically the answers identifying counter-arguments or weak points that call for further reasoning.

A textual interface should support rapid access to a number of commands:

- ***Sourcing***: where did the tool get the information for this answer?
- ***Adapting***: frame the answer to the specific knowledge of the user.
- ***Grounding***: how is the answer related to the ongoing dialogue?
- ***Expanding*** ("More/Less/Overview"): provide increasing details to the full explanation, allowing the user to decide whether to go further or to go deeper on some topics.

An example of a SAGE-ARS interface (1)

Marco (a 14 years old Italian teenager living in Italy) uses a chat system, and his father, Giulio, wants to remove Marco's subscription without Marco's consent because he is concerned about Marco's privacy when online. An automated decision-making system based on SPINdle* rejects Giulio's request because of the Italian legislative decree 101/2018. What if Giulio wants an explanation of the automated decision?

Giulio's (who?) request to remove Marco's (who?) profile was denied, because of the Italian legislative decree 101/2018 (what?).

This decision was taken by an automated process called SPINdle (what?) starting from a set of known facts (…more…).

SPINdle (what?) reasoned over a LegalRuleML representation (what?) of the GDPR (what?) which is a hierarchy of rules (…more…).

(*) SPINdle is a logic-based AI tool

An example of a SAGE-ARS interface (2)

Marco (a 14 years old Italian teenager living in Italy) uses a chat system, and his father, Giulio, wants to remove Marco's subscription without Marco's consent because he is concerned about Marco's privacy when online. An automated decision-making system based on SPINDle* rejects Giulio's request because of the Italian legislative decree 101/2018. What if Giulio wants an explanation of the automated decision?

Giulio's (who?) request to remove Marco's (who?) profile was denied, because of the Italian legislative decree 101/2018 (what?).

This decision was taken by an automated process called SPINdle (what?) starting from a set of known facts (…more…).

*SPINdle (what?) reasoned over a LegalRuleML representation (what?) of the GDPR (**...hide...**).*

- *The General Data Protection Regulation (GDPR) is Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (…more…).*

The LegalRuleML representation (what?) consists of a hierarchy of rules is a hierarchy of rules (…more…).

(*) SPINDle is a logic-based AI tool

An example of a SAGE-ARS interface (3)

Marco (a 14 years old Italian teenager living in Italy) uses a chat system, and his father, Giulio, wants to remove Marco's subscription without Marco's consent because he is concerned about Marco's privacy when online. An automated decision-making system based on SPINDle* rejects Giulio's request because of the Italian legislative decree 101/2018. What if Giulio wants an explanation of the automated decision?

Giulio's (who?) request to remove Marco's (who?) profile was denied, because of the Italian legislative decree 101/2018 (what?).
*This decision was taken by an automated process called SPINdle (what?) starting from a set of known facts (**...less...**).*

- *The decision taken by SPINdle is composed by a set of logical conclusions (…more…), the premises on which the rules have been applied (…more…), and the following hierarchy of rules used to get those conclusions:*
  - ***R1**: "if X is adult (what?), then X obtains consent (what?)" (… source …)*
  - ***R2**: "if X age is less than 14, then X does not obtain consent (what?)" (… source …)*
  - ***R3**: "if X age is less than 14 and X lives in Italy (where?), then X obtains consent (what?)" (… source …)*
  - ***R4**: "if X does not obtain consent (what?), then X's profile is removed (how?)" (… source …)*
  - *"**R2** rebuts (how?) **R1**" (…source…)*
  - *"**R3** rebuts ("Lex specialis derogat generali" …more… is applied **...hide...**) **R2**" (… source …)*

(what if?)

SPINdle (what?) reasoned over a LegalRuleML representation (what?) of the GDPR (what?) which is a hierarchy of rules (…more…).

An example of a SAGE-ARS interface (4)

Giulio's (who?) request to remove Marco's (who?) profile was denied, because of the Italian legislative decree 101/2018 (what?).

This decision was taken by an automated process called SPINdle (what?) starting from a set of known facts (...less...).

- *The decision taken by SPINdle is composed by a set of logical conclusions (...more...), the premises on which the rules have been applied and a hierarchy of rules.*
- *The set of premises is (...less...) :*
  - ***X** is called Marco (...source...)*
  - ***X** is 14 years old (...source...)(what if?)*
  - ***X** is Italian (...source...)(what if?)*
  - ***X** is resident in Italy (...source...) (what if?)*
- *The hierarchy of rules used to get the conclusions is (...less...) :*
  - ***R1**: "if X is adult (what?), then X obtains consent (what?)" (... source ...)*
  - ***R2**: "if X age is less than 14, then X does not obtain consent (what?)" (... source ...)*
  - ***R3**: "if X age is less than 14 and X lives in Italy (where?), then X obtains consent (what?)" (... source ...)*
  - ***R4**: "if X does not obtain consent (what?), then X's profile is removed (how?)" (... source ...)*
  - *"**R2** rebuts (how?) **R1**" (...source...)*
  - *"**R3** rebuts ("Lex specialis derogat generali" ...more... is applied ...hide...) **R2**" (... source ...)*

(what if?)

SPINdle (what?) reasoned over a LegalRuleML representation (what?) of the GDPR (what?) which is a hierarchy of rules (...more...).

Marco's
ed
'2018.

A different example

Explaining Decisions on Loan Application

Welcome *Mary*

What is a FICO score?

Question: What is a FICO score?

Answer:

- Whether you have a credit card or a charge card, the most important factor in building or improving your FICO score is using credit responsibly. That means paying your bills on time and using your credit only when needed. If you can do those things consistently, you should be well on your way toward maintaining a good score. [More..]
- For other types of credit, such as personal loans, student loans and retail credit, you'll likely want to know your FICO Score 8, which is the score most widely used by lenders. [More..]
- Learn more about the history of FICO Scores. [More..]
- FICO Scores is: Credit bureau risk scores produced from models developed by Fair Isaac Corporation are commonly known as FICO Scores. FICO Scores are used by lenders and others to assess the credit risk of prospective borrowers or existing customers, in order to help make credit and marketing decisions. These scores are derived solely from the information available on credit bureau reports. [Less..]

Pertinence	Source	Document
68.87%	FICO Scores is: Credit bureau risk scores produced from models developed by Fair Isaac Corporation are commonly known as FICO Scores. FICO Scores are used by lenders and others to assess the credit risk of prospective borrowers or existing customers, in order to help make credit and marketing decisions. These scores are derived solely from the information available on credit bureau reports.	MyFICO - glossary

Conclusions

Conclusions

SAGE-ARS in explanations lets the users explore a possibly overwhelming explanatory space.

This provides a reasonable support for goal-orientedness by letting users choose their own paths, pace, focus, language.

The interface is not a passive question/answering machine, but actively supports and suggests explorations, sourcing, groundings, and expansions.

Although the explanatory space is still the same huge and overwhelming Exhaustive Explanatory Closure with answers to all possible questions, it provides a manageable, navigable, explorable overview of the information that fosters effectiveness, efficacy, and satisfaction.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Credits:

Fabio Vitali[1], Francesco Sovrano[2]

[1] Department of Computer Science, University of Bologna, Bologna (I)

[2] Faculty of Informatics, Università della Svizzera Italiana, Lugano (CH)

fabio.vitali@unibo.it

francesco.sovrano@usi.ch