Politecnico di Milano – School of Design

Digital Interaction Design

A.Y. 2021/2022

Understandable AI and Explanation Usability

a fine-grained analysis of the state of art of Explainable AI visualisation from AI novices perspective

Francesca Macolino - 969390

Relatore: Marco Ajovalasit

Sommario

Abstract:

In the last few years, we have assisted to and extended and shared effort in the scientific community: build AI-powered systems embracing the Human Centred AI discipline which aim is on amplifying, augmenting, and enhancing human performance in ways that make AI systems reliable, safe and trustworthy.
In this direction, Explainable AI, the branch of artificial intelligence that encompasses the methods and processes that enable users to understand and trust the results and output created by machine learning algorithms, is one of the most promising fields.

Even if literature about explainable AI techniques is growing and growing every day, since the necessity to cope with the explainability-performance trade off of the most effective deep learning algorithm or the new regulations that made mandatory providing 'right for explanations', only in the last few years the XAI research community has embraced a more broader and multidisciplinary view on the topic shifting from serving only data scientist and domain experts and adopting an end user-centred approach. From this efforts literature agreed on the necessity to improve explanation effectiveness in terms of understandability and usability. AI novices, user with without any or little previous experience in the AI field are the user group most disregarded by them that's why they has been selected as the main target of this work.

This thesis covered the topic of Understandable AI and Explanation Usability from AI novices' perspective and contributed to the literature on designing explainable AI user experiences in the context of explainable interfaces providing actionable insights for practitioners to design explanatory narratives serving user needs, expressed, according to the question driven design approach for explainable AI, as question users may have in mind while they are seeking for explanations.

The methodology has exploited a participatory design approach which involved 10 semi structured interviews with AI novices to answer to the following research questions: What is the AI novice reasoning at the first interaction with explanation types? What information are easily caught, what mental model they inform, what is their perceived usefulness and their intention of use? and Explanatory forms/explanation type can convey the information needed to AI novices to answers the prototypical questions given by the question driven design approach?

The discussion has covered the first research question providing insights in terms of user friendliness and perceived usefulness and highlighting user's opinion in terms of applicable context of use and possible strategies to improve their visualisation in order

to overcome possible misleading factors. Additionally, from this analysis we have been able recognize patterns of user-explanation interaction for AI novices.

For what concern the second research question, from the first user study focused on the capability of explanation type to serve user explanatory needs (namely, answer the 10 prototypical questions) resulted the a fine-grained explanatory forms analysis, an actionable resource for XUI designers which list, for each of them, what question are able to answer differentiating the one given in an efficient manner i.e., providing complete or partial answer and if they are given with effectiveness aka if the information grasped to answer them are directly or indirectly given, thus, more or less cognitive demanding.

Additionally, we have reversed the explanatory forms' results providing an additional fine-grained question analysis which provide, for each of the 10 prototypical questions what the most suitable explanatory form are to answer them in order to fulfil user needs (i.e. providing complete answer). For each question the elements extracted from the form to answer them are summarised differentiating what ones can provide complete or partial hints to serve the whole XAI community in the development of new XAI techniques able to provide the most effective explainable experiences even for AI novices, in the near future.

1. Explainable AI

1.1 Introduction

In the last few years, we have assisted to and extended and shared effort in the scientific community: build AI-powered Systems embracing the Human Centred AI discipline which aim is on amplifying, augmenting, and enhancing human performance in ways that make AI systems reliable, safe and trustworthy.

In this direction, Explainable AI is one of the most promising field which has grown and grown in the last years pushed from the capillarity of the development of AI powered solutions as well as the need to accomplish the new regulations spread up to control and ensure the so called Trustworthy AI, such as the AI Act from the EU commission in April 2021 which states the necessity of the 'right of explanation' especially for AI solutions involved in high stake scenarios.

But let's take a step back, what is Explainable AI?
The terms explainable AI refers to the branch of artificial intelligence that encompasses the methods and processes that enable users to understand and trust the results and output created by machine learning algorithms.

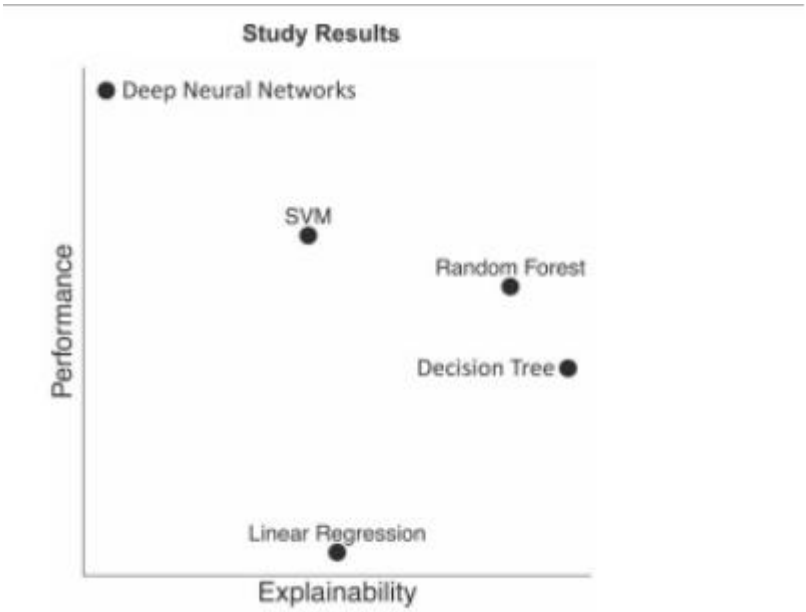
And why we need explainable AI?
While the origins of AI can be traced back decades, only in the last years as grown a broad and proven agreement on the capability for such machine intelligence AI powered systems to improve many aspects of human life, especially in the assisted decision-making field. It has been demonstrated that the capabilities of AI-based applications achieve a high degree of performance in complicated activities compared to humans, establishing them as a critical component for future industrial development (West, 2018).

According to The User Centric AI approach we intend such systems as able to enhance and increase users performance under the so-called AI augmentation, acting as companion in the decision making process thanks to the shared effort between designers and developers to build effective human-AI collaborations, in comparison to the opposite concept, the AI automation, which refers to the AI application in which the human is completely substituted by the machine.

It’s the purpose to ensure this effective Human-AI collaboration the key reason behind the need to make such systems explainable and transparent (Goodman and Flaxman, 2017) and a way to achieve that is providing explanations exploiting explainable AI techniques thus developing explainable AI systems.

1.2 Explainable AI techniques

The techniques of explainable AI, which are the domain of data scientists and AI engineers, are mainly classified based on the approach used. At present, there are mainly two types of solutions for explainable AI techniques: ante-hoc, involving models explainable by design and thus model-based, or post-hoc, used afterwards in a new round of training usually using reverse engineering methods on machine learning models to explore how a specific output can be generated given a single input enlighting the process the machine followed to took decision and thus model-agnostic. There’s a proven threshold between Performance and explainability of models that may be used to build AI based decision support system: traditional algorithm (such as linear regression, decision trees, and Bayesian networks), belonging to the category explainable by design, nowadays cannot compete with the performance accuracy reached by the modern neural networks: to overcome their ‘black box’ nature in the last years we have assisted to the development of more and more post hoc methods, in order to overcome the lack of explainability of this algorithm and guarantee their performativity in terms of output accuracy.



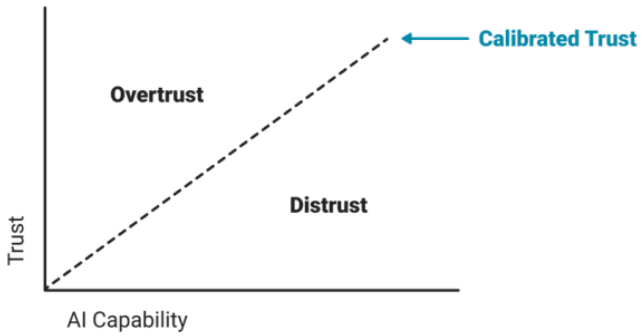
Performance explainability trade-off

Ensure models performativity is only the iceberg peak of why we need explainable AI: AI infused product are increasingly being use in sensitive areas and high-stakes scenarios (e.g. autonomous cars or military drones) and to ensure an effective human—AI collaboration in order to supervise it with the intended purpose we need to know how they make decisions generally [providing global explanations] and in a particular event [providing local explanation]. Additionally, in these scenarios embracing sensitive areas, as anticipated, current laws and regulations already require a precise level of explainability. As a matter of facts, the field of legal argumentation and reasoning deals with the boundaries of explainability: as an example, The General Data Protection Regulation (GDPR) in the European Union already defines a right to explanation, and it necessitates reasonable explanation on the logic of the AI systems when a citizen is subject to automated decision making. Finally, explainable AI became necessary to detect bias especially for systems trained on historical data, in order to not reinforce the discrimination of the past.

1.3 Explainable AI and calibrated trust

Based on those promises it's already clear that designing explainable systems brings numerous benefits: not only in terms of improving the models on which they are based with regard to fairness, transparency and accountability, but adopting explainable AI techniques also plays a fundamental role for the users who are going to use them.

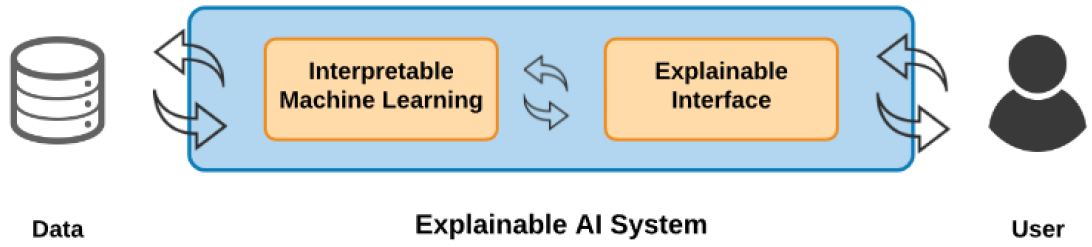
In fact, it has been amply demonstrated in the literature that explaining the functioning of AI based decision support systems through explanations that make the process behind the generation of the output understandable, helps in building of the so-called calibrated trust, which makes it possible to avoid the problems of algorithmic bias, which occur when users put too much trust in the system and not rationally questions the output, or of algorithm aversion, when the support and benefits that the system could bring are rejected because there is no trust in them.



From XAI, calibrated trust definition

1.4 Explainable systems

The DARPA XAI program illustrates the XAI process as a two-staged approach. It distinguishes between the explainable model and the explanation user interface (Gunning & Aha, 2019 cited by Chromik & Butz, 2021, pp. 1-3). Explainable AI systems consist not only of the explainable AI model but also of an explainable interface, where the user-system interaction takes place. Explainable interfaces design therefore plays a fundamental role in the field of AI augmentation, maintaining the so-called human-in-the loop, and, in decision support systems, allows the user to have the tools to understand how the system works and, by calibrating trust in it, be able to evaluate the suggestion or prediction it gives, before making the final decision.



From Mohseni et al.: Explainable AI systems

This thesis is focused on how to design explainable AI experiences within the context of explainable interfaces design to enhance an effective human AI collaboration in decision making support systems. The aim of this work is to evaluate the current state of art of explainability methods and explainable interface design.

2.XAI from HCI perspective

2.1 The shift between data science domain and HCI domain:

Broadening the user groups which can benefit from the development of Explainable AI systems moved the explainable AI field from the data science domain to the end user. Answering to recent calls to study AI systems from end users and XAI from the HCI perspective (Brennen, 2020; Weitz et al., 2019b). researchers in the field of XAI have increased their attention to the fact that the development of explanations from an AI-system requires a tailor-made approach considering the context of use in which the interaction takes place as well as a concrete analysis of the target groups to which the explanations are delivered and their needs.

Since then, the focus within XAI development has started to shift from a mainly technical approach to a more integrated sociotechnical one in which human-centered design (HCD) is paramount (e.g., Lim, Yang, Abdul, Wang, 2019; Mittelstadt, Russell, Wachter, 2019; Neerincx et al., 2019; Madumal, Miller, Sonenberg, Vetere; see Arrieta et al., 2020 for an overview of recent papers on HCD for XAI).

2.1.1 Terminology clarification

Before going deeper on the topic is worth to be mentioned that this shift of perspective has boosted the spreading of terminology referred to the XAI discipline, and nowadays there still a lack of a broader agreement on specific terms.

For the aim of this thesis we adopt the terminology as described by Arrieta et al. (2020):

•Explanation**:** an interface between human and system that accurately approximates the model of the system and is comprehensible to the human (Guidotti et al., 2018b).

•Explainability: the ability to deliver explanations. The model that is used by the system needs to be interpretable to be able to provide an explanation (Guidotti et al., 2018b).

•Causability: the ability the enable a user to achieve causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use (Holzinger et al., 2019).

•Interpretability: the ability to provide meaning to a human in understandable terms (Guidotti et al., 2018b).

•Transparency: a model is transparent if it is understandable by itself (Adadi and Berrada, 2018).

•Comprehensibility: the ability of a model to represent its knowledge in an understandable fashion (Adadi and Berrada, 2018).

•Understandability: the ability to make a human understand the model's function without the need to explain its internal structure or the algorithms that are used (Montavon et al., 2018).

2.2 XAI stakeholders

As introduced, the shift of the domain of the XAI research has been a direct consequence of the increasing awareness about the fact that user groups which can benefit from explainable AI application has grown and grown. Mohseni et al recognised, through an extensive literature review based on Meske et al., 2022 and Arrieta et al., 2020, five major stakeholder groups for XAI and a rationale for each one.

XAI stakeholder group	Explanation
Users affected by model decisions	As AI systems are widely implemented across services, people are constantly directly and indirectly affected by decisions made by various models in various contexts
Individuals using AI systems	An increasing number of tools and services include AI components. Examples are numerous, from online recommendation systems to anomaly detection solutions trying to block spam email
Managers and executive board members	In business firms, upper executive management has oversight into the AI systems used in their company
Regulatory entities	Various regulatory entities such as the European Union and individual governmental bodies are interested in controlling and legislating AI systems to protect citizens from potential harm that immature AI systems could cause
Developers such as data scientists and system engineers	Perhaps the most obvious target audience for XAI are the developers who create the AI models. They are responsible for ensuring that the models work effectively and in a desired fashion

From Mohseni et al: XAI stakeholders groups

According to literature there's a distinction between people who freely utilize AI systems and people who are impacted by the decisions made by AI systems when it comes to end users (Meske et al., 2022).

Also, two stakeholder organizations are monitoring AI systems from various angles: while managers and executive board members make sure AI systems fulfil their goal in the larger company environment, regulatory bodies ensure that they do it to complying regulations.

Finally, AI system developers are regarded as a separate stakeholder group (Arrieta et al., 2020; Meske et al., 2022) which mainly diagnose, supervise and detect system errors to improve systems**.**

According to any user centred design approach is critical to keep the audience in mind while putting transparent AI and XAI into reality (Parsa et al., 2020; Ribeiro et al., 2016). For instance, data scientists or AI auditors typically have more technical skills and knowledge of ML systems than the common public (Dodge et al., 2019; van der Waa et al., 2021; Weitz et al., 2019a). The objectives and motivations behind XAI design must therefore be made clear to the systems users, as well as the way of communicating the explanations themselves before starting the development phase.

Non-technical end-users, or end-users for short, include both laypeople and domain experts. For example, doctors employing AI-assisted technology in diagnostic tasks (Caruana et al., 2015; Holzinger et al., 2017; Jin et al., 2020 cited by Laato et al., 2022, pp. 8-10), judges using AI to support reaching a guilt decision (Kleinberg et al., 2017) and bankers using AI to help with loan application approval.

According to their latest literature reviews laypeople could be pinpointed as a key end user group in a multitude of studies but as they claim it’s clear that, in the current development of XAI research, the non-technical end-users are mostly disregarded, as a matter of facts not a single publication has mentioned “user" in the title. A few portions focus on the promotion of user-centric variables, such as adoption (Kwon, 2018; Meacham, 2019; Ming, 2018; Vellido, 2019), which, are all from the health context; fairness from the human-resources context and acceptance from meta explainability Datta, 2016; Goebel, 2018). Other variables are very related to technology and for experts, such as debugging and error resilience (Dimitrova, 2019; Theodorou, 2017), verifiability (Yeganejou, 2019), the performance of communication networks (Santos, 2019), among others.

Because, as anticipated, the primary focus of the XAI development has been debugging, comprehending, and enhancing AI models for technical users including data scientists, AI researchers, and developers (Dodge et al., 2019; Santos and Abel, 2019), the development of XAI techniques never took into consideration their precise needs by design. However,

XAI is critical across all areas where ML is used (Goebel et al, 2018), thus is needed to properly understand the user in order to promote an adequate explanation (Ribera and Lapedriza, 2019).

That’s the reason behind this work go deepen in understanding and improve end users’ needs and goal to design Explainable AI.

2.2.1 Explainable AI for AI novices

According to Mohseni et al. literature review on user group’s goal in explainable AI, there are four primary design goals for XAI system's end users who are AI novices:

•Transparency in Algorithms
In contrast to an incomprehensible intelligent system, a XAI system's immediate purpose is to aid end users in understanding how the intelligent system functions.
By offering comprehensible transparency for the intelligent algorithms, machine learning explanations enhance users' mental models of the underlying algorithms (Weller, 2017). Furthermore, by helping users comprehend model output better (Lim, 2015), transparency of a XAI system can enhance user interactions with the system (Kulesza et al., 2015).

•User Trust and Reliance:
By offering explanations, XAI systems can increase end-users' trust in the intelligent algorithm. Users can calibrate their reliance on the system's outputs and rationally question the system's suggestions. This result in a bond between the user and the AI system through the XAI design.
Recommendation systems (Berkovsky et al., 2017), autonomous systems (Wiegand et al., 2019), and crucial decision-making systems (Bussone et al., 2015) are a few examples of applications where XAI attempts to increase user reliance through its transparent nature.

•Bias Mitigation:
A dark – side - effect of intelligent systems may be unfair and biased algorithmic decision-making.
This may result from feature learning and biased training data and may lead to discriminatory algorithmic decision-making (Mehrabi, 2019). But implementing AI

techniques thus providing explanations End-users become able to check and see if those bias occurs.

Criminal risk assessment (Binns et al., 2018; Lee et al., 2019), loan and insurance rate prediction (Chen et al., 2019), and fairness assessment are a few examples of context of use in which XAI is applied for bias mitigation.
•Privacy Awareness:
Offering to end users a way to evaluate their data privacy is another objective in the creation of XAI systems. End users can learn through machine learning explanations what user data is analysed during algorithmic decision-making.

Based on this, Mohseni et al. synthesized for the five key objectives the underlying goals for explaining AI systems for end users: the increasing of (1) understandability, (2) trustworthiness, (3) transparency, (4) controllability and (5) the fairness of the system. To guarantee those benefits other studies also discussed general goals not particularly related to the ML system, such as usability, ease of use and satisfaction (Oh et al, 2018)

To the aim of this thesis to enhance AI novices XAI interaction central focus is posed to understandability and explanation effectiveness as a matter of explanation usability, easiness of use and satisfaction.

2.3 State of art of XAI user experiences for end users

To ensure that XAI systems meet users' needs evolved, in HCI research for XAI, the field of evaluation of explanations, which answers whether an explanation is good enough, and how to compare different explanations.

The HCI community is demonstrating visible efforts around designing and studying user interactions with explainable AI (Binns et al., 2018; Cai et al., 2019; Cheng et al., 2019; Dodge et al. 2019; Hohman et al., 2019; Kolcielnik et al., 2019; Lai and Tan, 2018; Rader et al., 2018). These recent studies largely focused on empirically understanding the effect of explanation features on users' interaction with and perception of ML systems, usually through controlled lab or field experiments. Notably, although explanations were found to improve user understanding of the AI systems, The reality of practical AI applications in sensitive areas reveals the inability of those systems to communicate effectively with their users (Erickson et al., 2008) and conclusions about its benefits for user trust and acceptance were mixed (Cheng et al., 2019; Kolcielnik et al., 2019; Lai and Tan, 2018;

Poursabzi-Sangdeh et al., 2018), suggesting potential gaps between algorithmic explanations and end user needs. As a matter of facts, the current AI techniques, exactly because developed by engineers with the primarly scope to debug systems are not enough to deliver explanations directly understandable by the diversified user group constituted by lay users. (Yang et al., 2021)

According to Naiseh studies on Human-AI interaction on collaborative human-artificial intelligence decision-making tools (Naiseh et al., 2021), users are subjected to two systematic errors while interacting with explanations: skipping or misapplying. The reason behind the former is the inadequacy of the current XAI solution to engage users' curiosity or meet the explanatory need that lead their explanation-seeking as well as the tendency of explanations to be too much cognitive demanding; for the latter the problem lay on the unintended purpose for XAI developer or designers to ensure their explanation are plain enough to be caught from a non-expert audience.

In fact, the AI academic community has been active in exploring mathematical approaches that can increase the explainability of models (e.g., LIME (Ribeiro et al., 2016), Shapley value (Lundberg and Lee, 2017), counterfactual explanations (Kim and Austin, 2016). Anyway, if such efforts remain predominantly based on computer scientists' perceptions of what constitutes an explanation there will always be a gap between explainability techniques and what it means to explain to real users (Wang et al., 2019; Liao, Gruen and Miller, 2020).

2.3.1 Understandable AI

The gap between the state of art of explainable AI techniques and non-technical end users is even more evident if we deepen the differences between what is explainable AI and the new emerging field in the HCI research of Understandable XAI. An understandable explanation is an explanation which provide a human with information that is extracted from and/or based on its internal model and makes a human understand (part of) the functioning of the model, to understand a given output. This thin difference between Explainable AI and Understandable AI is the key concept that has moved the research within XAI from a model-centred approach to a user-centred one. If explainability deals with extracting explanations from a system's model, which may be not inherently human-understandable, with the goal of deliver understandable AI an explainable model and an explainable interface is required to create explanations

that can be understood by humans (Holzinger et al., 2019 cited by Schoonderwoerd et al., 2021). Thus, while XAI is concerned with developing methods to make machine models transparent and traceable, causability is about measuring the quality of such explanations to increase causal understanding of a user (Holzinger et al., 2020). In that direction the human-centered design methodology provides methods to determine exactly what information is understandable and useful to humans and thus should be used in designing explanations from the system.

As the human use of computing is the subject of inquiry in HCI (Oulasvirta and Hornbaek, 2016 cited by Chromik and Butz, 2021, p. 2), this thesis tackles the challenge proposed to our discipline to “take a leading role by providing explainable and comprehensible AI, and useful and usable AI” (Xu, 2019) to “provide effective design for explanation UIs” (Xu, 2019).

But what we should design when we want to design effective explanations delivered through an explanatory User Interface?

3. Design explainable systems

3.1 Explanations as interactions

With all the promises in the first two chapters is clear how nowadays design explainable systems able to deliver effective explanation is an interaction design matter: actually, Mueller et al. (Mueller, 2019 cited by cited by Chromik and Butz, 2021, p. 3) consider an effective explanation to be “an interaction” and “not a property of statements”; Adadi et al. (Adadi , 2018) state that “explainability can only happen through interaction between human and machine” and Abdul et al. (Abdul et al., 2018) present research on interactive explanation interfaces as an important trajectory to advance the XAI research field.

The design of interfaces that “allow users to better understand underlying computational processes” is still considered a grand challenge of HCI research (Shneiderman, 2016) but, according to Shneiderman, is directly with XUIs the strategy to pursue the human-centered AI approach, which, as introduced in the first chapter, aims “to amplify, augment and enhance human performance” instead of automating it Shneiderman, 2020.

To describe Human XAI interaction, Miller frames XAI as one kind of a human-agent interaction problem where an ”explanatory agent [is] revealing underlying causes to its or another agent’s decision making” (Miller, 2019). As such, it is about the interplay between a human user and an AI agent that is mediated through an XUI.

As Shneiderman, we consider explanation user interface (XUI) as the sum of outputs of an XAI process that the user can directly interact with. He outlines two modes of XUI. Explanatory XUIs aim to convey a single explanation (e.g., a visualization or a text explanation). In contrast, exploratory XUIs let users freely explore the ML model behaviour (Shneiderman, 2020). They are most effective when users have the power to change or influence the inputs. Arya et al. (Arya, V., et al., 2019) distinguish between static and interactive explanations. A static explanation “does not change in response to feedback from the consumer”. In contrast, interactive explanations allow “to drill down or ask for different types of explanations [...] until [...] satisfied”.

Even if, as Ribera et al. mentions in their work, there is no agreement on a specific definition for an explanation, from their review results that some relevant points are shared in almost every definition. For example, many definitions relate explanations with "why" questions or causality reasonings. Also, and more importantly, there is a key aspect when trying to define what an explanation is: there are two subjects involved in any explanation, the one who provides it (the system), or explainer, and the one who

receives it (the human), or explainee. Thus, when providing AI with explainability capacity, one cannot forget about to whom the explanation is targeted and what are their needs and goal while seeking for an explanation.

If the explanation come out from the general interplay between the XAI system and the user, the focus of designing XAI user experiences become be the interactive qualities of the XUI itself. Vilone et al. define interactivity as “the capacity of an explanation system to reason about previous utterances both to interpret and answer users’ follow-up questions” (Vilone and Longo, 2020 cited by Chromik and Butz, 2021, p. 15). Moore and Paris (Moore and Paris, 1991) proposed that a good explanation facility should, among others, fulfil the requirements of naturalness (explanations in natural language following a dialogue), responsiveness (allow follow-up questions), flexibility (make use of multiple explanation methods), and sensitivity (provided explanations should be informed by the user’s knowledge, goal, context, and previous interaction). [Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces]

From this first two statements became natural considering explanations as interactive dialogues, conversation between the explainer and the explainee.

3.2 Explanations as conversations

The increasing demand of explainable AI systems and the different background of stakeholders of machine learning systems has highlight the need to propose the creation of different user-cantered explainability solutions, simulating human conversations with interactive dialogues or visualizations that can be explored.

The literature about considering explanations as conversations has roots in the work of Miller and Wang, the former, by conducting a literature review in social science on how humans give and receive explanations, identified a list of human-friendly characteristics of explanation that are not given sufficient attention in the algorithmic work of XAI, including contrastiveness (to a counterfactual case), selectivity, social process, focusing on the abnormal, etc. The latter, proposed a conceptual framework to connect XAI techniques and cognitive patterns in human-decision making to guide the design of XAI systems.

The critique of Miller (Miller, 2019 cited by Ribera and Lapedriza, 2019, p. 5) on current proposed explanations as being too static because not able to support "an interaction

between the explainer and explainee" is the milestone of the parallelism on explanations as conversations and lead us to the main desiderata that must be coped to deliver effective explanations.

An effective explanation should follow the cooperative principles of Grice [9] (Grice, 1975) and its four maxims:

- Quality: Make sure that the information is of high quality: (a) do not say things that you believe to be false; and (b) do not say things for which you do not have sufficient evidence;

- Quantity: Provide the right quantity of information. (a) make your contribution as informative as is required; and (b) do not make it more informative than is required;

-Relation: Only provide information that is related to the conversation. (a) Be relevant. This maxim can be interpreted as a strategy for achieving the maxim of quantity;

- Manner: Relating to how one provides information, rather than what is provided. This consists of the 'supermaxim' of 'Be perspicuous', and according to Grice, is broken into various maxims such as: "(a) avoid obscurity of expression; (b) avoid ambiguity; (c) be brief (avoid unnecessary prolixity); and (d) be orderly".

Notably, the first threee statements refer to the content of the explanation, while forth refers to the type of explanation.

With this premises is clear that designing Explainable AI system targeting end users, aimed to the deliver understandable AI can benefit from the theoretical frameworks developed for human communication-

We postulate that to make explanation effective we must ensure the explanations ability to answer questions users may have in mind while they are in search of explanation with great quality and quantity, thus with enough details to fulfil their needs without overwhelm them.

We believe that this may overcome the skipping user error found in the previous literature analysis.

To achieve the relation maxim is needed to precisely target the explanations provided according to the question and investigate what elements do not serve the forth maxim: thus, overcoming the misleading factors that may lead users to misapply explanations.

The social nature of explanation also maps to an essential requirement for interactivity in XAI applications (Krause, 2016 cited by Liao and Varshney, 2021, p. 14). User interactions do not end at receiving an XAI output but continue until an actionable understanding is achieved. In other words, as users' explainability needs are expressed in questions, they will keep asking follow-up questions until satisfied, thus engaging in back-and-forth conversations.

Miller reviewed several relevant theories including Hilton's conversational model of explanations (Hilton, 1990), which postulates that a good explanation must be relevant to the focus of a question and present a topology of different causal questions. Antaki and Leudar (Antaki and Leudar, 1992) extended this model to a wider class of argumentative dialogue for the common pattern of claim-backing in explanations. Walton (Walton, 2004) further extended this line of work into a formal dialogue model of explanation, including a set of speech act rules. From their work has born the research path on a question driven design approach for explainable AI user experiences.

3.3 Related work on XAI for end users and XAI user experience design

The literature is clear: explanation are conversations, XAI development should include social and behavioural science point of view and support human-AI dialogues taking place through Explanatory interfaces. The dialogue has to be interactive allowing the user to ask follow-up question until his explanation need is satisfied. Overall, the design of this explanatory experiences may serve different categories of users, all the XAI stakeholders thus need to be developed with a user-centred approach.

As a matter of facts, the theoretical approach presented since now has been deepened and criticized. The main point against it, given the underlying benefits to consider explanations as dialogues or conversations has been mark it as too abstract and not able to provide practical guidance on how to design explanatory interfaces for explainable user experiences.

Wang et al. conducted a review on explanation theory literature, and further provided a theory-driven, user-centered XAI framework that describes how the human reasoning process and explanation theories guide explanation system requirements (Wang, 2019

cited by Jin et al., 2021, p. 6). They suggested the XAI system should support reasoning while mitigating heuristics and bias. Their work is a first attempt in developing user-centered XAI design guidance, but again remained at a conceptual and abstract level, and lacked actionable guidance on how to practically implement explanation theories for context-specific tasks and needs. In their follow-up paper, Lim et al. (Lim et al., 2019), extended the framework by detailing the explanation types (input, output, certainty, why, why not, what if, how to, and when), and by proposing pathways to link these types to users' three explanation goals: filter causes, generalize and learn, and predict and control.

This explanation type taxonomy was first identified by Lim and Dey in 2009 (Lim et al., 2009), by surveying users' questions in crowdsourcing user studies for context-aware systems. Based on their work, Jin at al. (Jin et al., 2021) has defined 12 end user friendly explanatory forms which compose EUCA, End User Centred Explainable AI prototyping framework. They explored the XAI solution space by extracting the resulting explanation information from existing technical literature in AI, HCI, and information visualization fields via literature review, then they selected and summarized end-user-friendly explanatory forms based on the following criteria:

- The explanatory forms must be end-user-friendly, i.e., users are not required to have background knowledge in AI or machine learning techniques to understand the explanation.

- The explanatory forms must be generally mutually exclusive regarding their underlying information in order to compose a library of items, building blocks that represent the elemental explanation information, and their combination would not be redundant/repeated in an XAI system.

EUCA work included a variety of explanation goals, exploited in the user study they conducted with 32 interviewees, as the trigger point or motivation to check the explanation of an AI system. They based their findings on the correlation between explanatory forms and users' needs with quantitative and qualitative user study data, that may vary in different contexts or usage scenarios. Their results provide fine-grained details of end-users' requirements for different explanation goals.

Their work is a great contribution in providing practitioners with actionable insights about how and why choose an explanatory form instead of one another providing a table which analysed user friendliness level, pros, cons of using them from an end user perspective, UX/UI recommendations, applicable explanatory goal and algorithm

examples to implement each of them. The granularity level of their analysis is notably, as well their attention on following a participatory approach with end users, but the user friendliness level associated to each explanatory form is vague and the approach to consider explanatory forms as building block is not suitable to optimize the explanatory space and develop a proper narrative around the explanation experience reducing cognitive effort for users. Additionally, their explanatory goals, even if summarized from prior works miss the link between user needs and question to answer in a dialogue with the AI system proved to be effective to guide the design process for XAI user experiences.

On the contrary Liao et al. (Liao et al., 2020), further explored the idea of providing mapping guidance between users' requirements and explanation types to facilitate human-centered explanation design developing a question driven framework. Thought their first effort, based on prior HCI work using prototypical questions to represent "intelligibility types" (Lim and Dey, 2009), and social science literature showing that people's explanatory goals can be expressed in different kinds of questions (Hilton, 1990) they proposed to identify users' explainability needs by eliciting user questions to understand the AI (Liao, Gruen and Miller, 2020).

By interviewing 20 designers, they collected common questions users ask across 16 ML applications and developed an XAI Question Bank, with more than 50 detailed user questions organized in 9 categories:

• How (global model-wide): asking about the general logic or process the AI follows to have a global view.

• Why (a given prediction): asking about the reason behind a specific prediction.

• Why Not (a different prediction): asking why the prediction is different from an expected or desired outcome.

• How to be That (a different prediction) : asking about ways to change the instance to get a different prediction.

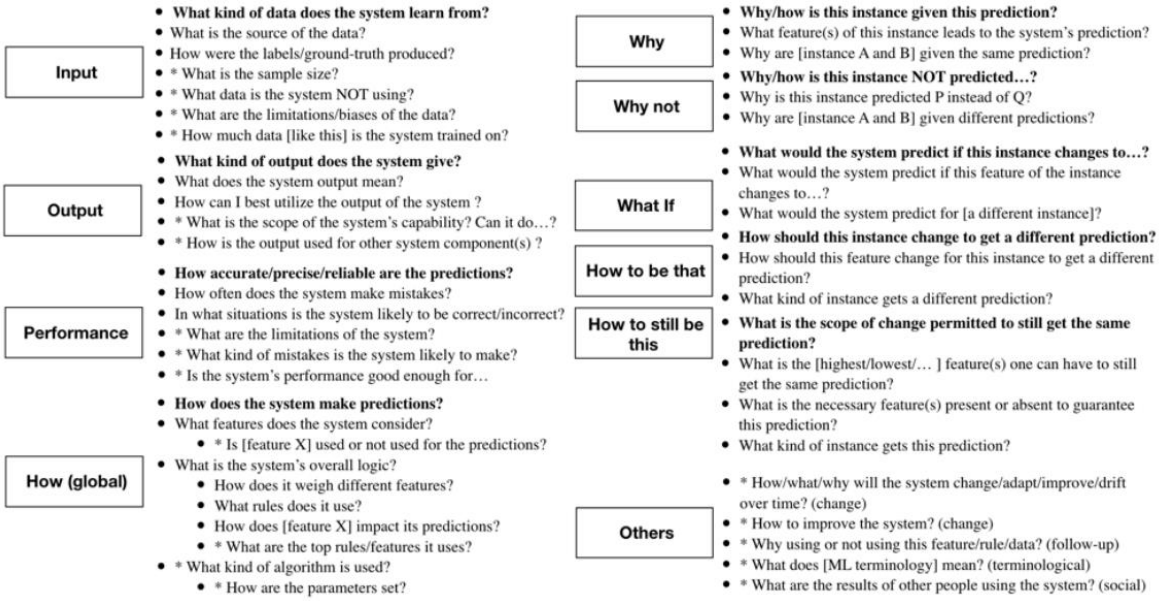
• How to Still Be This (the current prediction): asking what change is allowed for the instance to still get the same prediction.

• What if: asking how the prediction changes if the input changes.

• Performance: asking about the performance of the AI.

• Data: asking about the training data.

• Output: asking what can be expected or done with the AI's output.



The XAI question bank from Liao et Al.

In a follow-up work (Liao et al., 2021), they propose the complete question-driven user centred design method that starts with identifying key user questions by user research, then uses these questions to guide the choices of XAI techniques and iterative design. To facilitate this process and foreground users' explainability needs, they suggested to reframe the technical space of XAI by the user question that each XAI technique can address. For example, a feature-importance explanation technique can answer the 'Why question', while a counterfactual explanation can answer the 'How to be that question'.

Category of Methods	Explanation Method	Definition	Algorithm Examples	Question Type
Explain the model (**Global**)	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	**How**
	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	**How**, Why, Why not, What if
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	**How**, Why, Why not, What if
Explain a prediction (**Local**)	Local feature importance and saliency method	Show how features of the instance contribute to the model's prediction (including causes in parts of an image or text)	[61, 74, 83, 85, 101]	**Why**
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	**Why**, **How to still be this**
Inspect counterfactual	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature (often in a visualization format)	[8, 33, 36, 51]	**What if**, How to be that, How to still be this
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	**Why**, **Why not**, **How to be that**
Example based	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	**Why**, How to still be this
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	**Why**, **Why not**, How to be that

Table from question driven design approach: correlation between method and questions.

Additionally. they provide a suggested mapping between the question categories and example XAI techniques, focusing on techniques that are available in current open-source XAI toolkits accessible for practitioners (H2O.ai Machine Learning Interpretability, 2017; Model Interpretation with Skater, 2018; IBM AIX 360, 2019; Microsoft InterpretML, 2019).

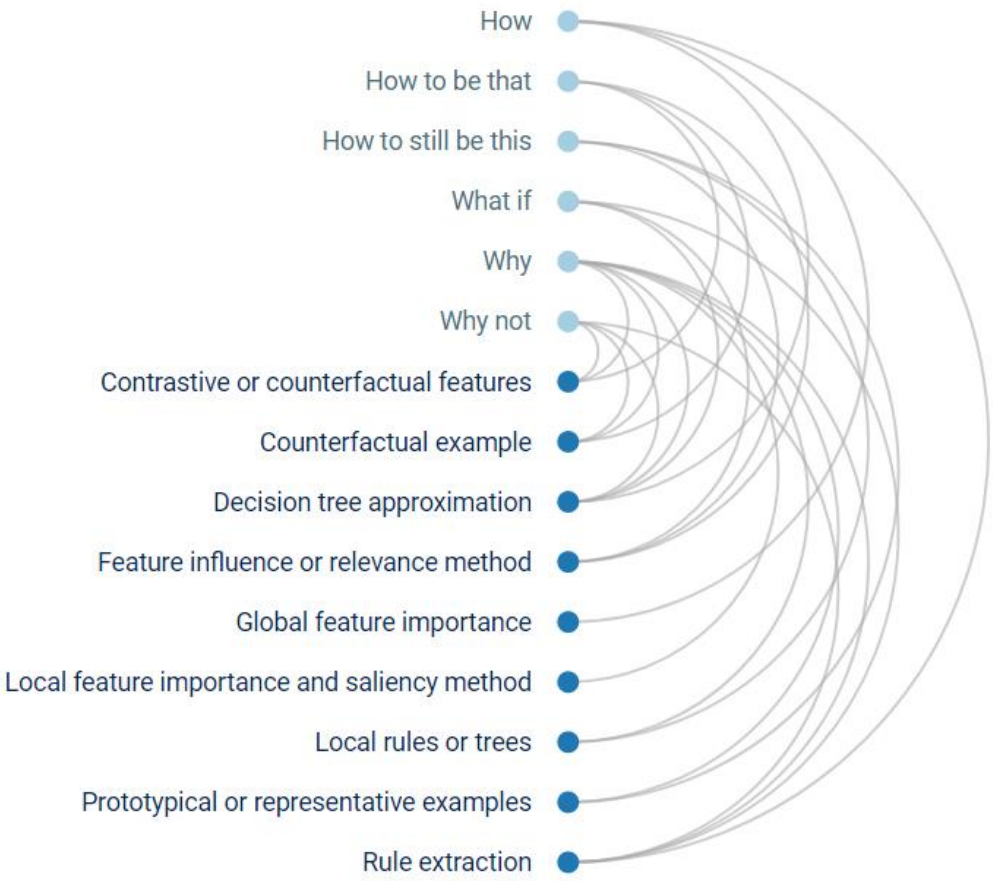
Question	Ways to explain	Example XAI methods
How (global model-wide)	• Describe the general model logic as feature impact*, rules† or decision-trees‡ • If user is only interested in a high-level view, describe what are the top features or rules considered	ProfWeight*†‡ [28], Global feature importance* [71, 105], Global feature inspection plots* (e.g. PDP [49]), Tree surrogates‡ [25]
Why (a given prediction)	• Describe how features of the instance, or what key features, determine the model's prediction of it* • Or describe rules that the instance fits to guarantee the prediction† • Or show similar examples with the same predicted outcome to justify the model's prediction‡	LIME* [89], SHAP* [72], LOCO* [63], Anchors† [90], ProtoDash‡ [47]
Why Not (a different prediction)	• Describe what features of the instance determine the current prediction and/or with what changes the instance would get the alternative prediction* • Or show prototypical examples that have the alternative outcome†	CEM* [27], Counterfactuals* [69], ProtoDash† (on alternative prediction) [47]
How to Be That (a different prediction)	• Highlight feature(s) that if changed (increased, decreased, absent, or present) could alter the prediction to the alternative outcome, with minimum effort required* • Or show examples with minimum differences but had the alternative outcome†	CEM* [27], Counterfactuals* [69], Counterfactual instances† [100], DiCE† [78]
How to Still Be This (the current prediction)	• Describe features/feature ranges* or rules† that could guarantee the same prediction • Or show examples that are different from the instance but still had the same outcome	CEM* [27], Anchors† [90]
What if	• Show how the prediction changes corresponding to the inquired change of input	PDP [49], ALE [10], ICE [44]
Performance	• Provide performance information of the model • Provide uncertainty information for each prediction* • Describe potential strengths and limitations of the model	Precision, Recall, Accuracy, F1, AUC; Communicate uncertainty of each prediction* [42]; See examples in FactSheets [11] and Model Cards [77]
Data	• Provide comprehensive information about the training data, such as the source, provenance, type, size, coverage of population, potential biases, etc.	See examples in FactSheets [11] and Datasheets [39]
Output	• Describe the scope of output or system functions. • If applicable, suggest how the output should be used for downstream tasks or user workflow	See examples in FactSheets [11] and Model Cards [77]

Table from question driven design approach: correlation between questions and XAI methods

Their results revealed rich details on users' needs for XAI but failed to show evidence (such as user studies) that the corresponding XAI methods will answer users' questions. Their framework directly guides the choice of explanation types based on user questions

(explanation needs) but, since the lack of user studies their correlation between explanation type and questions is based on assumptions.

Additionally, their results link multiple question to the same explanation type and viceversa.



Visualisation from UXAI: correlation between questions and methods

As user goal and needs may be conflicting with one another, designers of XUI “need to make trade-offs while choosing or designing the form of interface” (Tsai et al, 2019) but, as far as we know, there’s a lack of knowledge about based on what designers, exploiting the question driven framework and recognising user questions as user needs, should decide which explanation type use in build the explanatory interface.

It’s worth to be mentioned an additional work with valuable resources for practitioner who works on defining end user centred XAI experiences: UXAI. UXAI is an online resource seeking to surface critically informative, granular information otherwise buried in academic papers, in an approachable way that is more in line with current industry guidelines from IBM, Google and Microsoft. The website is divided into sections, working from a broad overview of AI to a tangible brainstorming tool. This last resource is clearly

based on Liao et al question driven design approach and their taxonomy question-explanation type. Deepening their work an additional question appear from the 9 category identified in the Liao et al. question bank: the How (under what conditions) question, associated to local explanation types. Surprisingly, while comparing the associations between question and explanation type, new matches results [appendix]. Even for their work is not possible to find any justification from a user study to validate question-explanation type pairings making even more prominent the necessity to conduct a study validate them with end user themselves.

From UXAI website: example of two toolkit brainstorming cards

4. Research gap and research questions

4.1 Understandable AI

From the literature review described in the first 3 chapters is clear how design explainable AI user experiences built around user needs is nowadays central in the HCI XAI research. As deepened in the third chapter, by and large, literature agrees that explanations in XAI systems are answers to a question, usually about the outcome of a computation. In the root of this research field, the question was expected to be focusing on the individual computation performed by the system (a local question) and to the causes of such outcome, so it could be phrased as a “why” or “how” question, and specifically a “Why did I obtain this result (as opposed to some other ones)?”.

As previously mentioned the state of art of explainable AI techniques often fails to deliver understandable and usable explanation for end users, able to overcome skipping and misapplying errors and, from the analysis of related work on XAI for end users, there’s a lack of evidences coming from user studies that the explanations that can be generated by the state of art of XAI techniques are directly linked to the prototypical questions identified from the literature on the intelligibility types exploited in the question driven design approach and UXAI toolkit.

Another missing element resulting from the literature is a fine grained analysis of Understandability of the current state of explainable AI technique that may ensure the intelligibility of the explanation we deliver to end users especially to an AI novice audience.

Because of that, the aim of this thesis is to investigate and validate the proposed association between explanation type and question they may answer and a fine-grained user study on users perceptions in terms of understandability while interacting with the state of art of explainable AI techniques.

For what concern the user group chosen to validate our hypothesis, since the lack of literature investigating non-technical-end user needs deepened in the chapter 2, the sample for the experiment has been composed of layusers without or with little previous experience in the XAI field or AI based decision making support systems, thus AI novices, postulating that, if the experiment will meet their mental models it will be possible generalize the insights in terms of understandability and even for a more expert audience.

RQ1: What is the AI novice reasoning at the first interaction with explanation types? What information are easily caught, what mental model they inform, what is their perceived usefulness and their intention of use?

4.2 Usable XAI

Thanks to the analysis of user perception on each explanatory form will be possible to understand what the misleading elements are and provide hints in terms of how improve the explanations visualisation as well as recognize patterns of user-explanation interaction for AI novices.

RQ2: Explanatory forms/explanation type can convey the information needed to AI novices to answers the prototypical questions given by the question driven design approach?

We assume that this activity may confirm Liao et al. results and thus more than one explanatory form will result able to answer the same questions. In the process of designing explanatory interfaces following the question driven design approach, to not cognitive overload users and avoid the skipping error, is clear that designers have to took decisions to provide an explanatory form instead of one another, but as highlighted in the section 3 there's no clear reference in literature on how of why base this choice.

Tackling the challenge to inform designers to develop effective explanatory interfaces we want to investigate explanatory forms from a usability perspective based on the definition of usability of explanatory narrative from Sovrano (Sovrano et al.2020):

'We consider an explanatory narrative as a sequence of information (explanans) to increase understanding over explainable data and processes (explanandum), for the efficiency, effectiveness, and satisfaction of a specified explainee that interacts with the explanandum having specified goals in a specified context of use.'

We built our experiment on their definition intending as the sequence of information the information conveyed by each explanatory form able to answer the question seeked by the user i.e. the question to answer is the specified goal of the interaction. To evaluate the interaction usability we rely the definition of usability as the combination of effectiveness, efficiency, and satisfaction, as per ISO9241-210, 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" (Norris et al., 2005).

For what concern the evaluation of efficiency to explain ('accuracy and completeness with which users achieve specified goals') we will investigate the quality of information, i.e. if the explanatory form provide enough element to fulfil the need to provide a

complete answer to the question or only some partial hints. On the other hand the effectiveness (“resources used in relation to the results achieved, usually including time, human effort, costs and materials.”) of providing an answer will be evaluated according to the ability, of the explanatory form to provide an answer in a direct or indirect way thus if requiring additional cognitive effort compared to an answer given at a first sight. The satisfaction will be evaluated on the overall interaction in terms of explanatory form easiness to be interpreted, so easiness to extract the information needed to answer the question and how this question is given (completely or partially, directly or indirectly).

For the reasons mentioned before the specific explainee is a lay users with no previous experience with XAI methods and techniques, thus, an AI novice. The context of use will be set thanks to the exploitation of a scenario-based experiment in which AI powered system can be used to boost decision making support tasks.

The results will be analised with the purpose to cope with the following hypothesis:

H1: different explanations addressing the same question has different level of effectiveness.

H2: different explanations addressing the same question has different level of efficacy.

H3: different explanations addressing the same question has different level of satisfaction.

The experiment will lead us to understand what explanatory form are more suitable to be chosen to answer a precise question and why.

Additionally, all the insights derived by the explanatory forms analysis will be exploited to understand the state of art explainable AI techniques to meet the user’s question-seeking needs and drive the focus of XAI community to develop new user-needs-informed explanations.

5. Metodology

In this study, we aimed to investigate the level of Understandability of the state of art explainable AI techniques and methods from an AI novice perspective as well as to validate the question driven design approach for explaninable AI and provide practitioners with actionable insights on which base their decisions to design effective explainable AI experiences through XUI.

The thesis experiment followed a participatory design approach methodology involving a round of ten semistructured interviews composed by open and closed questions to guide the conversation and generate quantitative and qualitative insights, a scenario-based task aimed to introduce a possible context of use of an AI based decision support system and the exploitation of the think aloud technique aimed to investigate participants mental model built around the scenario settings.

The 10 online semi-structured interviews, lasted 1.53 minutes in average and were structured as follow: the first phase has been aimed to introduce the research space and experiment goal and investigate participants information about their age, gender, educational background, familiarity with AI and related concepts, as well as their attitudes towards the use of AI in decision-making. Constraint for the participants selection has been select only the ones without prior experience in the field of XAI and AI based decision support systems.

The following phase consisted in a user test settled in a scenario-based task delivered with the auxilium of materials showed in a sequence of presentation slides (appendix). Participants were asked to imagine themselves as the protagonist of a scenario in which an AI system is used to help them take a decision in a daily context. The two scenarios used in the experiment has been selected between the fourth used in the user study conducted by Jin, et al. to validate the EUCA Framework. The reason behind this decision is that the scenarios had to serve the same experimental scope: should include examples of AI augmentation applications for an AI novices audience. As a matter of facts, the decisions involved in EUCA scenarios don't require any domain knowledge to be taken. Our experiment targeted 10 persons and, according to literature in qualitative research the sample of users for interviews should be between 5 and 8 to collect valuable insights. That's the reason why, among the fourth proposed by EUCA we have proposed to our interviewers two of them, to allow a generalisation and to avoid any kind of scenarios bias in analysis that followed:

•House task: users use AI to get a proper estimate of their house price.
•Health task: users use AI to predict his/her diabetes risk.

These tasks are critical decision-making scenarios, because their decisions have significant consequences on one's health and life (Health Task) and finance (House Task). Additionally, using two of the EUCA scenarios allowed us to compare our results with their as well as give the chance to improve our research in the future in exploring explanatory narratives correlating explanatory goals and question-seeking, more details can be found in the further development section.

Finally, we have referred to EUCA scenarios because of their use of publicly available datasets to prepare the related explanatory forms: so, the experiments participants, were able to interact with realistic and or seemingly explanations as generated running different XAI techniques.

Experiments participants were assigned to one out of the two scenarios randomly. At the end of the scenario introduction the participants have been provided of information about the input used by the AI system in the scenario to generate the corresponding prediction.

After the scenario introduction was needed to introduce some vocabulary needed to deepen the context of the research and establish a common ground of preliminary knowledge intended to inform participants to be able to take over the interview and provide valuable insights. For that reason participants have been provided with definitions of explainable AI, prediction, explanation, explanatory UI, explanatory form and question type. The last ones has been showed in details and some time has been dedicated to familiarize with them and, eventually, solve any doubt since, later on, they was going to constitute the central part of the conversation.

Established the needed common ground the experiment turned out into a fine-grained analysis of each of the explanatory forms selected combining the explanation type presented in the question driven design approach and UXAI with the EUCA framework ones. They were presented to participants in a randomized sequence to avoid bias in the following up analysis.

Based on the previously mentioned combination of QDDA-UXAI-EUCA 14 end user friendly explanatory form have resulted: Feature relevance, Rule flowchart, Decision rules, Feature shape, Similar example, Decision tree, Feature interaction, Counterfactual example, Typical example, Global feature importance, Feature importance, Dataset, Output accuracy, Performance.

For each scenario the 14 explanatory forms have been prepared with contextualized information and were presented to participants as possible component of the XAI

interface that may be built by designer to help them in make decision in those scenarios.

This experiment phase, aimed to guide the participants analysis of each explanatory form followed the same structure. At a first sight, participants have been asked to think out loud deepening what information they were able to extract from the explanatory forms In order to get what mental model they were able to inform and if those information were easy to retrieve. This qualitative information has been exploited to determine the user friendliness level of each explanatory form and identify, if occurred, elements that may make it hard to understand, according to literature, factors of misapplying. Other questions investigated if the explanatory form has been considered useful to support the prediction given by the system and if yes, why. Those qualitative data has been analysed to determine the qualitative perceived usefulness of each explanatory form. After having collected all those direct and spontaneous information about the perceived usefulness and easiness to use of each explanatory form we moved to ask to what, within the 10 questions proposed by the XAI question bank from Liao et al. may be completely or partially answered with the information retrieved by each explanatory form and why. Each question has been reframed to fit scenarios context (e.g. from ‘how it works?’ to ‘how does the system estimate house prices?’ or ‘how does the system predict diabetes risk?’).

The results of this experiment phase have been aggregated and analysed both from a qualitative and quantitative point of view: the former has been exploited to understand what elements served the participants to find an answer to the questions; the latter to determine a scale of probability that different user’s may find the question answered from each explanatory form and to determine the usability properties (efficiency and effectiveness) of each explanatory form in the context of use of answering to a precise question. The quantitative analysis must be intended only as a support to the qualitative one since the user sample involved in the experiment is not enough to conduct comprehensive quantitative research.

Combining the results of this analysis, we were able to determine what question, within all the one answered by the explanatory form is more likely to be answered, in a complete or partial manner and, thanks to the analysis of why the user determine these answering properties if it is directly or indirectly given thus requiring more or less cognitive effort. We postulate that great usability is achieved if the explanatory form resulted easy to be interpreted and the answer resulted as given by most of the participants in a complete and direct way thus providing a piece of an effective

explanation informing XUI designers to shape the explanatory narrative following user's need.

Additionally, this analysis has served to help XAI community to solve the second user error found in literature: skipping explanations. According to [nohsemi] We claim that the curiosity needed to deepen explanations and to not skip them lay in the designers' capability to provide explanation serving user needs [answering the question they have in mind] in a usable format [exploiting the most usable explanatory form(s) to build the XUI].

During the analysis, the same data used to develop the fine-grained explanatory form analysis: user's correlation between explanatory form and questions answered, has been exploited to analyse the state of art of the ability of the XAI techniques/explanatory form to answer the prototypical questions to AI novices, generating qualitative insights to understand what pieces of information give user's elements to answer completely or only partially to questions and ways of interaction that require more or less cognitive effort to provide answers according to the direct or indirect way to provide hints. This analysis composes the fine-grained analysis of the question presented in the third chapter of the discussion. Each question has been analysed determining which explanatory forms are most suitable to be chosen to answer it for XUI designers following the same usability prioritisation [easy to be interpreted + providing a complete and direct answer]. We claim that those results may inform the XAI research community with actionable insights to lead next development of XAI techniques.

Experiments materials can be found in the Appendix.

6. Discussion

The analysis of experiment results is divided in three sections: AI novices – XAI interaction, Explanatory form analysis, and question analysis.

The first one summarizes general insights gained from the user study about AI novices - XAI interaction in the context of AI based decision support systems.

The second one cover firstly a general evaluation of the state of art of explanation visualisations comparing them according to the user study findings before going deeper into the fine-grained analysis which cover the quantitative/qualitative analysis of each explanatory form in terms of understandability (easiness to be interpreted, perceived usefulness and intention of use) and usability (ability to provide answers to the prototypical question proposed by the questions driven design approach according to the usability paradigm). In the analysis are highlighted found inconsistencies and contribution to literature, as conclusions for each explanatory form are listed the questions – the needs – that its able to fulfil, following the already mentioned usability scale from an AI novice point of view.

Lastly, the third one deals with the evaluation of the state of art of explainable AI technique to answer the protypical question given by literature. The discussion again starts with the general findings come from the user study on user perceptions of question highlighting correlation between question and explanatory goals before moving to a detailed analysis of each of the question which discuss what elements, pieces of information, are exploited by AI novices to catch complete or partial answers to them and what way of interaction make the process directly or indirectly provided. As conclusions, for each question the most usable explanatory form available in literature is provided.

Discussion

AI novices - XAI interaction

6.1.1 AI novices – XAI interaction patterns:

Qualitative analysis of participants claims during the natural interaction with the explanatory forms pointed out common patterns and general knowledge about how AI novices interact with explanations. The main insights have been elaborated as follows:

Correlation between user friendliness and perceived usefulness and question answering

There's a direct correlation between the user friendliness and the perceived usefulness and the explanatory form capability to serve user's needs, as a matter of fact If participants was not considering an explanatory form useful or easy to interpret it resulted as not able to answer to any or really few prototypical question (P06 claimed that the similar example explanatory form was useless to support system prediction, and when directly asked what answer was able to give, no one has been selected. The same for P04 about the feature interaction explanatory form.)

Interactivity affect explanation perceived utility and effectiveness

As find in literature Interactivity is a factor that affect perceived utility: as an example the feature relevance explanatory form has been appreciated especially for the possibility to direct interact with it and intervene to change parameters (P08: 'this one can be useful to me because I can intervene! I can program it I can see what changes if I change the data, I can snoop on it!).

Combining explanatory form is a way to fulfil user's explanation needs.

As literature claims there not a one-fits-all solution when you deal with explanations, the information conveyed by different explanatory forms are complementary and may be combined to fulfil explanation needs (P01 referring to feature shape: "Nice combo with the graph from earlier [feature importance]." Or P10: "in decision rules all the possible and imaginable answers should be formulated and so I might get lost but then I would like to have in the side a

schematic thing that would make me lose less [suggesting to match decision rules and the decision tree]”.; P05 commenting similar example: “It tells me a lot of information because it gives me other houses with similar characteristics and so it gives me more information than the range, but it gives me a very high variability, maybe then I would also like to have the different characteristics [counterfactual example] and I would understand the comparison better.”
Interestingly, participants suggested to merge information from explanatory form belonging to the same category [rules, feature, examples], confirming that, since explanations can overlap in terms of intent or scope the XAI research community should focus to develop new solution that can boost explanation effectiveness having care to propose the right amount of information that not cognitively overwhelm users.

Correlation between questions and information extract that help to answer them.

As the literature about question driven design approach suggests, different question and relative answers are correlated, as an example participants found what if question as a different way to see the how to be that and how to still be this one. Additionally some information extracted from explanatory form are able to provide hints to answer different questions especially for the Why and Why not question (P05 and P06 agreed:” feature shape answer to the Why not question for the same reason is able to answer to the why question”; P07 on decision tree: “Why yes, it tells you the values of the characteristics; Why not, yes again for the values”).

Notably, overall, the why question is harder to be answered than the why not one, probably because accepting a why explanation requires an higher level of confidence in the system capability than the other (P02: It doesn't tell you 'why' like none of the others [explanatory forms] because probably to really know why you would have to be the system designer/developer; P04:” Decision tree doesn't answer to the why question but to the why not yes, I trust more the why not but to get a complete answer to the why I would like more information, only the

parameters is not enough”; P08 commenting similar example “Why [question] yes I get it in relation to others, but only partially, there are no sufficient dat. To why not [quesiton] on the other hand answers completely”).

Not all the explanatory form has the same level of capability to answer questions.

As the fine-grained analysis which follow in the discussion has confirmed, some explanatory forms are more able than others to answer to the same question. During the experiment, participants, after having seen some explanatory form, has begun to compare their ability to answer question and, thus, expressed preferences. That resulted by their evaluation on the level of completeness of the given answer (P01, commenting the ability of feature interaction to answer how it works question: “same response as feature shape but less partially than before” or commenting the rule flowchart:” How to be that: you see the characteristics as they might be changed, but the hierarchy tree [decision tree] does it better because it tells me all of them”; P04 at the 13th explanatory form: “What if? Yes, probably is the one that answer better”; P05 commenting feature interaction and before seeing the rules based explanatory forms: ”Yes, this can help but I would prefer a different visualisation like: if house bigger than 75 square metres the price would be higher”; P07 on counterfactual example:” that was what I could test with feature relevance and it's smarter because it gives you the ones you can actually change.”; P08 at the output accuracy, the 14th explanatory form: “How confident yes, completely answered and much better than all the others”) as well as the cognitive effort needed to get it (P05 on rule flowchart: "I get the answer from the Flow of decision, a bit like the tree [decision tree], maybe a little more directly").

Additionally, accordingly to the law of quantity from [], participants disregarded to have the same information given multiple times through different explanatory forms (P07 seeing the similar example: "What data is answered, but I already knew that information from the input. I don't need this additional explanation, I guess, information given where already knew before”)

User friendliness and understandability directly depend on the visualisation method.

Explanatory forms are characterized by different visualisation method which affect explanation effectiveness and the overall understandability. In general schemes are preferred, because are easier to be interpreted (P04 while interacting with the rule flowchart: ”Again arrows [referring to Decision Tree], schemes are easy to be interpreted”; P01 commenting decision rules: “They are just cells of the ends of the tree conditioned to two features, the tree was easier, so time consuming to read everything, it seems to me not very complete.”). Then pure text and images, that resulted still easy but requiring more cognitive effort to get interpreted. (P04 and P05 agreed while commenting Decision rules "writing in letters makes it more complicated for me to understand at least." “It is easy to interpret but it is very heavy, because there would be many cases, so something schematic would be better”; P05 on similar examples:" Easy to interpret, a little less complicated than graphs."). Math graphs resulted the harder to be interpreted, because [as euca has already underlined] require user to have a higher level of math literature to extract information from them ([P07 on feature shape: “the graph is not much useful just write that the 75 kg feature had had diagnostics between 52% and 80%, and you had solved”; P09 unable to understand the dataset explanatory form and interpreted wrongly the feature interaction one commenting decision rules claimed: “If I have all the case histories I can compare the rules and understand, easier to interpret than graphs information”-

Flow of explanation matter for explanation effectiveness

During the experiment analysis resulted clear that the order in which the explanations are presented to the users and their position in the interaction flows influence their overall understanding. This means that designing explainable AI user experiences require additional efforts to understand users need and make the explanatory narrative as flexible as possible in order to meet all of them for all the possible of users as well, as said before, allow user to get answers to follow up questions until they are satisfied. (P01 asked about database while seeing the first explanatory form [output accuracy] in order to better understand the system behind; both P04 and P10 has some difficulties in understanding the global feature importance because both of them has already

interacted with the local one; P05 suggested to have the confidence and data information in a side page to deepen the reliability and accuracy of the system before starting to use it)

Additionally, local and global explanation has different role's in informing user's mental model, referring generally on the system functioning (global) or directly to the predicted outcome (local). If XUI designers don't take into consideration this differentiation while designing the explanation interaction flow user's may get confused missing the explanation effectiveness goal (P09 on the decision tree: "But is not directly related to my house, this confuse me"; P01 on system performance: "I think it is marginal information, because since what I get from the system is an estimate so we know when it is accurate. But so here we're talking about the system and not the prediction, it's something that has to be said a priori, it's not something that has to do with the prediction and so I think it has to be said a priori.").

Lastly, to design effective explanatory user experiences is important to test the interaction flow and understand what better improve the user AI collaboration, as P02 pointed out: "Do they always put the explanatory form with the prediction? Because if they give it to you before times you can already get an idea of the possible result".

Domain/previous knowledge counts

Users tent to overcome their doubts due to their previous experience on the topic and the task, especially to answer questions like 'how to be that', 'how to still be this' but even why questions: P05 commenting on feature importance: "How to be that: I can refer to the more weighted parameters, here I am more confident from my personal experience"; P07 Answering about questions in similar example: "How to still be this not [answered by the form] but I can intuitively base on my past experience that I can decrease my body weight"; P08 on similar example how to be that: "yes because in comparison to the other but I know because I know how diabetes works"; P09 on Similar example: "help me to understand with what confidence because I compare with what I already know".

Vocabulary importance

In general, to drive understandability and explanation effectiveness is important to have care of the terminology choices, in our experiment the main misinterpretations come out from explanations about algorithm performance, accuracy metrics and labels on graphs.

The more you interact with explanation the more you understand the system and trust it

Overall, it's worth to be mentioned that our experiment, which let participants explore and freely interact with all the possibility that the explanatory space can provide, show evidence about the benefit that explanations provide in informing users' mental models on AI systems functioning and enhancing their trust on them in a step by step process. Even if, as anticipated in the analysis of participants level of confidence on AI infused product for assisted decision making, at the beginning of the interaction most of them was suspicious and 'scared', as the experiment proceeded, by each new explanatory form presented their level of understanding of the system grew as did their confidence and awareness of limitations and opportunities in using the system in the real world. One participant, even for not so effective or useful explanatory form claimed: "Regardless however it is useful, extra information is always good".

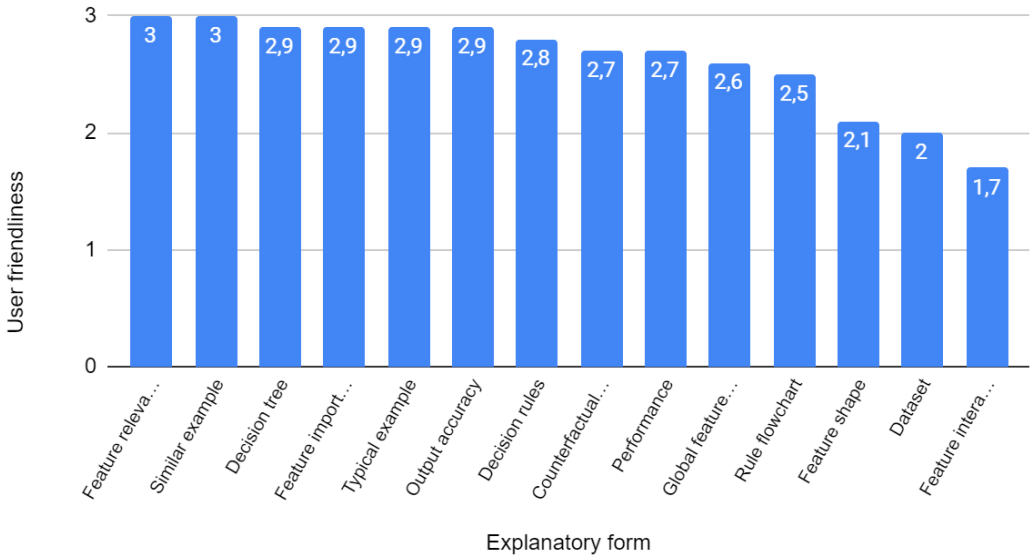
Discussion

Explanatory forms

6.2.1 General insights

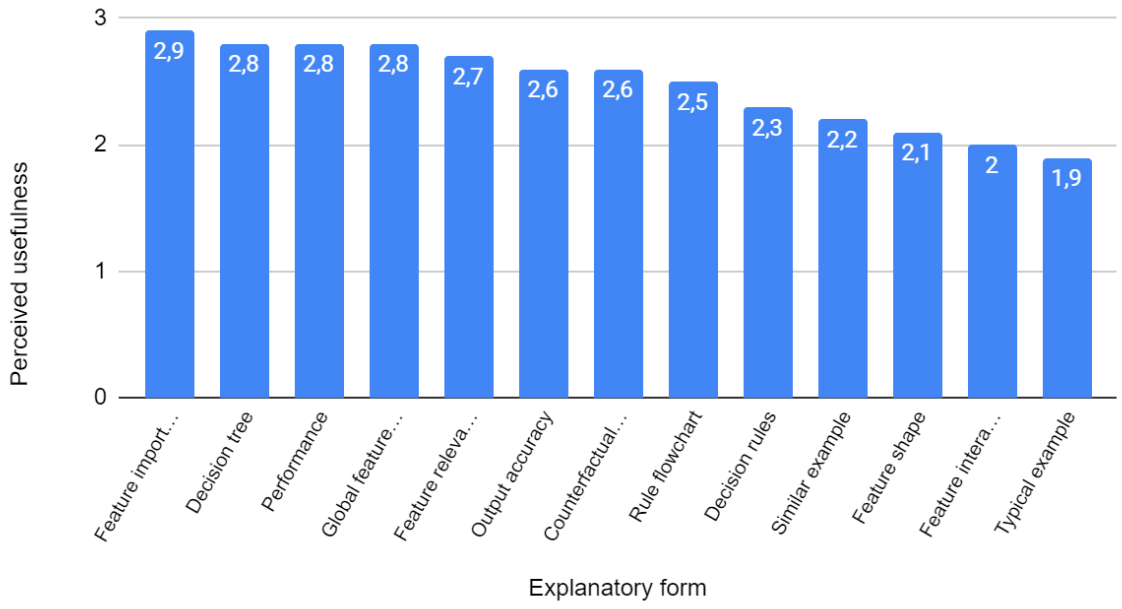
The analysis of users' interaction with the 14 explanatory forms extracted from the literature reveals insightful results on their overall understandability for an AI novice audience as well as their perceived usefulness and intention of use. All those findings are discussed in the fine-grained explanatory forms analysis which follow. This research must be intended as able to provide mainly qualitative insights to inform the XAI research community. The quantitative analysis performed must be considered only as a support for the qualitative one since the sample of respondents is not enough to consider the results able to represent comprehensive qualitative research.

With regards to the user friendliness of explanatory forms extracted analysing user comments on the easiness of interpretation, most of them got great results. The ones resulted less easy to be used in the context of use to get insightful information in the decision-making support scenarios has been the feature shape, the dataset, and the feature interaction explanatory forms. The reason why is directly linked with one of the findings presented in the previous chapter about the visualisation method used to convey information: these three explanatory forms are represented with math graphs which requires previous knowledge to be correctly read and interpreted.



Explanatory forms user friendliness comparison

Directly linked to the easiness to be interpreted we found as the less perceived as useful the same explanatory forms previously mentioned [feature shape, feature interaction and dataset], highlighting the direct correlation between easiness of being understood and perceived usefulness of explanations. In addition to those one, either the similar and typical examples hasn't resulted with a great level of perceived utility, suggesting the possibility that explanation through example is not the best one to provide meaningful information to support users in the decision-making process.

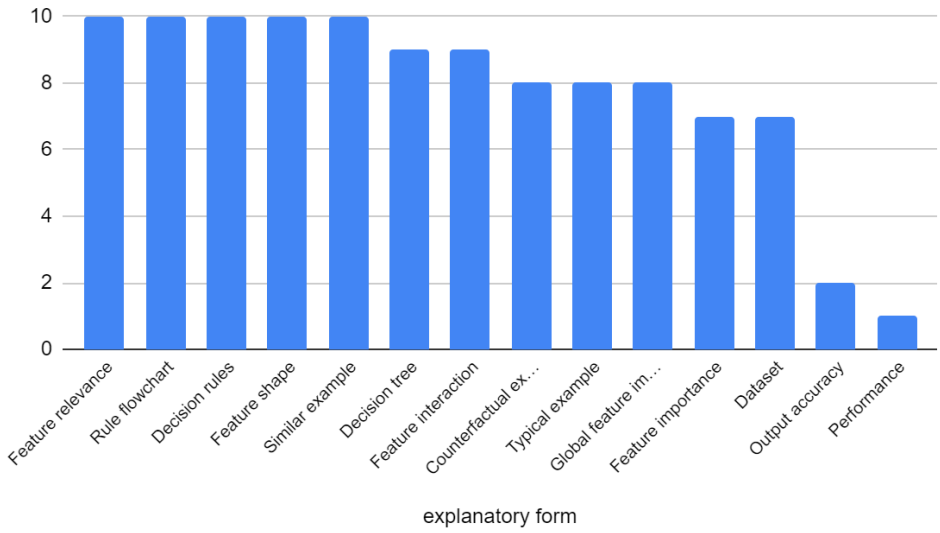


Explanatory forms perceived usefulness comparison

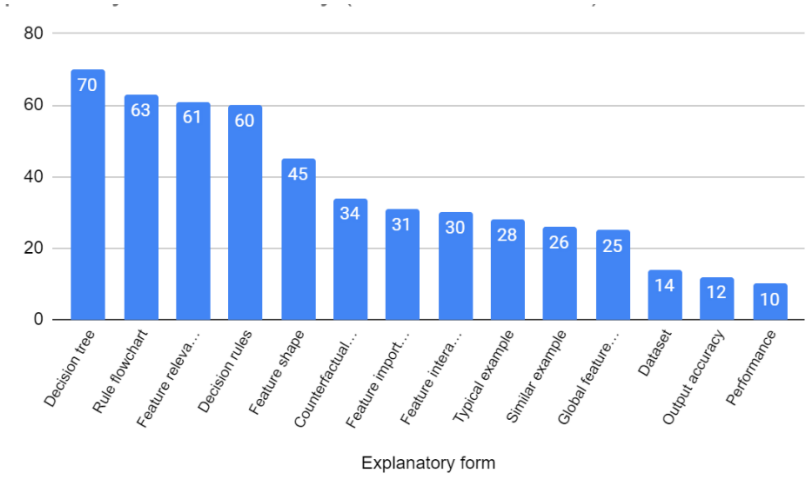
For what concern the explanatory from ability to answer to the 10 prototypical questions proposed by the question driven design approach for explainable AI user experiences the results have been elaborated merging together a qualitative and quantitative analysis to highlight, for each of them what among the questions are most frequently considered answered by experiment participants. Additionally, we were able to extract if the question given has been considered complete or partial and if the way of interacting with the explanation is able to provide the answer directly or indirectly thus requiring additional cognitive effort.

Overall, the 14 explanatory forms analysed resulted as able to serve user needs: in average 8,71/10 experiment participants were able to extract an answer, while interacting with them, at least one out of the 10 prototypical questions proposed.

Overall the state of art of XAI techniques, represented through the explanatory forms resulted as efficient in terms of meeting user needs, most of the them cover a wide range of questions. In details, 5 explanatory forms covered all the needs represented by the questions (feature relevance, rule flowchart, decision rules, feature shape and similar example), all the other covers most of the question, at least 7 out of 10 and only the two related to the confidence level of the system and the prediction [the output accuracy and the performance score] due to their focalized purpose cover only 1 or two of questions (not surprisingly, the how confident one.)



Explanatory form flexibility: question based



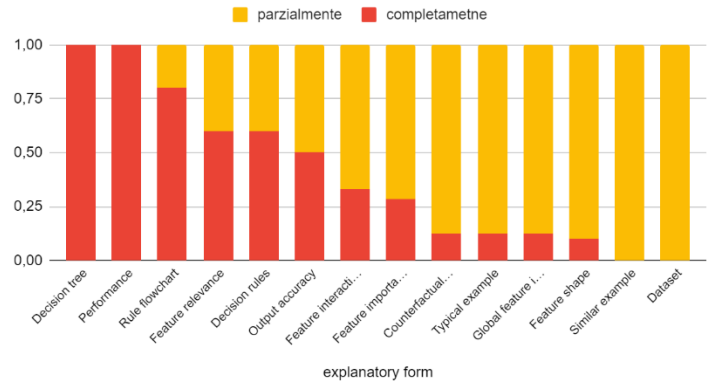
Explanatory form flexibility: mentions based

Analysing the sum of mentions of question answered from experiment participants we can evaluate the explanatory form flexibility, so their ability to convey information that make them suitable to meet most of user needs (answer questions). At the top of the list we find the decision tree, which totalized 70 mentions among the experiment participants making this explanatory form the most flexible. Notably, this list doesn't take into consideration the quality of the answer provided (complete or partial; directly or indirectly)

Of course, not all the question covered by the explanatory forms are answered with the same level of details. The following table and visual representation highlight the result in terms of completeness or partiality of the answer among the one answered by each explanatory form thus their overall effectiveness.

explanatory form	N° question answered	Completely	(%)	Partially	(%)
Rule flowchart	10	8/10	80%	2/10	20%
Feature relevance	10	6/10	60%	4/10	40%
Decision rules	10	6/10	60%	4/10	40%
Feature shape	10	1/10	10%	9/10	90%
Similar example	10	0/10	0%	10/10	100%
Decision tree	9	9/9	100%	0/9	0 %
Feature interaction	9	3/9	33,3%	6/9	66,7%
Counterfactual example	8	1/8	12,5%	7/8	87,5%
Typical example	8	1/8	12,5%	7/8	87,5%
Global feature importance	8	1/8	12,5%	7/8	87,5%
Feature importance	7	2/7	28,6%	5/7	71,4%
Dataset	7	0/7	0,0%	7/7	100%
Output accuracy	2	1/2	50%	1/2	50%
Performance	1	1/1	100%	0/1	0%

The following images highlight if the information conveyed by the explanatory forms are considered able to answer completely or partially to the questions covered by them. As we can see notice the decision tree and the performance cover all the answers given in a complete manner, on the contrary the similar example and the dataset only partially.



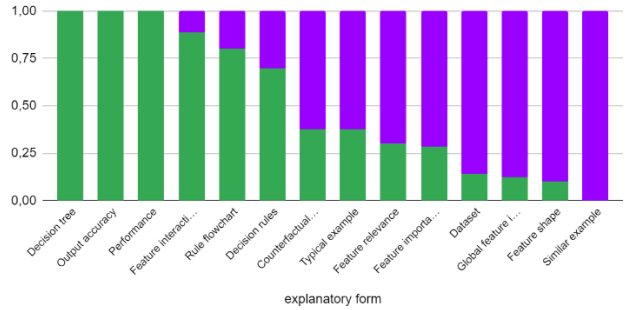
Distribution of partial (orange) or compete (yellow) answers given

Following the same reasoning, not all the question covered by the explanatory forms are answered convey information able to inform users with the same level of efficiency. The following table and visual representation highlight the result in terms of directness or indirectness in informing the answer among the one answered by each explanatory form.

explanatory form	N° question answered	directly	(%)	indirectly	(%)
Rule flowchart	10	8/10	80%	2/10	20%
Feature relevance	10	3/10	30%	7/10	70%
Decision rules	10	7/10	70%	3/10	30%
Feature shape	10	1/10	10%	9/10	90%
Similar example	10	0/9	0%	9/9	100%
Decision tree	9	9/9	100%	0/9	0%
Feature interaction	9	8/9	89%	1/9	11%
Counterfactual example	8	3/8	38%	5/8	63%
Typical example	8	3/8	38%	5/8	63%
Global feature importance	8	1/8	13%	7/8	88%
Feature importance	7	2/7	29%	5/7	71%
Dataset	7	1/7	14%	6/7	86%
Output accuracy	2	2/2	100%	0/2	0%

Performance		1	1/1	100%	0/1	0%

The following image highlight if the information conveyed by the explanatory forms resulted able to inform users directly or indirectly to answer the questions covered by them. As we can see notice the decision tree the output accuracy and the performance ones cover all the answers given in a direct manner, on the contrary the similar example gives all the answers indirectly.



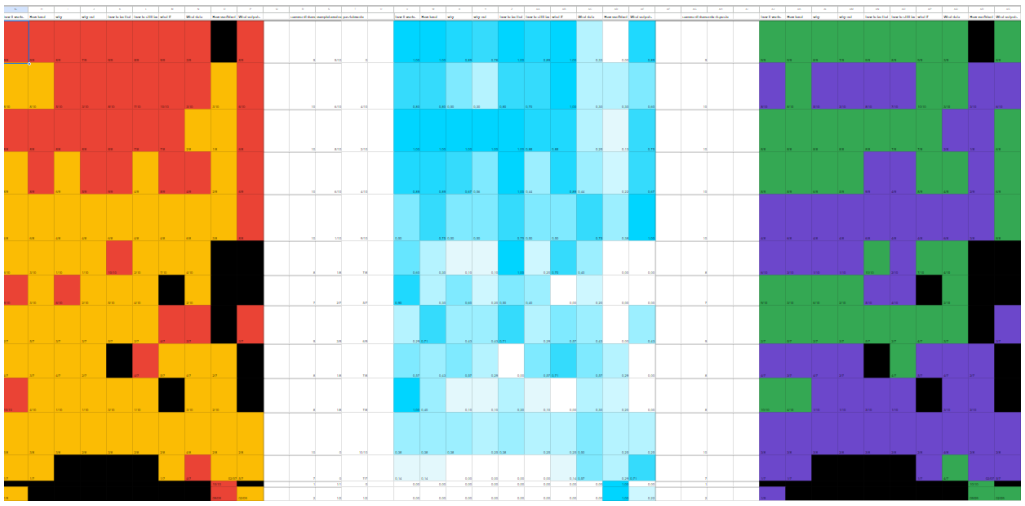
Distribution of direct (green) or indirect (purple) answers given

To conclude with an overview of the state of art of explanatory forms ability to answer prototypical questions has been realised three tables which visually show it that can be found in the appendix.

The first one differentiates them in terms of ability to provide complete or partial answers.

The second one highlight, with the help of a gradient, the percentage of participants which evaluated the explanatory form (row) to answer a question (column) so representing the probability that user would find the answer interacting with the explanatory form.

The third one shows their ability to convey information to answer each question directly or indirectly.



6.2.2 Fine grained analysis:

Thanks to the qualitative analysis of the data resulted from our experiment we were able to elaborate a fine-grained analysis of each explanatory form. The following pages provide the detailed analysis of each of them structured as follows: the first section is dedicated to a description resulted from the literature review: definition and questions answered according to related works. The second section covers the explanatory form analysis in terms of understandability form AI novice perspective: their User friendliness level is accompanied with some highlights on the the identified factors of misapplying turned into suggestion to improve the proposed visualisation and the analysis about their perceived usefulness and applicable context of use. Lastly, the third section covers the quantitative and qualitative analysis of the prototypical questions answered by each explanatory form: what questions are more probably answered for an AI novice audience and why, and if these answers are completely or partially and directly or indirectly given.

It's worth to be specified that, the raw data resulted from the experiment has been filtered and evaluated as follow:

To determine user friendliness and perceived usefulness level the participants answers to the questions 'the explanatory form is easy to be interpreted?' and 'the explanatory form is useful to support the system prediction?' has been synthetized in three numerical level between 1 (no) and 3 (yes), and the average has been calculated to determine the final score

To determine what question-type are answered by each explanatory form only the question mentioned as answered by at least three of respondents has been selected and quantitatively evaluated, and to determine if the question is partially or completely

answered in a direct or indirect way was been considered the majority resulted by respondent analysis.

The raw data of the experiment result can be found in the appendix

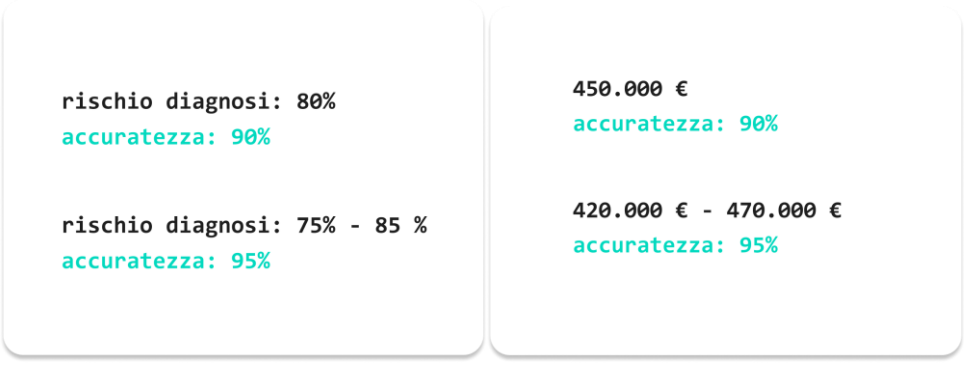
Output accuracy

The output from AI models is usually probabilistic and that the certainty score shows the case-specific level of confidence in the prediction.

One way to express the certainty of the AI model in its prediction is through a category system, such as "high" or "low" certainty. Another way is to use numbers, such as a percentage or a probability score. The model can also provide n-best alternatives, which are the top n predictions with the highest certainty scores. Visualizations can also be used to indicate the certainty of the prediction.

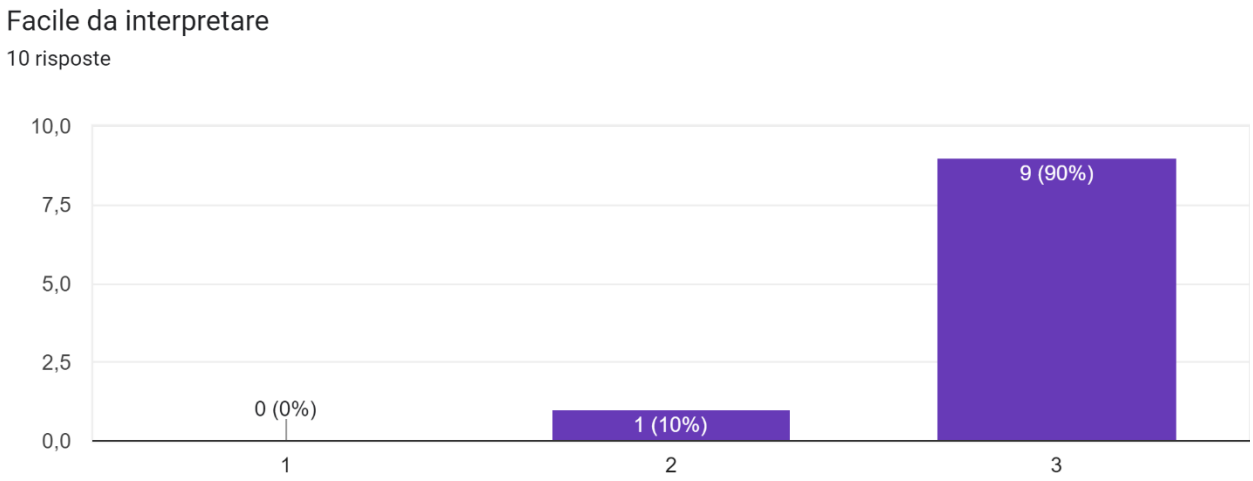
The certainty level of the model can vary depending on the input and the specific task. It is also dependent on the quality of the data and the design of the model.

In our experiment the output card contains prediction information, including a point prediction, a prediction range, and the corresponding uncertainty levels expressed by a percentage.



Understandability

User friendliness and perceived usefulness



The model confidence explanatory form has been evaluated easy to be interpreted by 9 out of 10 participants, the difficulties encountered has been deepened in the factor of misapplying section.
At the question ‘what information this explanatory form is able to convey’
Participants was able to extract multiple information from this explanatory form: analysing them can provide us insights on how the explanatory form is interpreted and what can be the perceived usefulness by users.

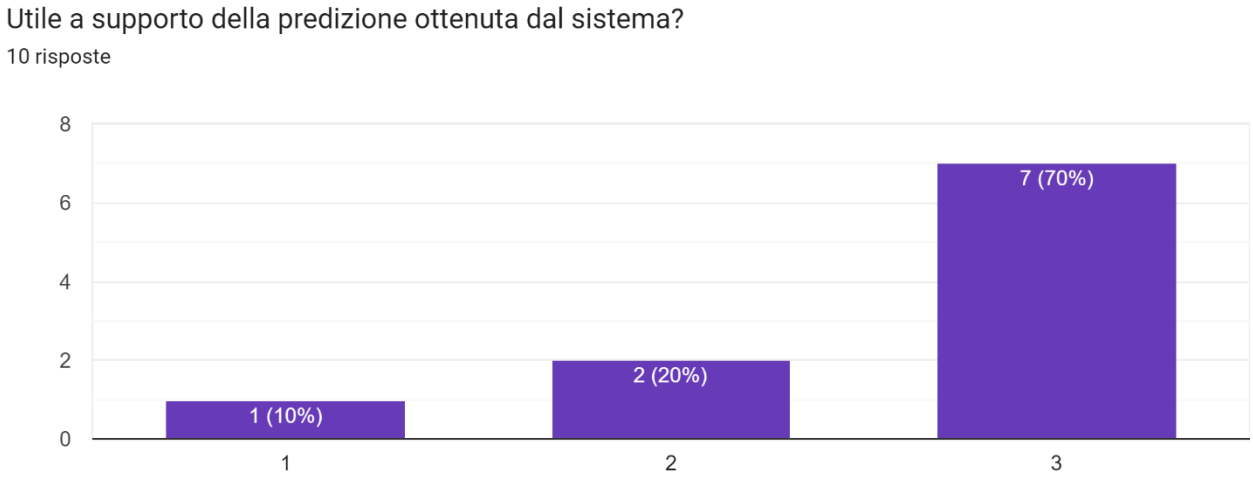
Participants mostly agreed that accuracy is a measure of the confidence of the prediction. To participants was intuitive that the accuracy presented in the form of a range, had higher certainty level than the point prediction. That is why P07 considered it unnecessary to have been provided with both.

This way to provide a measure of the reliability of a prediction, has been compared P08 to how usually are communicated weather predictions. This means that that way of indicating how likely a prediction is to be correct generally matches user’s mental models, confirming the results on the easiness to be interpret.

Even if in general this explanatory form has been considered easy to be interpret and participants has matched literature definition of output accuracy it is important to underline that accuracy may be not a well-known concept and sometimes is needed to be explained. (P06: ‘What is accuracy? truthfulness?’, P09: ‘I don’t know what accuracy means, so for me this is completely useless’)

This results into a design recommendation: clearly define and explain the concept of accuracy to users, as it may not be a well-known term for some of them.

Applicable context of use



According to participants, output accuracy may be a useful tool for evaluating the reliability of the given prediction and that can be used to compare different predictions to make decisions.
Participants also suggested that output accuracy gives an indication of the trustworthiness of the prediction and can provide peace of mind when using the system. However, the text also mentions that accuracy may not be useful in certain cases, such as when the system is already known to be reliable, and P06 highlighted the assumption that the market doesn't allow low accurate prediction tools.
Overall, participants were able to extract from this explanatory form knowledge about the probabilistic nature of AI predictions, thus, talking about accuracy increase their awareness of possible errors and lead users to calibrate their trust in the prediction.

Explanation Usability

Literature claims

In the literature about question driven design approaches the model confidence explanatory form as been mentioned only in the UXAI work which claimed that is able to answer to **a why, why not and how confident** questions

Study results

Nine participants out of ten found this explanatory form able to answer to at least one of the protypical questions analised.

How confident **9/9** **Complete** **Direct**

Output accuracy helps user to answer to how confident question in a complete and direct way. Participants agreed that accuracy is directly link to confidence and it is a measure of the precision of the predictions made by the AI model. Accuracy is considered a precise measure of the reliability of the prediction, giving the user an idea of how much to trust the prediction. Compared to other way to understand how confident the system is, about the given prediction, output accuracy has been mentioned as better than the other methods.

Conclusions

Question type	Mentions	Efficiency	Effectiveness
How confident	**9/9**	**Complete**	**Direct**

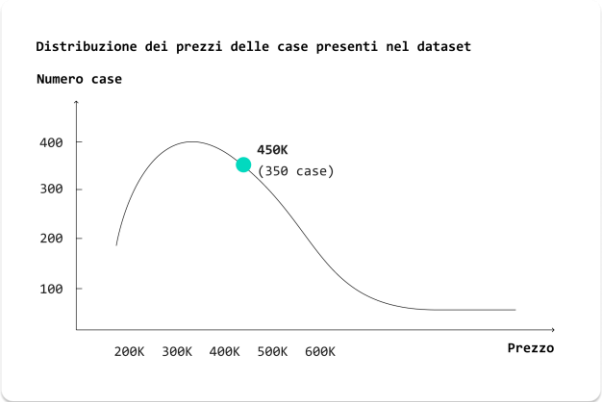
Confirming literature findings the model confidence explanatory form is able to answer to the **how confident** question. **Inconsistently with literature, No** one of the interviewers mentioned that this explanation can provide any clue to answer to a why or why not question.

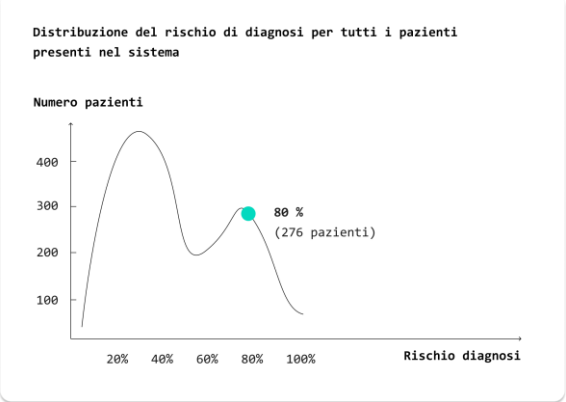
Dataset

Dataset information, such as the metadata on the dataset where the model is trained, show what information the system has access to in order to make decisions. It can help end-users understand the model and identify potential flaws in the data. Interacting with dataset information may happens that users would like to check their own data point within the training data distribution and use it as a dashboard to navigate, identify,

and filter interested instances such as similar, typical, and counterfactual examples, to compare what are the same and different features between their input and the interested instances.

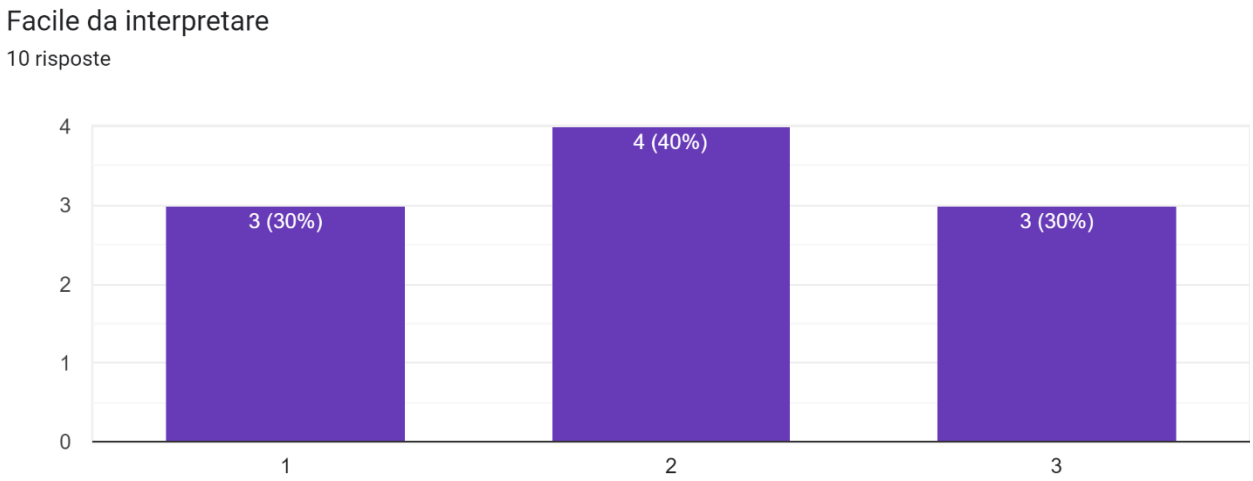
In our experiment the dataset card contains information on the training dataset distribution of the prediction outcomes.





Understandability

User friendliness and perceived usefulness



Overall, the dataset explanatory form has been considered not so easy to be understood by half of participants, the reason are deepened in the factors of misapplying section

Analyzing what participants are able to extract from the graph is easy to explain why the explanatory form is considered not so easy to be interpreted and, consequently, perceived as useful.

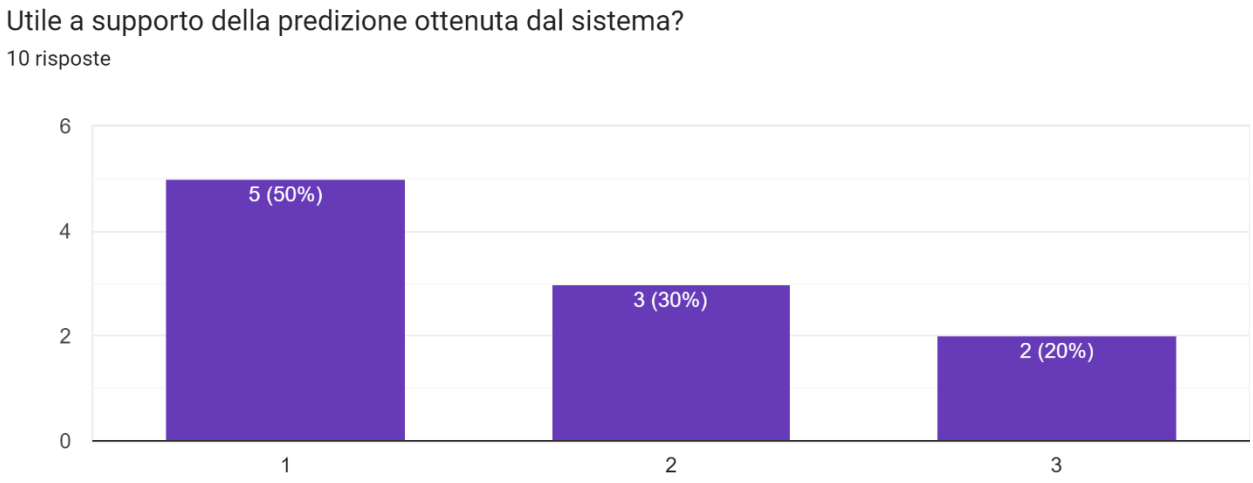
First difficulties were in understanding what ‘distribution in data’ means. That’s why a dataset may be not a well-known concept and to meet different AI expertise probably is needed to be gurther explained. In addition, presenting a dataset in the form of a graph or chart, requires a higher AI/math/visualization literacy, as a matter of facts participants had difficulty understanding the movement or trend in the graph and underlined the hardness to find information such as the total number of patients directly.
Anyway, Once extracted this preliminary information, it is also been mentioned the need to be able to filter the data according to their own interests to accomplish a desire to be able to compare their own data to the data in the graph in order to understand their prediction or position in relation to others.

It’s worth to be noted that users without previous knowledge about AI algorithm has the tendency to consider dataset data as the number of Instances that have been evaluated by the model and not the ones used to train it. As a matter of facts talking about the output that the system can provide P01 stated ‘since they are the results provided since now’. Thus, in terms of output, the dataset information may mislead the users making them think that only the output provided by the chart are the ones that the model is able to provide.

As design recommendation to provide effective dataset explanation we found:

●Clearly labeling and explaining the title of the graph or chart, so that participants can better understand the data being presented.
●Providing additional information such as the total number of instances in the dataset.
●Allowing for filtering of the data according to the participant's interests.
●Help users to compare their input data to the other in the dataset in order to understand their prediction in relation to others.

Applicable context of use:



Since the explanatory form was not so easy to be interpreted participants had difficulties to extract information and consequentially hasn't find it so useful.

Anyway, some applicable context of use of the dataset explanatory form has been mentioned such as:

●Deepening the task-context by exploring the dataset instances

●Comparing their own data to the data in the graph to understand their position in relation to other

●Understand the reliability of the system

Explanation Usability

Literature claims:

In the literature about question driven design approaches the dataset explanatory form as been mentioned only in the UXAI work which claimed that is able to answer to a **what data** question

Study results:

Due to lack of userfriendliness of the explanatory form only 7 out of 10 found the dataset explanatory form able to answer to at least one of the protypical questions analised.

What outputs 5/7 Partial Indirect

Half the participants of the experiment has declared that the dataset explanatory form is able to answer to a what output question, that's because they could see the results on the Y-axes of the graph. Still, this question can be only partially answered because, as P05 stated: 'I can see the price range, but it is not certain, in the sense that this is the dataset but maybe [my result] could also go out of range'. Again has to be underlined that the poor AI literacy of some user can led them to think that this graphs shows all the possible results.

What data 4/7 Partial Direct

The dataset explanatory form was not associated to the what data question, as the literature suggests, by a lot of partecipants. This may be due to multiple factors such as the difficulties encountered on the interpretation of the graph and the previous knowledge about what a dataset is in the context of AI models. Anyway, for those who associated the explanatory form to the What data question, its ability to answer it is considered not so complete again due to the nature of the visualisation: partecipants expresses the need of interact with the to fully answer the question so we can assume that if it's given by a system then the question would be completely answered.

How confident 2/7 Partial Indirect

Even if it has been mentioned only by 2 participants up to 7 the dataset explanatory form may be able to answer to 'How confident question' by providing information about the number of Instances that have been evaluated by the model. Knowing that the model has been trained on a large number of them, participants feel that it increases the reliability of the predictions made by the system. It's worth to be noted that users without previous knowledge about AI algorithm has the tendency to consider dataset data as the number of Instances that have been evaluated by the model and not the oned use to train it .Additionally, the ability to filter the data according to their own interests allows them to compare their own data to similar cases in the dataset and make more informed decisions. This gives them a sense of confidence in the predictions made by the system. Anyway, this information about the confidence has been considered by participants partial and indirectly extractable.

Conclusions:

Question type	Mentions	Efficiency	Effectiveness
What outputs	5/7	Partial	Indirect
What data	4/7	Partial	Direct
How confident	2/7	Partial	Indirect

The dataset explanatory form is able to answer to the **what data** question. Additionally, the dataset form has been evaluated by some participants as able to provide even hints to other question: the ‘what output’ and ‘how confident’.

Confirming literature findings, the dataset explanatory form has resulted as able to answer to a what data question even if for less than the half of participants. This result highlights that, at the state of art of the dataset explanatory form, end users find difficulties in understanding the value that may provide in terms of explanation quality and usefulness. Two more question enriches the literature about this form: the’ ‘what output’ and ‘how confident’

Feature relevance

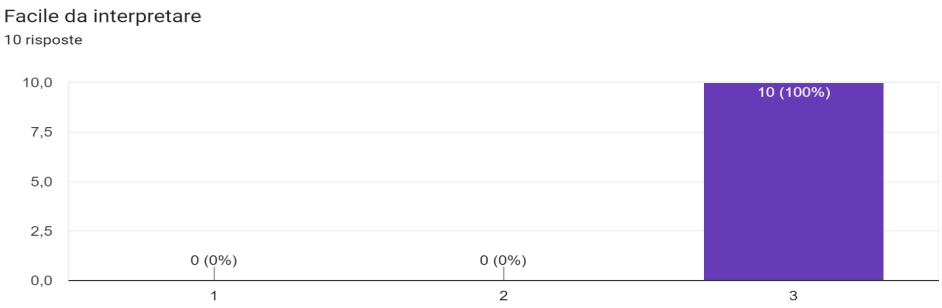
Feature influence or relevance show how the prediction changes corresponding to changes of a feature (often in a visualization format). Is the output of the homonym XAI method aimed to understand the contribution of each feature or attribute of the input data to the output of a machine learning model. This is typically done by showing how the prediction changes corresponding to changes of the input features. Usually is interactive and help the user to identify the most important features, as well as any data that may be having a disproportionate effect on the model's predictions. The preliminary goal that lead to its

development is to understand how the model is making its predictions and to help identify potential issues or areas for improvement.

In our experiment the feature influence or relevance explanatory form was composed by the list of characteristics taken into account to predict the outcome and their corresponding values, and the prediction given by the model. To let participants grab the interactivity of the explanatory form some ‘’chevron’’ has been added near the value of the characteristics according to the drop-down list standards in interface design.

Understandability

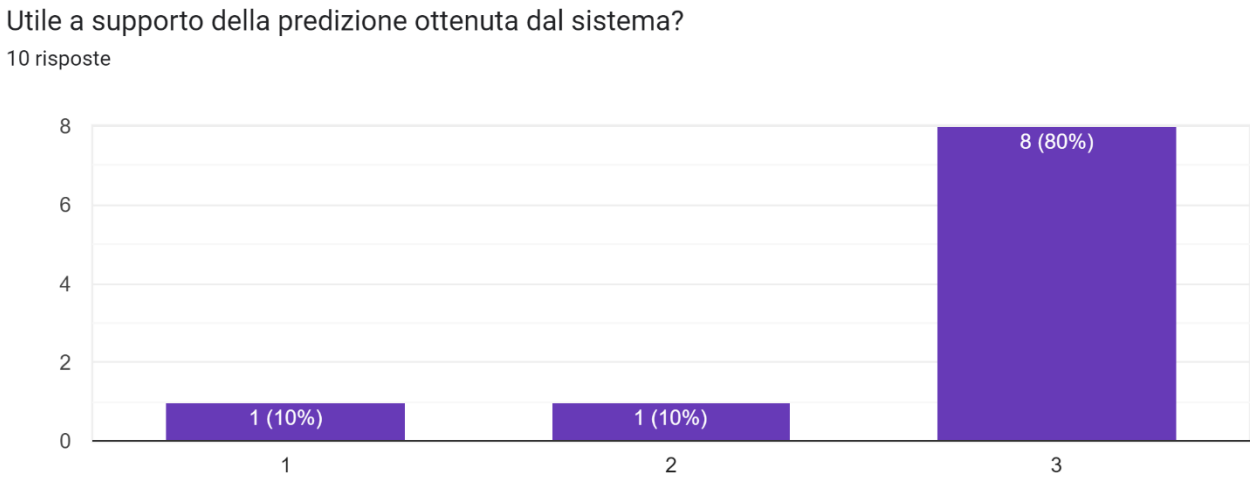
User friendliness and perceived usefulness



This explanatory form has been evaluated easy to be understand by all the experiement partecipants (10/10). As a matter of facts they were able to correctly extract the information about the characteristics of the input, getting to know what are the parameters that the system use the evaluate it and they recognized the possibility to interact with them to see how the result changes.
While describing the information they were able to extract by the interaction with this explanatory form they mentioned the possibility to understand how different parameters influence the output according to the changes, and hypothesising changes in their input understand how to improve the prediction obtained by the system.

Even we can't speak about factors of misapplying some suggestion to improve has been given let the user interact only with the characteristics that can be changed in the context of use, provide additional information about what parameter is weighted more than one other in changing the outcome (in the diabetes risk scenario 'the dangerous ones'), in this way adding a counterfactual goal to the explanatory form

Applicable context of use:



The feature relevance explanatory form can be useful in several ways according to the partecipants:
●it can aid in planning for future actions by understanding the impacts of different input data on the model's output.

●it can provide an understanding of the factors that the model takes into account and how they interact to produce the output.
●it can be used to identify the most important features that are being considered by the model.
●it can be used to understand how the model is making its predictions and can aid in understanding the inner workings of the model.
●it can be used to experiment with different input values and see how it affects the model's output in real-time, which can help in understanding the system better.
●it can be used to understand how to reach a certain risk or prediction outcome.
●it can be used to understand what changes can be done in the input data to obtain a different outcome.

Overall, the feature relevance explanatory form is seen as useful in providing insights into how the model is making its predictions, understanding the factors and attributes that the model takes into account, and providing a way to experiment with different input values and see how it affects the model's output.

Explanation Usability

Literature claims
In the literature about question driven design approaches the model confidence explanatory form as been evaluated in the first work about a question driven design approach for XAI as able to answer mainly to a **what if** question, but less precisely even to the **how to be that and how to still be this questions**. Lately, the UXAI work added two more questions that may be answered by the form: **why and why not**.

User study results
All the ten participants found this explanatory form able to answer to at least one of the protypical questions analised.

How it works ● **8/10** ● **Partial** ● **Indirect**

The feature influence and relevance explanatory form is able to give some general information on how it works, but it does not provide specific details that why has been

evaluated mostly as partially able to answer the question. Thus, the explanatory form may help to understand the general logic of the system by showing the parameters that are taken into consideration, and some users may be able to infer more information by interacting with the system and experimenting with different parameters.

How **8/10** **Partial** **Direct**

The feature influence and relevance explanatory form can answer the 'how' question by providing information by showing the features and corresponding values and helping in understanding that depends on the correlations between features. However, some of the responses indicate that understanding the exact calculation and reasoning behind the predictions may not be clear.

Why **5/10** **Complete** **Indirect**

The feature influence and relevance explanatory form can partially answer the "why" question by paying attention on the values of the different parametres that generate the AI output, but it may not be completely clear or accurate. Some partecipants claimed that’s not enought to fully understand how the calculation is performed.

Why not **3/10** **Complete** **Indirect**

The feature influence and relevance explanatory form may be able to partially answer the "why not" question. The reasons for this include the ability to understand the reasoning behind a decision by changing parameters, and the ability to adjust values to arrive at an answer, but not necessarily understanding the specific factors that carry the most weight.

How to be that **8/10** **Complete** **Indirect**

The feature influence and relevance explanatory form can be used to understand how to obtain a precise result, and thus what changes make to the input to achieve it, by adjusting values of features by trial and error. This activity can even help to identify specific data points as influential. Anyway, the process described is perceived by user as risky, as P01 claim ‘I could blame one feature more than another’. Generally speaking the process to gain an answer to the question is considered as indirect and time-consuming so, for most of the users, not so efficient.

How to still be this **7/10** **Complete** **Indirect**

The feature influence and relevance explanatory form can be used to answer the question of "how to still be this" by experimenting with different combinations of values and using trial and error method. However, some respondants express concerns about potentially inefficient or unhealthful methods (in relation to the context of the diabetes risks) that may be used in the process.

What if **10/10** **Complete** **Direct**

The feature influence and relevance explanatory form is able to completely answer "what if" question by trial and error thanks to the direct ability to modify parameters.

How confident **3/10** **Partial** **Indirect**

Few respondents claimed about the feature influence and relevance explanatory form to give some partial and indirect information to answer to the ‘how confident question’. In particular P05 stated: 'making all these changes would help me a lot to understand how the accuracy of the modification would have a different and therefore more accurate and reliable result' the others two respondents only appreciated the range in the results.

What outputs **6/10** **Complete** **Indirect**

More then the half of respondents claimed the what output question may be answered through the interaction with the feature influence and relevance explanatory form, by the way the process to get all the possible results has been of course described as too effortful

What data **3/10** **Partial** **Direct**

Few respondents found the feature influence and relevance explanatory form, able to answer to the ‘what data’ question. This answer was given by partecipants to which having information about the features taken into consideration by the model to take decision was enough to consider the question answered.

Conclusions

Question type	Mentions	Efficiency	Effectiveness
What if	10/10	Complete	Direct
How to be that	8/10	Complete	Indirect
How	8/10	Partial	Direct
How it works	8/10	Partial	Indirect
How to still be this	7/10	Complete	Indirect
What outputs	6/10	Complete	Indirect
Why	5/10	Complete	Indirect
Why not	3/10	Complete	Indirect
How confident	3/10	Partial	Indirect
What data	3/10	Partial	Direct

The feature relevance explanatory form is able to answer completely the the **what if** question but the results highlight a great ability to answer even to the How it works/How, How to be that and How to still be this questions. Less mentioned the why, why not how confident and what data questions.

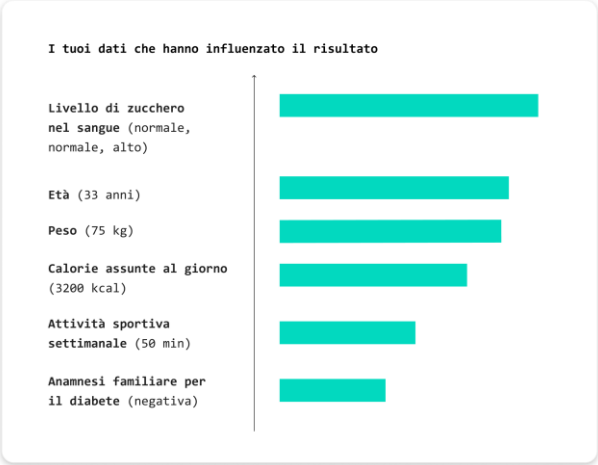
Confirming literature findings, the model confidence explanatory form is able to answer completely the **what if** question. Most of participants agreed either to the ability of the form to answer to the **How to be that** and the **How to still be this question**. Despite the UXAI findings only the half of participants considered the form able to answer to a **why**

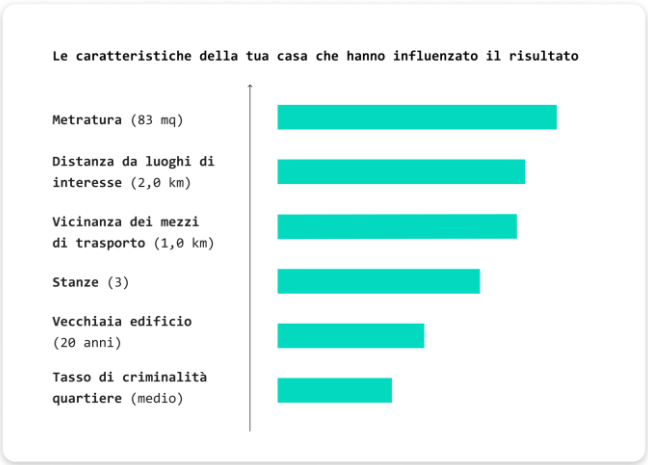
question and only three to a **why not** one. Outside of literature findings our experiment shows that the feature influence explanatory as able to provide hints even to the **how it works, how, what outputs,** and less frequently to **the how confident and what data questions.**

Local Feature importance

Local feature importance and saliency explanatory form is used to explain how a model's prediction is made by identifying which features of an input are important to the decision and their contribution to the prediction. This can include a list of key features and their importance scores for a specific task or a visual representation, such as a color map, to indicate important parts of an image for recognition. These methods assume that the prediction is explainable by linearly addable important features.

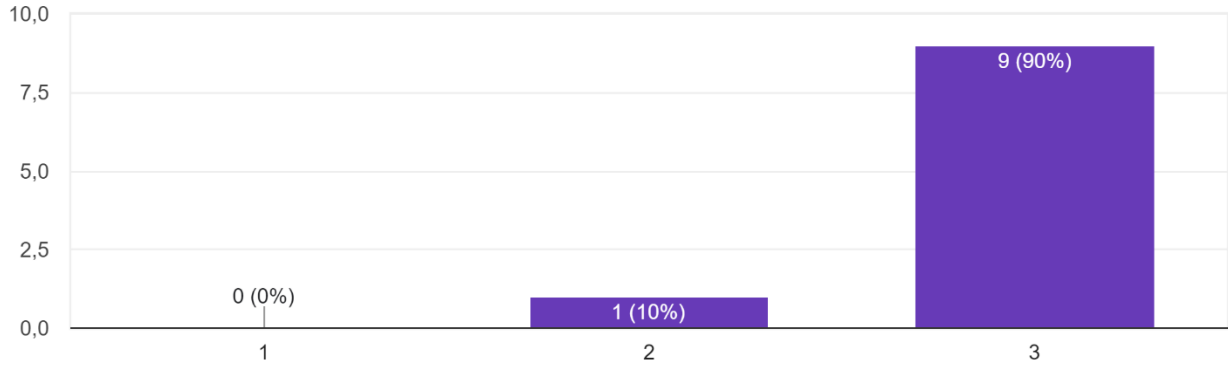
In our experiment the Local feature importance and saliency explanatory form showed the list of the characteristic of the user input with their value in bracket and a bar chart to represent their importance score to the prediction





Understandability

User friendliness and perceived usefulness



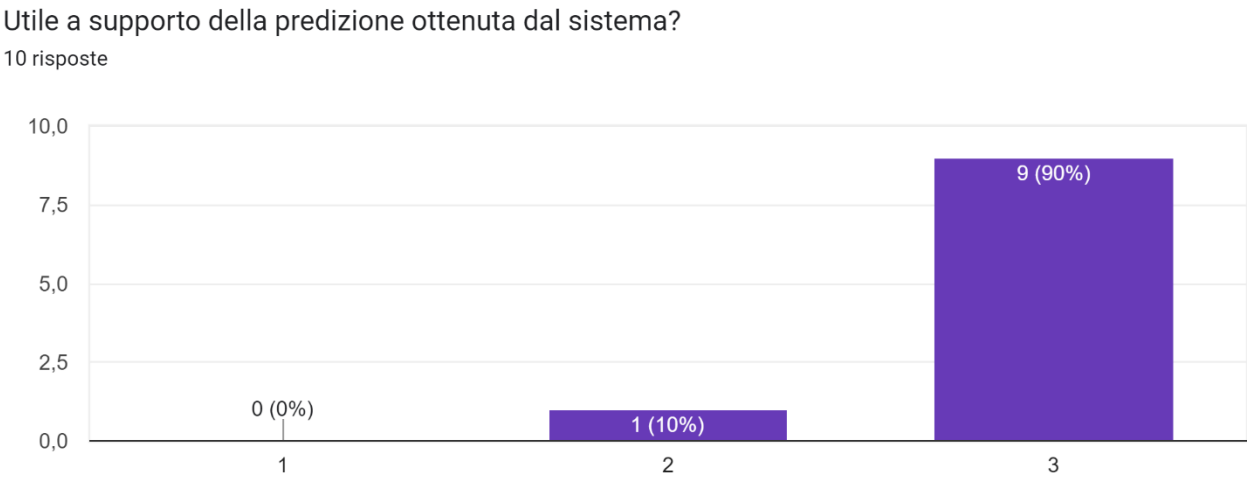
"This explanatory form is easy to be interpreted? Yes = 3, No = 1

The feature importance explanatory form has been evaluated as easy to be interpreted by 9 out of 10 participants to the experiment.

Pertecipants to the experiment found the form easy to interpret, as a matter of fact they were able to understand his main goal: explain the weight or influence of different features on a prediction or outcome.

Even if we can't speak about factors of misapplying the lack of units of measurement for the importance scale make user doubt about the precise weight of the features, thus, to improve the visualisation this information can be added directly on the lines which convey the importance or in hover when the mouse enter in. In addiction, if the task can have a numerical output, partecipants expressed their interest in knowing if the influence is positive or negative.

Applicable context of use:



The participants generally found the feature influence explanatory form to be useful in supporting the system output. They noted that it can help in understanding the weight of different features in determining the final result. Identifying which features are most important and influential in the output, it's useful to inform what changes can be made to improve the given output. Some participants also mentioned that the explanatory form can provide a better understanding of how the system works and how data is used in calculations. One participant expresses the usefulness that the explanatory form would have ifyou would like to learn about the system main goal.

Overall, the feature influence explanatory form is seen as helpful in identifying which features and data are most impactful and in modifying the output.

Explanation Usability

Literature claims

In the literature about question driven design approaches the feature importance explanatory form has been mentioned in the first work from Liao et al. as ablet to only aswer the **why** question. Lately UXAI mentioned it as able to answer to the **why not and how to be that question.** Euca, agreeing with Liao, **mentions** the form as able to answer **why** question but even tha a **how one.**

Study results

All the ten participants found this explanatory form able to answer to at least one of the protypical questions analised.

How it works 9/10 Complete Direct

The feature importance explanatory form helps users understand the logic behind the model showing the weight of each feature, which gives users an understanding of how the model is making its predictions. However, it should be noted that the form does not necessarily provide a complete understanding of the model's workings, and it may not be clear that the given weights are consequential of the consequences of your input data and not the overall logic behind the system

How 3/10 Partial Direct

The feature importance explanatory form helps users understand how certain parameters influence the result, but it may not provide full understanding of all other cases.

Why 6/10 Complete Direct

The feature importance explanatory form allows users to understand the "why" behind a result combining insights from their feature values, and their weights, or other factors that led to a certain outcome. It also gives users a clear understanding of which factors are considered most important and how they contribute to the final result. This information can provide users with greater confidence in the logic and accuracy of the result.

How to be that 5/10 Partial Indirect

Allowing users to see which features have the greatest weight or impact on the result, the feature importance explanatory form may suggest to users which parameters, if modified, would consequentually lead to another outcome. To 5 respondants to 10 changing the most relevant parametres would improve the result, anyway, this indirect process has been recognized to provide only a partial answer to the how to be that

question since don't provide any hints about the result that would be obtained through the change of the feature. It's worth to be mentioned that there's not a linear correlation between the change of a feature and the result, and, for not experienced user, this is probably needed to be explained, maybe through a tooltip in the interface.

How to still be this 4/10 Partial Indirect

Allowing users to see which features have the lower weight or impact on the result, the feature importance explanatory form may suggest to users which parameters, if modified, would probably not change the result significantly. As for the how to be that question, some participants express the conviction that change one feature would automatically change the prediction, misunderstanding the principle behind AI systems functioning.

Conclusions

Question type	Mentions	Efficiency	Effectiveness
How it works	9/10	Complete	Direct
Why	6/10	Complete	Direct
How to be that	5/10	Partial	Indirect
How to still be this	4/10	Partial	Indirect
How	3/10	Partial	Direct

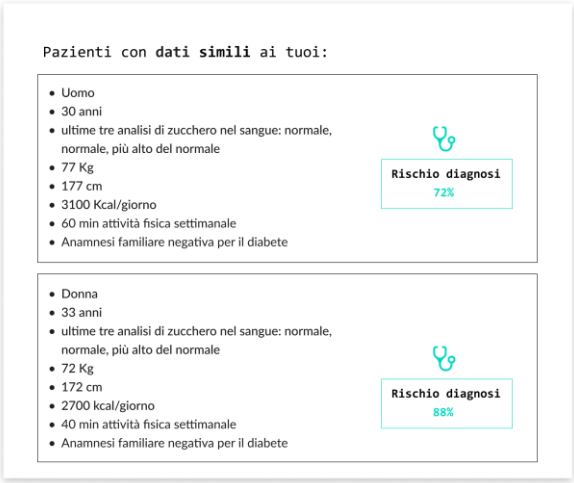
The feature importance explanatory form has been evaluated has able to greatly answer to the **how it works** question. Then, in order of frequency, we found the Why, How to be that, How to still be this, and How questions

The experiment findings mainly confirm the literature about what question the feature importance explanatory form is able to answer: as Euca suggested user found the form

able to provide multiple hints about the general functioning of the system and how it is able to predict outputs. Confirming Liao results more than half participants found the form able to provide why answers, and, confirming UXAI addition half of them mentioned the How to be that question. Additionally, the form has been evaluated as able to provide hints about the how to still be this question. Contradicting literature findings, only two out of 10 participants mentioned the why not question, that according to our methodology is not enough to be considered in the list of question answered by the form.

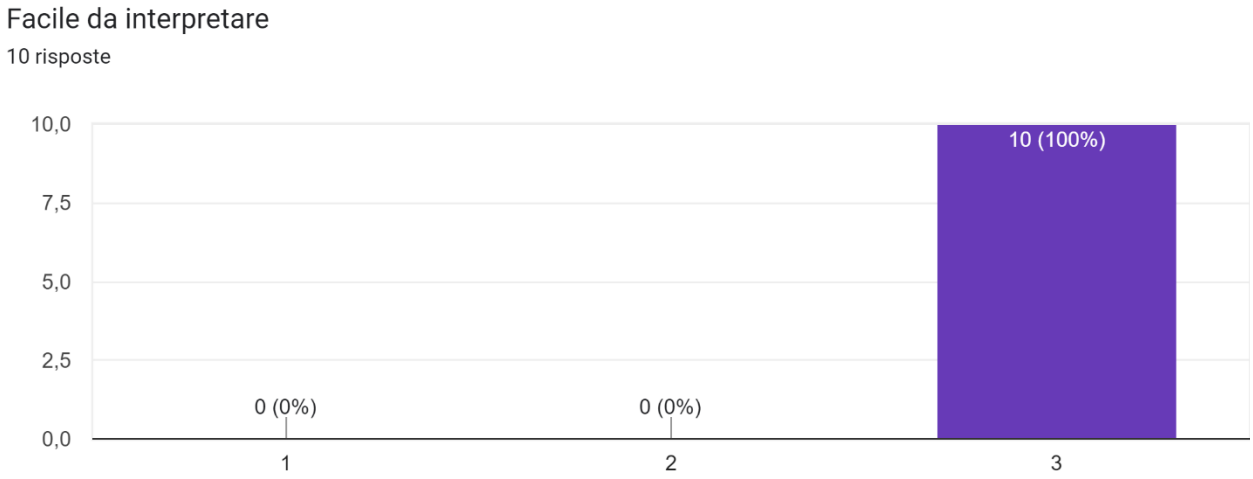
Similar example:

Similar examples are instances that are similar to the input data in terms of their features, and they have the same record as the prediction.
In our experiment the visualisation of the similar examples comprehends information about the main features or characteristic of the instances provided as similar to the partecipants input and their prediction result



Understandability

User friendliness and perceived usefulness

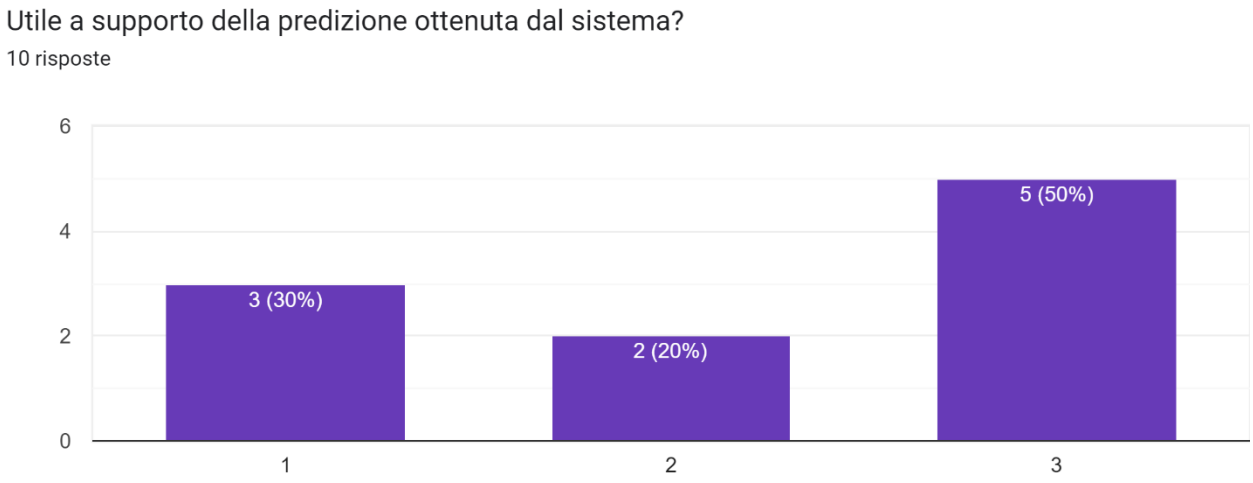


The similar example explanatory form has been evaluated as easy to be interpret by 10/10 participants at our experiment.

Seeing the similar example explanatory form participants focused their attention on the characteristics of the similar instances provided identifying their properties. Some found the visualization useful for understanding the variability and correlation between data and outputs, while others wanted more information and context to better interpret the results. Anyway, most of participants agreed on the intention to comparing their own data to that of others in order to better understand their own prediction. It's worth to be underlined that the first interaction with this explanatory form has been conflictual for a good amount of participants which struggled to find it truly useful in the purpose of supporting their predicted outcome.

No element in the similar example explanatory form has been recognised as able to be misunderstood or misapplied by the participants. Anyway, to still improve the form visualisation and maybe affect the overall perceived usefulness one possible direction can be allow the user to navigate and explore all the feature of the similar examples provided.

Applicable context of use:



As the quantitative analysis suggest overall partecipants haven't identyfied the similar example as an explanatory form useful to support the system predicted outcome. Especially because the main information extracted, the characteristics on which is based the algorithm functioning has been mentioned as already known by the couple input-output. Anyway, this explanatory has been mentioned as useful to:

●Compare properties of the different instances and their predicted outcome

●Understand what the correlations between data may be and partially understand how the evaluation is performed by the system

●In some cases, can allow for imagining changes to obtain a different outcome

Explanation Usability

Literature claims:

In the literature about question driven design approaches the similar example explanatory form as been evaluated in the first work about a question driven design approach for XAI as able to answer mainly to a **why** question, but less precisely even to the **how to still be this question.** Lately, the UXAI work added the why not question to the list of question that may be answered by the form.

Study results:

How it works 3/8 Partial Indirect

Participants had mixed opinions on whether the explanatory form helps them understand how the system works and only 3/10 has answered positively but claiming that the answer provided is mostly indirect and partial. Those ones mentioned that they were able to understand from the similar examples that the system functioning is based on parameters and that it compares their data with similar cases, and this partially helps them understand the general logic.

How 3/8 Partial Indirect

Even for the how question only 3/10 partecipants has answered positively and again claiming that the answer provided is mostly indirect and partial The answer has been mentioned to may be deducted analysing the characteristics of the instances with the similar prediction but at the same time acknowledging that there may be variations within a certain range.

Why 3/8 Partial Indirect

The similar example explanatory form has been evaluated by partecipants as not able to fully answer the "why" question (3/10, partially and indirectly). Some information about it can be given comparing the characteristics of the examples provided but they mentioned a lack of knowledge about how parameters values can vary to stay within a range, incomplete data, and a lack of sufficient data in relation to other factors.

How to be that 3/8 Partial Indirect

The similar example explanatory form has been evaluated by partecipants as able to provide some hints to answer the "how to be that" question by 3/10 participants. This would have been done by comparing input data with the given examples but only thanks to the use of previous knowledge about the task of the scenario. This confirms the low result in terms of participant which agreed about the ability to answer the question.

What data 4/8 Partial Indirect

Four out of ten participants expressed the opinion that the similar example explanatory form is able to partially answer to a what data question. To answer, the information were found in the list of features, but as a participants claimed: ‘I had them already because of my inpu. Another participant mentioned the the question may be partially answered because ‘the example should came out from the database’.

Conclusions

Question type	Mentions	Efficiency	Effectiveness
What data	4/8	Partial	Indirect
How it works	3/8	Partial	Indirect
How	3/8	Partial	Indirect
Why	3/8	Partial	Indirect
How to be that	3/8	Partial	Indirect

The similar example explanatory form has been evaluated has able to partially answer to all the question extracted by the experiment analysis: What data, How it works, How, Why and How to be that.

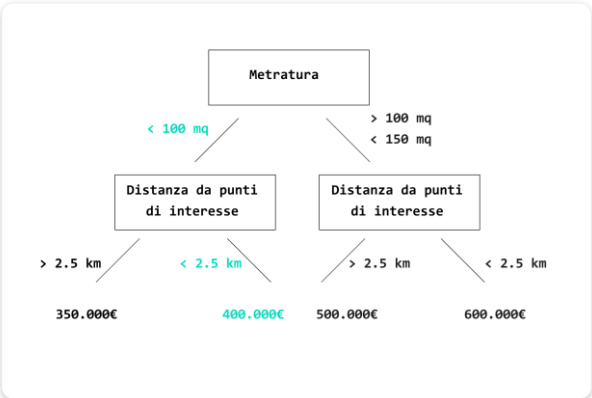
Overall, the similar example explanatory form has been recognised as able to answer to a very few amounts of question compared to the other explanatory forms analysed during the experiment, this result, confirm the direct answer from the participant about the perceived usefulness of it. Comparing the question answered to the one suggested by literature no participants has mentioned it as providing some answer to a why not or how to still be this question; for the last one, the reason is probably because the visualisation presented some examples with the same output of the input instance and some slightly different. Probably due to same reason few participants found the form

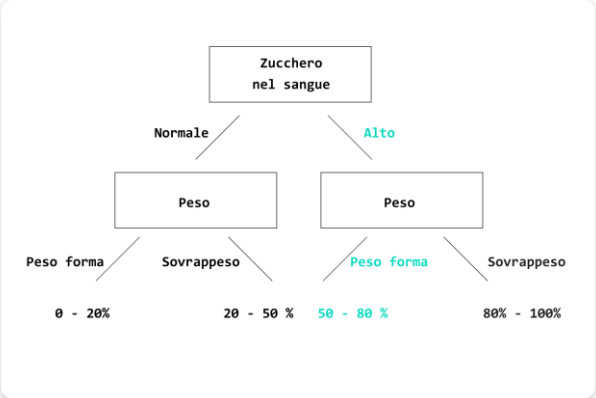
able to answer to a how to be that question. The experiment results suggest even that the similar example explanatory form may provide to some users hints about the How it works, How and what data questions.

Decision tree

A decision tree approximation explanatory form is a graphical representation of a model where the rule is represented as a tree structure, with branches representing the decision pathway and the leaves representing the outcome. The decision tree is used to approximate the model in a way that makes it interpretable.

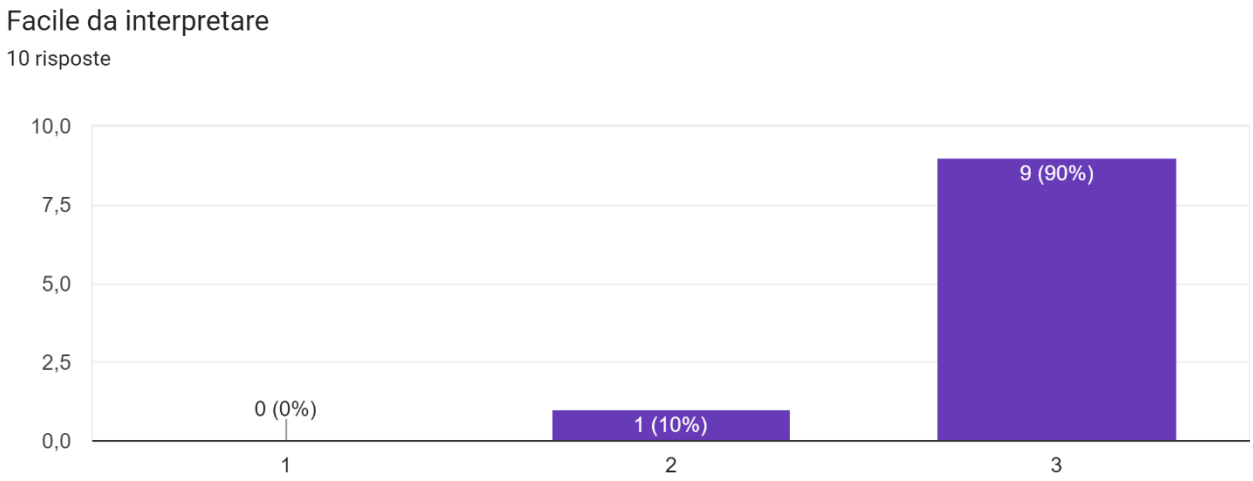
In our experiment the decision tree was simplified representing four possible paths and their corresponding predicted outcome ranges according to two particular characteristics taken in consideration by the scenario related AI based decision making support system. The case-specific path followed to provide the participant prediction was highlighted with a different colour.





Understandability

User friendliness and perceived usefulness



The decision tree explanatory form has been evaluated as easy to understand by 90% of participants (9/10)

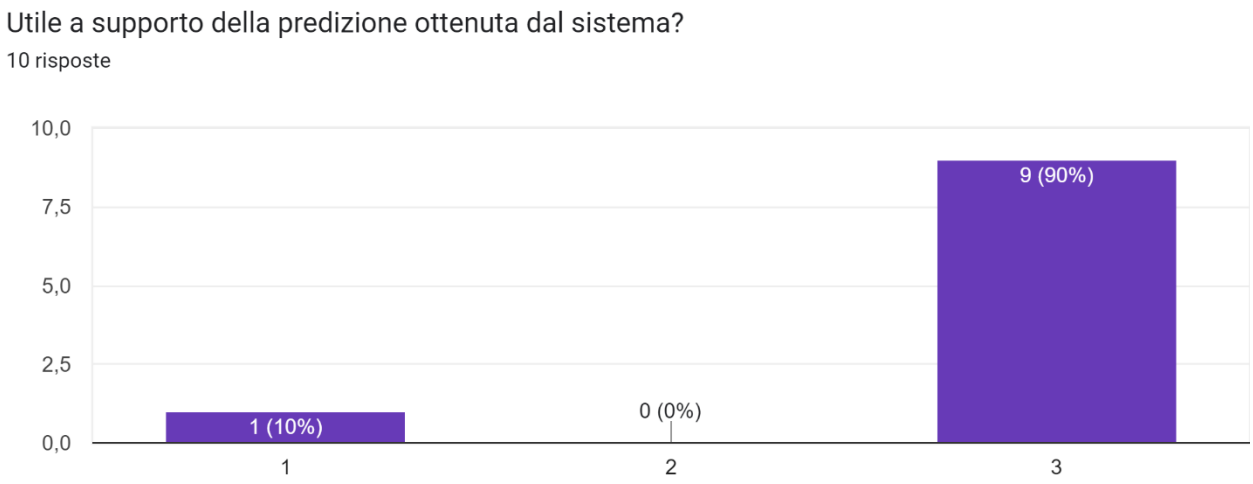
In general, this explanatory form has been appreciated in relation to the amount of information that it is able to convey, as one participant said ‘it gives me all the information that I need’, underlying that more information are even a matter of satisfaction, in the context of explanation seeking.

Some of the key pieces of information that participants were able to extract from the decision tree explanatory form visualization are:
●The logic and flow of the decision tree, which helped them understand how it worked and how to navigate it
●The importance of certain features in relation to the decision tree and how they affected the final outcome
●The range of percentage values based on the input data, which gave them a sense of the precision and reliability of the results
●The flow of decision-making and how the system arrived at a final result
●The factors that increase or decrease the final outcome
●The relevance of certain factors in relation to the final outcome.

With regards to the analysis of misapplying factors, some participants mentioned that the visualization was too standardized and not specific to their input, this suggests that the highlighted path for the ‘decision’ related to their feature was not so visible. Another factor of misapplying is linked to the approximative nature of the form, one participant

claimed: ‘I too could arrive at the same result with this data’ while discussing about the information it is able to convey, so the suggestion is to explicit in the visualisation or with a label that this explanatory form only an approximation of how the system perform its evaluation.

Applicable context of use:



In general participants agreed on the useful nature of the decision tree approximation In supporting the system predicted outcome
In particular, the following context of use has been mentioned:
- Helps participants understand the prediction by giving insight into the system's logic, by highlighting the features that led to the result and in general providing reasoning behind the outcome,
- Helps participants understand how reliable the system is allowing them to perceive it as rational and trustworthy

Explanation Usability

Literature claims

In the literature about question driven design approaches the decision tree approximation has been mentioned both by UXAI and the work from Liao et al. as able to answer expecially to a how it works question, but even to **a why, why not and what if ones.**

Study results

Nine out of ten participants found this explanatory form able to answer to at least one of the protypical questions analised.

How it works 9/9 Complete Direct

The participants found the decision tree approximation explanatory form helpful in understanding how the system works in a mostly completed and direct way. They appreciated the visibility of what features were more influential according to the order of the flow and the ability to see the logic behind the results. Although it does not explain the exact reason for each decision, it gives a general idea of the decision-making process. The parameters and values were also noted as helpful elements in the explanatory form.

How 9/9 Complete Direct

The participants suggest that the decision tree approximation explanatory form helps users understand under quat condition does the system predict an outout by allowing them to see all parameters in reverse, thus providing a complete understanding and justification of the prediction process.

Why 8/9 Complete Direct

Partecipants explained that the decision tree approximation explanatory form can provide answers to the "why" question mainly completely and directly by showing the values of the features and following the parameters. However, some partecipants expressed the need of more details to completely answer the question, one in particular, claimed that information about the values range which belong to the branch's labels (e.g. alto, peso forma diabetes risk context can hinder their understanding,

Why not 7/9 Complete Direct

The summarization of the participants answers suggests that the decision tree approximation explanatory form can help users answer the 'why not' question by

providing the values of the features and allowing users to navigate through the tree structure to understand the decision making process. As for the why question a best practice should be don't hide value ranges with general labels not clear for not expert In the task topic task. (e.g alto, peso forma)

How to be that 9/9 Complete Direct

The explanatory form is completely able to directly answer to the how to be that to all the partecipants. To address the answer the decision tree approximation guide users through the caracteristics and their values letting them to follow alternative paths.

How to still be this 8/9 Complete Direct

According to participants the decision tree approximation explanatory form allows to be aware of how maintain their current prediction by checking the value of the parameters while considering multiple paths, through the tree structure. However, some users find the process to be not so efficient. The form assists in finding the answer by providing a systematic approach to evaluate various characteristics and make decisions based on their relative weight.

What if 9/9 Complete Direct

All the partecipants agreed that the decision tree approximation is able to completely and directly answer to the what if question thanks to the tree structure that allow to follow different path and delve deeper in the prediction for all the possible combination of feature values

What outputs 8/9 Complete Direct

The participants agreed that the decision tree approximation is able to completely and directly answer to the what output question setting current expectation of the possible results that can be provided by the system

What data | 3/9 | Complete | Direct

Three to nine partecipants claimed that the decision tree approximation is able to provide hints for the what data question, especially because of the transparent communication of the feature and the corresponding values that compose the tree branches.

Conclusions

Question type	Mentions	Efficiency	Effectiveness
How it works	9/9	Complete	Direct
How	9/9	Complete	Direct
How to be that	9/9	Complete	Direct
What if	9/9	Complete	Direct
Why	8/9	Complete	Direct
How to still be this	8/9	Complete	Direct
What outputs	8/9	Complete	Direct
Why not	7/9	Complete	Direct
What data	3/9	Complete	Direct

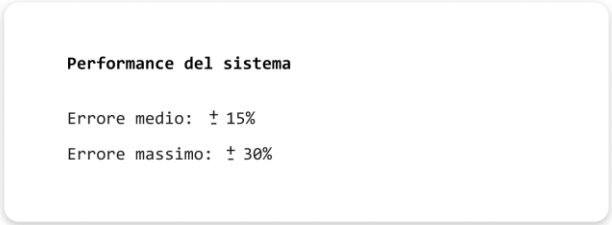
The decision tree approximation explanatory form has been evaluated as able to answer to most of the possible question that users may have in mind while interacting with an AI based decision support system. All participants found it able to answer to an **How it works, How (under what condition), How to be that, and what if** questions, whether

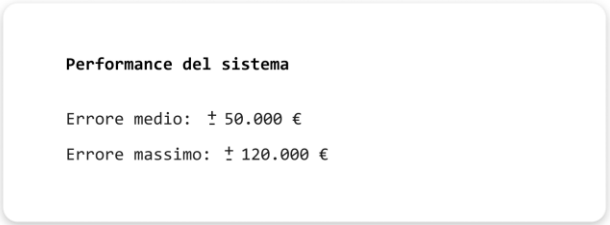
most of the found in it answers to a **Why, How to still be this, What outputs and Why not** questions. Only few participants mentioned even answers to a what data question.

Our experiment confirms previous findings for what concern the **Why, Why not What if and How it works** questions as well as enriches the literature about the decision tree approximation explanatory form: based on our findings it may provide even great hints to answer to **How to be that, How , How to still be this,** and **what outputs** questions. Lastly, may provide answer to a **what data** question.

Performance:

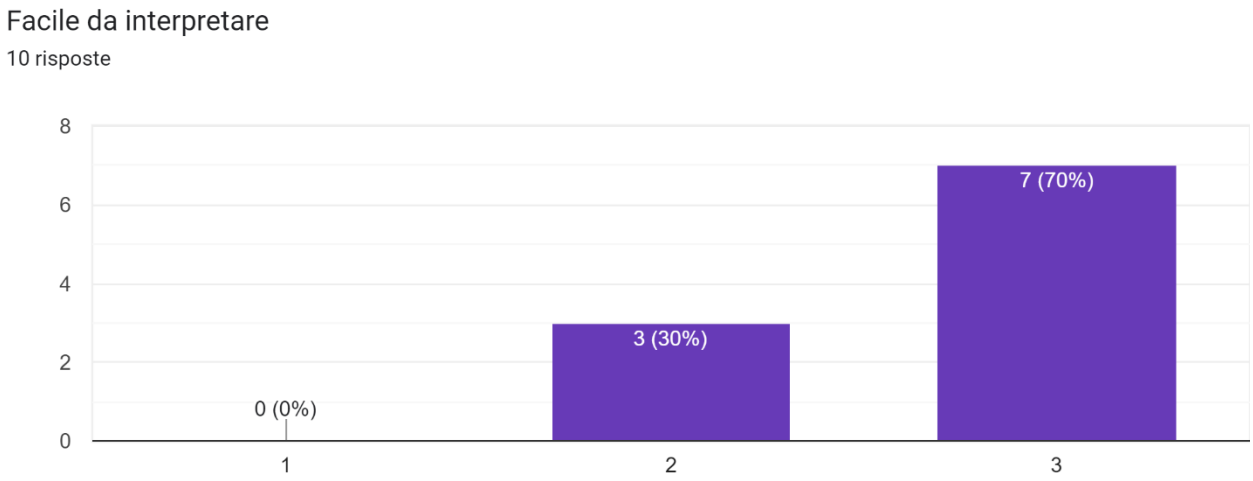
Performance metrics, like accuracy, confusion matrix, ROC curve, and mean squared error, provide a general understanding of the quality of a model's decisions, and allow users to have realistic expectations of the model's abilities. They give a broad picture of the model's overall performance.
In our experiment the performance explanatory form showed the system performance in terms of average error and maximum error.





Understandability

User friendliness and perceived usefulness

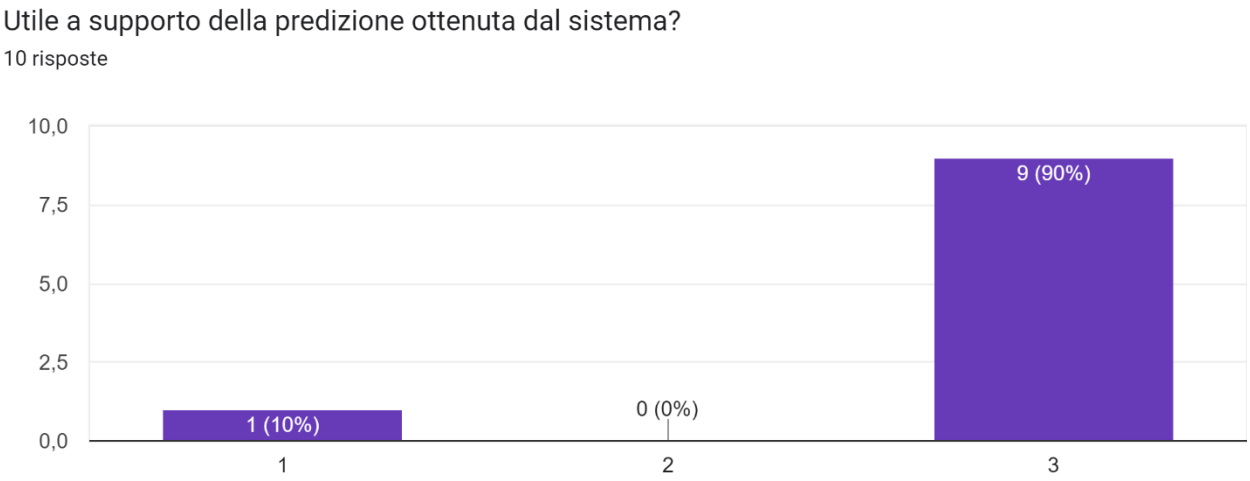


The decision tree explanatory form has been evaluated as easy to understand by 70% of participants (7/10), the three remaining has some difficulties.

Participants were able to extract from the explanatory form visualisation information on the reliability and accuracy of the system predictions They understood that the system can make errors with a determined average and a maximum value.

Some participants found challenging to interpret the difference between the average and maximum error, implying that this definition can be out from vocabulary domain of less literature people. One participant claimed: ”perhaps an error graph (box plot type) would be easier and to also understand how much the error affects the estimate, the perception of the error on the estimate could be offset from the value because 50,000 may seem a lot”. As suggested an improving in terms of easiness to understand could be use a graphical representation and provide clarification on the scale of the error. This has been confirmed by another participant: “Better percentages of more or less”. Since participants also expressed confusion about the meaning of average and maximum error add a tooltip with a definition could be a nice to have.

Applicable context of use:



Most of participants has found the information extracted useful as a support the predicted outcome of the decision support system (9/10)
In particular they found the performance for useful to:

- understand the reliability of the AI decision support system thus providing a sense of security about the system.
- Overall, understand how much they can trust the AI.

The only one participant which find the explanatory form as not useful provided hints about the moment of interaction in which it would be better to show it, as he suggested, it can be considered a marginal information about the overall system and, since it's not referring to the given prediction it may be provided in a side page at the beginning of the system interaction. One other participant even though has declared the explanatory form as supportive and useful expresses the concern that might be useless sometimes, especially when the error rate is low, because may lead to an unjustified under trust.

Explanation Usability

Literature claims

In the literature about question driven design approaches the performance explanatory form has been mentioned only in the UXAI work which claimed that is able to answer to **a why, why not and how confident** questions

User study results

All the ten participants found this explanatory form able to answer to at least one of the protypical questions analised.

How confident **10/10** **Complete** **Direct**

The partecipants all agreed on the ability of the performance explanatory form to answer the how confident question mostly completely and directly.

Conclusions

Question type	Mentions	Efficiency	Effectiveness
How confident	10/10	Complete	Direct

The explanatory form has been recognised as greatly able to provide answer to the How confident question.

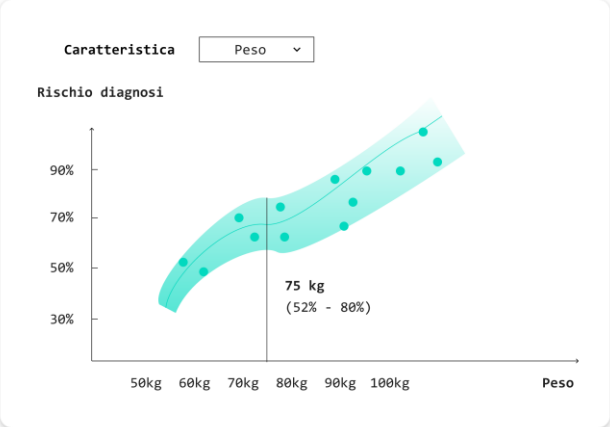
Our results confirm the literature findings about the ability of the system performance explanatory form to answer to an How confident question. On the contrary its ability to provide hints to answers to even ‘why’ and ‘why not’ questions did not obtain confirmation in the results of our experiment.

Feature shape

The feature shape explanatory form displays the connection between a specific feature and its result. For example, it could demonstrate the relationship between the size of a house and its expected price. This form is typically represented through either a line graph (for continuous features) or a bar graph (for categorical features) to visualize if the relationship between the outcome and feature is simple or more complicated, such as being linear or monotonic.

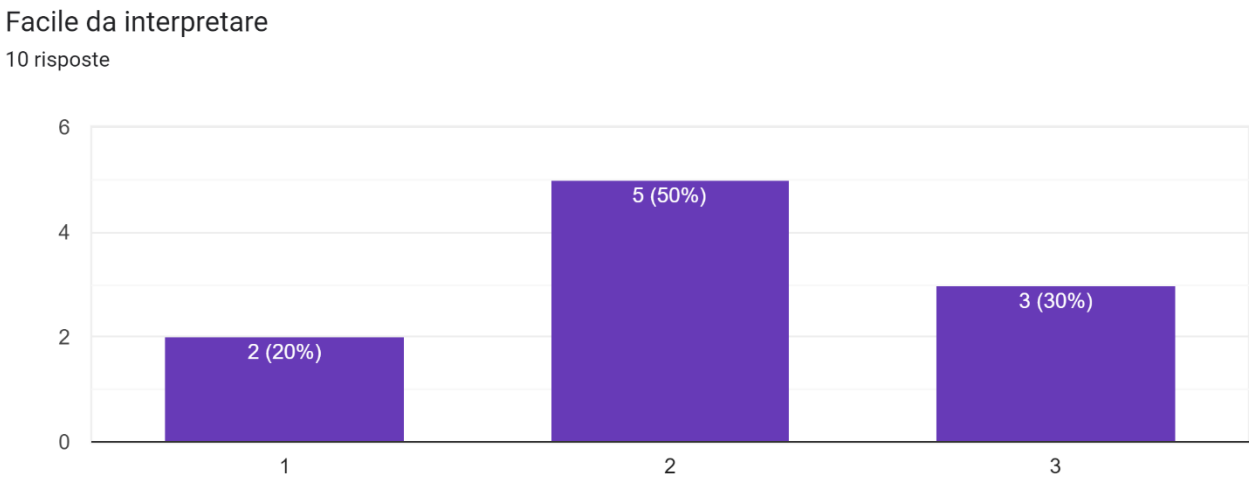
In our experiment the feature shape explanatory form showed a combination of the two common visualisation techniques: the line plot highlighting the average of the outcomes of instances with the feature value and a scatter plot highlighting the values of particular instances. A button with the standard visualisation of a drop-down list was added on the

top of the visualization, to let users understand the interactivity nature of the explanatory form, due the user possibility to change the feature of reference. Additionally, the max of the area covered by the graph-values along the two axes has been highlighted with a linear gradient, to have a glimpse on the variation of the possible results according to the feature value: in particular, the colour gradient intensity varied accordingly to the predicted outcome frequency.



Understandability

User friendliness and perceived usefulness



Participants has a different perception about the easiness of understanding the feature shape. Only three out of ten participants have evaluated it as easy, all the others has difficulties in understanding what the line plot signified (one participant completely

misunderstood and commented that the line was the representation of his input values) and most of them were confused about what the scatter plot points signified. After a while mostly of the participants agreed on the good amount of information that the explanatory form can convey but, at the same time, this abundance of hints made them cognitive overloaded lead them to consider as complex the overall evaluation about the form interpretability.

In general, participants were able to extract information regarding the correlation between feature values and predicted outcome from the feature shape explanatory form visualization.

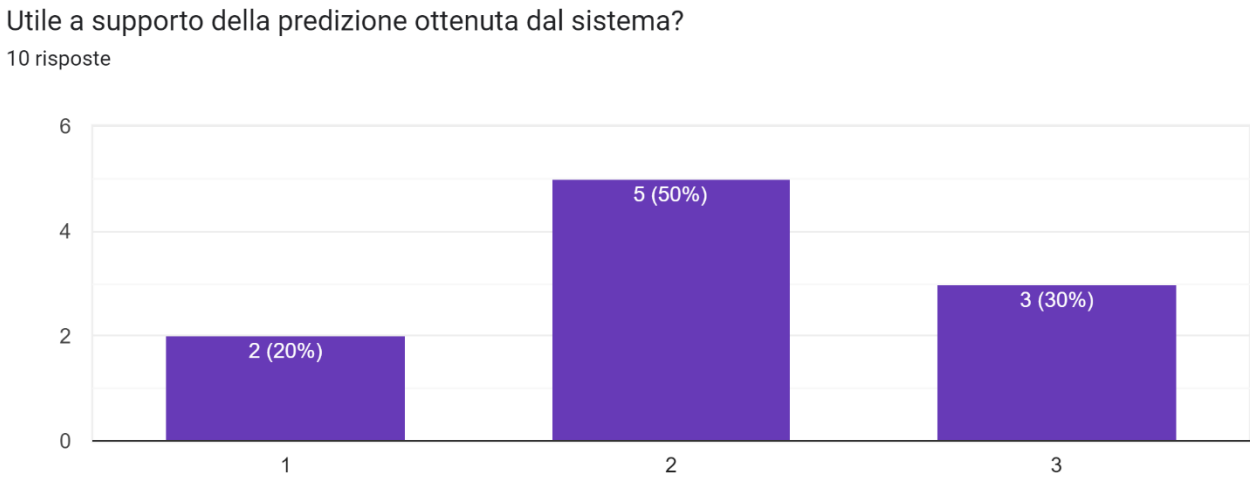
Even if some participants found the visualization to be difficult to interpret and not very intuitive, others appreciated the ability to see the variability of different features and compare the predicted outcomes. The visualization was seen as useful in understanding the feature relevance and the general trend and variability of various features by using it, but it has been highlighted the difficulty to interpret it rigorously making the forn lose the benefit of being considered as potentially complete in terms of information conveyed.

Overall, the visualization was seen as a tool for changing features and getting a better understanding of their variations and impact affecting the overall result, but not very useful for simple explanations.

For what concern the evaluation of the factor of misapplying, the correlation between the data showed in the form and the system dataset has not been clearly identified by the half of participants: not understanding the proper meaning of the line plot and the points of the scatter plot. In this direction may be useful for improving user understanding had additional information displayed through an hover interaction on the form elements.

It’s worth to be noted that anyone between participants commented the gradient-coloured area.

Applicable context of use



The explanatory form has been considered as useful to support the system outcome by 3/10 partecipants. Due to the difficulties in the interpretation and understandability the explanatory form the others participant expresses some concerns about the perceived usefulness, especially because it requires a lot of concentration and time to be understood. Another factor that affects the perceives usefulness is the relation with the data with which the system was trained on, considered as not always relevant to match users interests by one participant.

Given some preliminary knowledge in data and statistics the explanatory for became perceived as more useful and complete: as one participant has stated:
“If one knows about data analysis, one will find this explanation very comprehensive. One can see for each characteristic how much the data is estimated. Anyway, the values are clearly displayed.” Anyway, most of the participants agreed that the form clarifies how each parameter affects the prediction, even If in not a satisfactorily clear and direct way. At the same time, one participant claimed that seeing only one feature at the time is useless and suggested to increase the usefulness by showing all the features together in the same graph.

Explanation Usability

Literature claims

This explanatory form has not been previously studied in the field of question seeking in human XAI-interaction.

User study results

Eight out of 10 participants found this explanatory form able to answer to at least one of the protypical questions analised. This data directly correlate with the found difficulties in interpret it.

How it works 4/8 Partial Indirect

Four interviewers up to 10 suggested that the explanatory form helps users to answer partially to the question how it works by allowing to see all the characteristics relevant for the calculation. Has been underlined that a complete understanding of the logic may not be possible since the weight of each characteristic. is not given. Another factor that makes the question answered not exhaustively is the fragmented nature of the information conveyed since they regard only one feature at the time.

How 6/8 Partial Indirect

Six out of ten participants have found the feature shape explanatory form partially helpful in answering the "how (under what conditions)" question. As a participant explained that's because: "I look at the 80% line [in the output axe of the graph], for each characteristic, the range of values it refers to but I struggle to correlate the various characteristics". This difficulty in correlating the various feature explains even why other participants considered this process as unintuitive and time spending claiming that the answer is provided in an indirect wat. To truly answer they would like to see multiple variables and a more complete picture at the same time.

Why 4/8 Partial Indirect

Four out of ten participants claimed that the feature shape explanatory form is able to answer to the why question, but only partially and indirectly. This is done through the ranges provided for each feature, if the users value fit on it, then the answer is considered as given. Anyway, participants mentioned that the 'decomposed' nature of the information provided doesn't allow them to have a complete, and thus, satisfying, answer.

Why not 4/8 Partial Indirect

As for the why question four to ten participants claimed that the feature shape explanatory form is able to answer to the why not question, again only partially and indirectly. The process to answer it is done with the same information as before: for the intended result the corresponding range are searched for and compared to the input values. Even in this case the 'decomposed' nature of the information provided doesn't allow them to have a complete, and thus, satisfying, answer.

How to be that 6/8 Partial Indirect

Six out of ten participants found the feature shape explanatory form able to give partial and indirect hints for answering to "how to be that" question. As the users mention factors that concur to not have a complete answer are the fact that they only have one feature at a time and therefore since they are unsure of the weight of each feature in predicting the outcome they can only suppose it. Thus, they recognize trial and error as a method for understanding how to achieve a desired result, but again they expressed the concern of not being sure if the relationship between feature is linear.

How to still be this 4/8 Partial Indirect

As for the "how to be that" question only 4 participants up to 10 found a way to extract information to provide an "how to still be this" answer. Even in this case the feature shape explanatory form is able to answer partially and indirectly as the process has been described as complicated and time-consuming. The reasons are again the ineffective way to show only one feature at the time and the lack of knowledge about which feature weighs more or weights less in the system calculation.

What if 4/8 Partial Indirect

Again, only 4 participants up to 10 declared that the feature shape explanatory form is able to provide an answer to a 'what if' question partially and indirectly. The process to achieve it is recognized as time-consuming because only allow you to see one feature at the time without knowing which feature weighs more or less in the system calculation.

How confident 3/8 Partial Indirect

Three participants out of eight found partial and indirect information to answer to an ‘how confident’ question. This is done by providing the range of predicted outcome for each value feature, suggesting the ‘not absolute’ nature of the prediction given by the system and by checking the average value of each feature. This information, as one participant claimed, ‘make me understand that there is a reasoning behind the functioning of the system’. However, as a factor not given to fully answer the question, has been mentioned the unclear correlation between the predicted values and what has been happened ‘in the real world’.

What outputs 8/8 Complete Direct

The question ‘what output’ is considered as answered by 8 to 8 participants mainly in a complete and direct way thanks to the range of value thaty can see in the Y-axe in the graph. However some participants claimed that the answer can be only partial again due to the fact that is related to one feature at the time.

What data 6/8 Partial Indirect

The feature shape provides answers to the "what data" question, to six participants up to ten, but not always in a clear and concise manner. For some participants the answer is given by the list of characteristic considered, for others, the points with the instances in the dataset provides elements to understand the data. However, some participants were insicure about their answer at the point that one claimed that he can try to answer the question by ‘pure hypothesis’, not trusting at all his capacity to understand the information provided by the form.

Conclusion

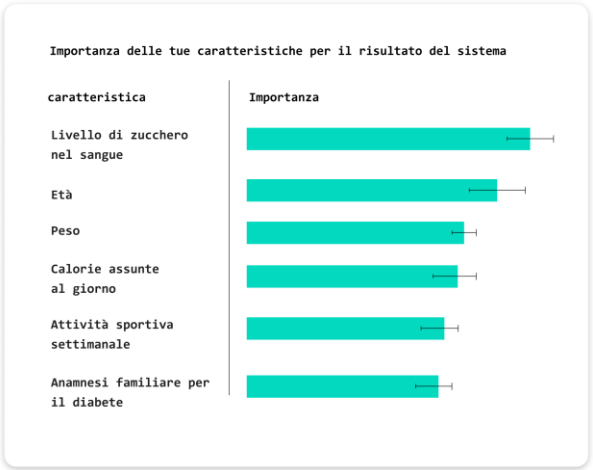
Question type	Mentions	Efficiency	Effectiveness
What outputs	8/8	Complete	Direct
How	6/8	Partial	Indirect
How to be that	6/8	Partial	Indirect
What data	6/8	Partial	Indirect
How it works	4/8	Partial	Indirect
Why	4/8	Partial	Indirect
Why not	4/8	Partial	Indirect
How to still be this	4/8	Partial	Indirect
What if	4/8	Partial	Indirect
How confident	3/8	Partial	Indirect

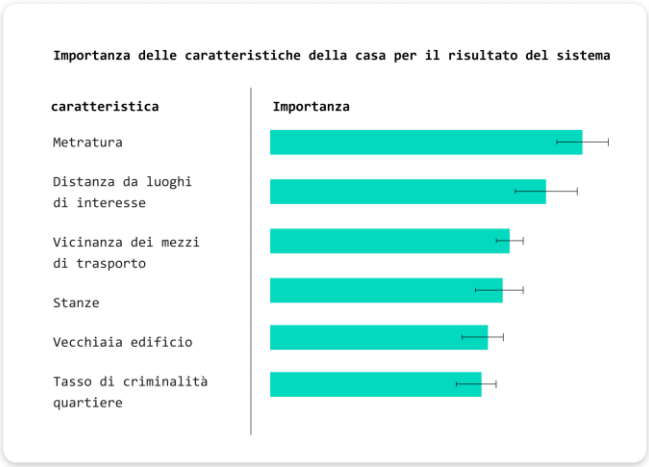
The feature shape explanatory form has resulted as able to answer to all the question proposed by the question driven design approach for explainable AI: Listed in order of participants mentions: What outputs, How (under what conditions)/How to be that/what data, How it works/why/why not/How to still be this/what if/How confident. Not all of the answer that the form may provide has the same level of efficacy and effectiveness as a matter of fact, due to the difficulties mentioned in the previous analysis, most of them has been recognised by participants as partially given. As a consequence, the explanatory form results as characterised by a great flexibility in answering to question user may have in mind while interacting with AI based decision support system but, for what concern the overall usability, there's still room for improvement.

Global feature importance

The global feature importance explanatory form is a visual representation of the weights assigned to various features used by the model to make predictions. It highlights the key features and their importance scores in determining the outcome. The form provides an explanation of how the prediction is made assuming that is done through the linear combination of important features. It allows for a clear understanding of which features are important for the decision and what are their attributions to the prediction.

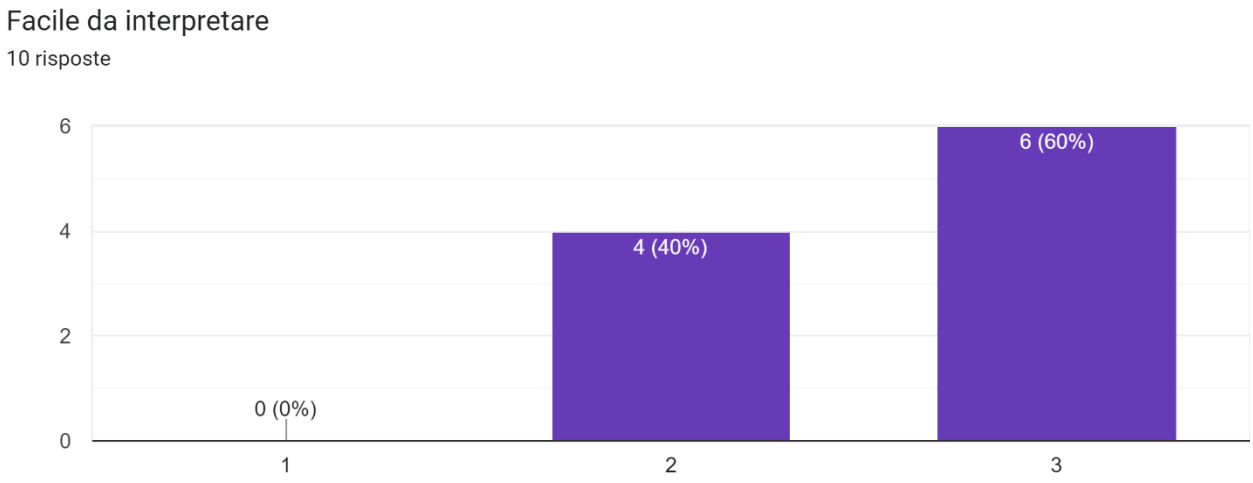
In our experiment the global feature important explanatory form was composed by a horizontal bar chart listing all the features that has an impact on the system prediction, ordered by their average importance. Additionally, according to standards in box plot visualisation on the average value has been added the graphic sign called whisker to indicate the dispersion (also called variability, scatter, or spread) ending in the maximum and minimal value that the feature importance score may have in the prediction.





Understandability

User friendliness and perceived usefulness

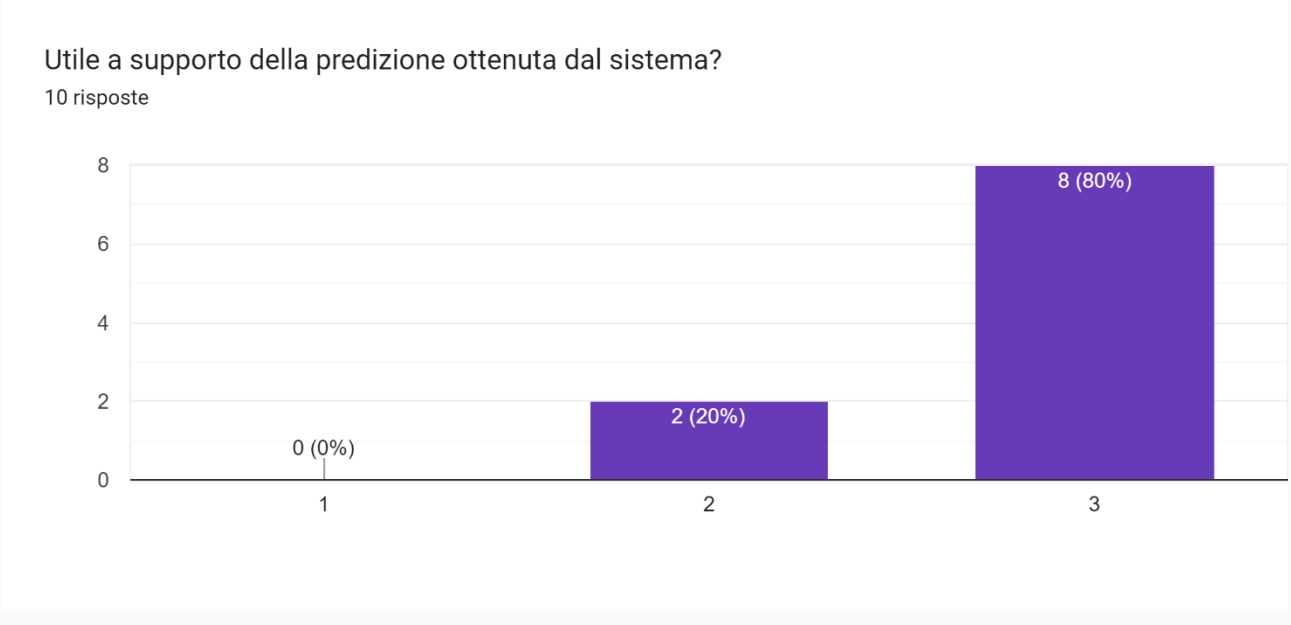


Overall Partecipants found the global feature importance easy to be interpreted, even though less than the half of partecipants has some difficulties in understanding the whisker sign on the graph consisting in half the information conveyed by it.

The participants in the interviews were able to extract information about the importance of the various features calculated to provide the prediction in each scenario. They understood that the longer the bar in the visualisation, the more important the feature is. Some of them noted that the black line indicates the variability of the feature weight in the system performance, but for others this was an element that made them doubt about their ability to correct interpret the form. Overall, they appreciated that the visualization helped them understand which features have the most impact on the final result.

Overall, the main factor of misapplying has been identified in the whiskers on the graph visualisation, but this hasn't affected the overall perception in terms of perceived usefulness. Anyway, since the sign convey important information in terms of variability of the importance that a feature could have, that allow users to better understand how the system works, an effort in make it more user friendly should be done. In this sense a suggestion for improvement may be to add the maximum and minimum score at the end of the sign while you hover it with your mouse.

Applicable context of use



The global feature improrttance explanatory form has been evaluated as a good way to support the prediction provided by the system in the following context of use:

- Understanding the most influential features that affect the estimate and understand the corrective actions to take in daily life based on factors like sugar level, calorie
- Participants believe the form is useful because they know the hierarchy of weight that the features have and thus understand the evaluation better.

Explanation Usability

Literature claims

In the literature about question driven design approaches the model confidence explanatory form as been mentioned both in the Liao and al.contribution and in the UXAI work. Bot of them consider it able to answer to the **How it works and How (under what condition)** questions

User study results

All the experiment participants found this explanatory form able to answer to at least one of the protypical questions analised.

How it works 10/10 Complete Direct

All the participants agreed on the capability of the global feature importance explanatory form to answer directly to the how it works question. This is perceived thanks to providing the average weight of each feature. That information let participants be aware that different feature may have thus understanding different impact on the prediction result. Half of them claimed that the explanatory form conveys enough information to have a complete answer to the question whether three of them considered it as partial.

How 4/10 Partial Direct

Four to ten participants found, thanks to the explanatory form, an answer to the How (under what conditions) question, but only in a partial way. They explained they were able to do some hypothesis in answering this question based on what they knew about the value of the feature of their input. This result can even have been conditioned by some personal bias due to previous knowledge in the scenario topic: as one participant claimed I'm able to understand how I get this result because I'm overweight [diabetes scenario – input value].

How to be that 3/10 Partial Indirect

Participants have mixed opinions about the capability of the global feature importance to provide an answer to the how to be that question. Overall, only three to then of them found some hints to answer it making the information extracted for this purpose partial and indirect. As one participant claimed: "It only tells you which are most important or not to change the risk diagnosis, but not precisely how to be that, still be this, change the result", so the only answer you may get is "probably by decreasing the characteristics with high importance".

What data 3/10 Partial Indirect

Three to ten participants found some information about the 'what data' question because they claimed they has been provided with the feature list. One of them recognized that this has no added value since 'but I had them even before' referring to the input information provided with the prediction at the beginning of the experiment.

Conclusions

Question type	Mentions	Efficiency	Effectiveness
How it works	10/10	Complete	Direct
How	4/10	Partial	Direct
How to be that	3/10	Partial	Indirect
What data	3/10	Partial	Indirect

The global feature importance explanatory form as been evaluated as completely able to answer to the **How it work** question by all the participants. Even though significatevely less in terms of number of mentions the other question answered by the form resulted the **How (under what conditions), How to be that and What data.**

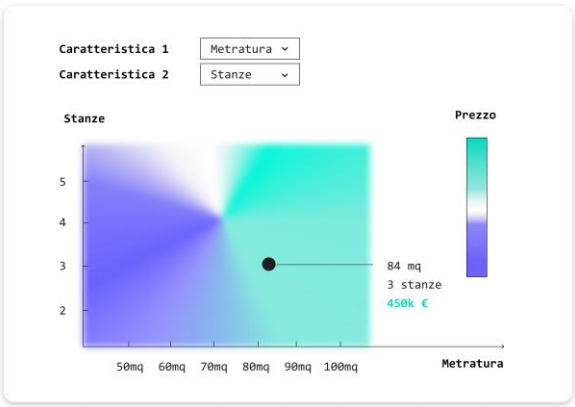
Compared to the literature our results confirm the ability of the form to directly provide answer to the How it works question. On the contrary, our findings don't confirm a great link between the How (under what condition) question and the form as UXAI has suggested. In addition, our study enriches the pull of question answered somehow by the form with the How to be that and what data one.

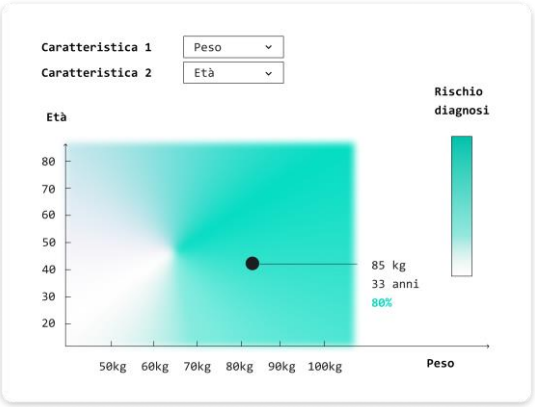
Feature interaction

The Feature Interaction Explanatory Form considers the non-linear effects that occur when features interact with each other on the outcome. This form extends traditional feature analysis by considering the combined effect of two or more features on the outcome, rather than simply examining the individual impact of each feature. By considering the interactions between features, this form provides a more complete understanding of the relationships between features and the outcome. Thus, the explanatory form can be considered as an extension to the feature shape one.

In our experiment the Feature interaction explanatory form presented a 2-axis graph (representing the two characteristics main character of the analysis and their scale of

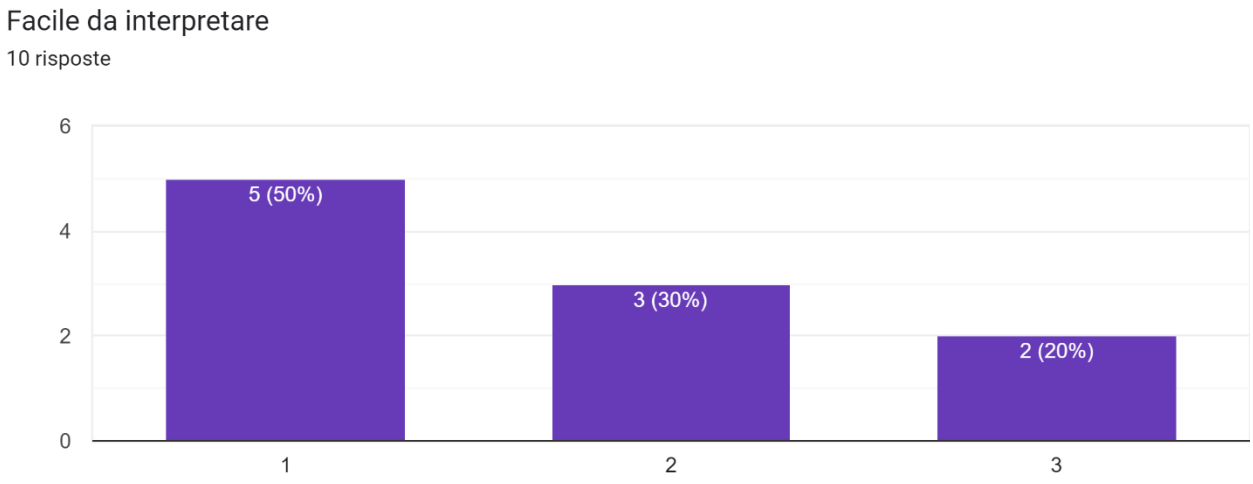
values). The predicted outcome range has been graphically represented through a gradient coloured scale, illustrated near the graph and used to fill the graph area. The input features and the corresponding outcome has been explicated with a text and highlighted graphically with a point in the graph. To suggest that the analysis can involve all the features considered by the AI based decision support system, the two characteristic considered were displayed in a button designed following the standard visualisation of a drop-down list and thus highlighting the interactivity nature of the explanatory form.





Understandability

User friendliness and perceived usefulness



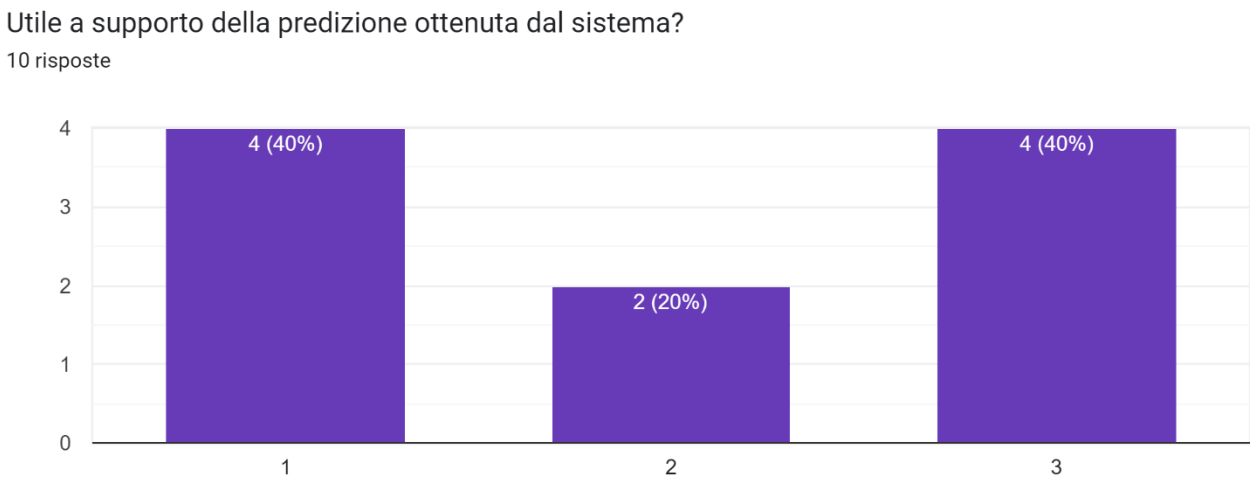
Overall the feature interaction explanatory form has resulted as hard to be interpreted by the experiment participant. As a matter of fact only 2 up to 10 has no difficulties to retrieve information from it, all the others has some difficulties and half of them has directly expressed their hardness-to-understand it.

As previously mentioned, in general, the participants had mixed opinions about the feature interaction explanatory form: some participants found it difficult to understand especially in terms of cognitive effort and the time needed to interpret the information presented. Despite these difficulties, overall, the participants was able to understand that the feature interaction explanatory form shows the relationship between different features and their impact on the outcome. Participants understood the relation between the colour gradient and the predicted outcome; they recognized that the dot represented the input instance and the possibility to change the feature taken into consideration to generate it. As mentioned by one participant the explanatory form' correlates two parameters together so it makes me realise that there are correlations between characteristics that I may not have known before'.

For what concern factors of misapplying analysis and design recommendations, overall, the participants had differing opinions about the effectiveness of the form in presenting the information and suggested different ways in which it could be improved. Most of the participants found hard to understand the precise value of the prediction at a first glimpse due to the use of the colour gradient, described as not able to convey precise information because of the difficulty to separate the different shades of colour. At this regard some of them suggested to use more colours or to differentiate the different value through some geometrical shapes like squares or rows.

Additionally, since it relates three variables at the same time some participants suggested that the form could be improved by using 3D visualization. Overall was mentioned that the explanatory form is quite overwhelming in term of information provided and would be better to improve the visual hierarchy of different elements to better highlight the most important information. However, others felt that the form was too confusing and that a simpler format, such as a line graph, would be more effective in showing the relationship between features. Lastly, some claimed lack of the precise values on the prediction colour gradient scale.

Applicable context of use



Since the difficulties in interpreting and extracting information the perceived usefulness of the form doesn't present so many results. Anyway, here some possible context of use identified by participants:
Since the form correlates three variable at the same time it has been recognized the usefulness in evaluating multiple parameters simultaneously.
At this regard, one of the participants claimed: 'May be useful if I held one characteristic constant and saw how the other changes, I would understand how the risk diagnosis changes' suggesting the possibility to understand better how the result may change variating one characteristic with respect to one another. Or, as another suggested: 'if I identify two characteristics, I have the power to change I can use this explanatory form to understand how to optimize my efforts'. Additionally, the form may arise awareness about the interlink of different features in the outcome calculation because: 'Make me understand that there are correlations between characteristics that I may not have known before.'

Explanation Usability

Literature claims

The feature interaction explanatory form hasn't never taken into consideration in previous literature on a question driven design approach for explainable AI

User study results

Due to the analysed difficulties in interpreting the form only 7 out of 10 participnats found this explanatory form able to answer to at least one of the protypical questions analised.

How 5/7 Partial Direct

The feature interaction explanatory form has been recognized as able to answer to the How (under what condition) question by 5 out of 7 respondents. All of them thought that the question may be answered only partially and indirectly because by considering only two features at the time they lose the overall vision, even if they recognized that interacting with the form, changing the features considered they would be able to have it back. For another participant the main obstacle to completely answer the question is the lack of a precise range of value for the output scale.

Why 3/7 Partial Direct

Only 3 to 7 participants found the feature interaction explanatory form able to answer to a why question, and only partially. They claim that they can get an explanation due to the the relation of feature and outcomes but only relying on two characteristics at the time.

Why not 3/7 Partial Direct

As for the why question only 3 to 7 participants recognized the explanatory form able to answer partially to a why not question. Even in this case only relying on two characteristics at the time make the question answered only partially.

How to be that 5/7 Partial Direct

Five out of seven participants suggested that the feature interaction explanatory form can provide some level of understanding about how to achieve a certain result, based on two features in average, but not necessarily for all other relationships. Thus, It provides partial and indirect answers, with the potential to understand more, but the confusion caused by the complexity of the information makes it difficult. The users can see how

their possible actions impact the results and make changes to achieve a desired outcome. However, there is still some uncertainty and a need for further clarification.

What if 4/7 Complete Direct

Four out of seven participants claimed that the feature interaction explanatory form provides them with the information they need to answer "what if" questions. This is achieved by allowing the users to condition their analysis on two or more features, giving them all the information, they wanted, helping them understand relationships by the colour change in the graph, and allowing them to see the results of possible changes by moving around in the graph. Anyway, the process to get this information make the answer provided only indirectly and the piece of information highlighted by participants were considered enough to fully answer the question only by half of them.

What outputs 3/7 Complete Indirect

Three out of ten participants claimed about the capability of the explanatory form to answer at the ‘what outputs’ question. Anyway, as one participant highlighted: ‘the amount of result could be more than the reality because I think this kind of analysis consider all the possible feature combinations even if they are not realistic [such as being height 170 and weight 40 kgs].

What data 3/7 Complete Direct

Three out of seven participants recognized the feature interaction explanatory form able to answer at the ‘what data’ question. For them, this has been provided by the samples that can be found moving around the graph and because of the list of all the possible features taken into consideration to provide the outcome.

Conclusions

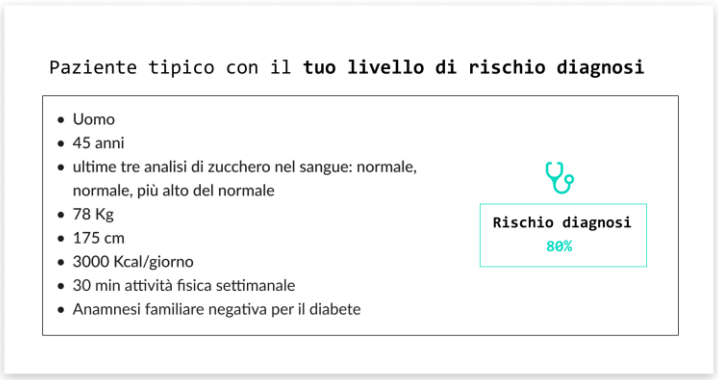
Question type	Mentions	Efficiency	Effectiveness
How	5/7	Partial	Direct
How to be that	5/7	Partial	Direct
What if	4/7	Complete	Direct
Why	3/7	Partial	Direct
Why not	3/7	Partial	Direct
What outputs	3/7	Complete	Indirect
What data	3/7	Complete	Direct

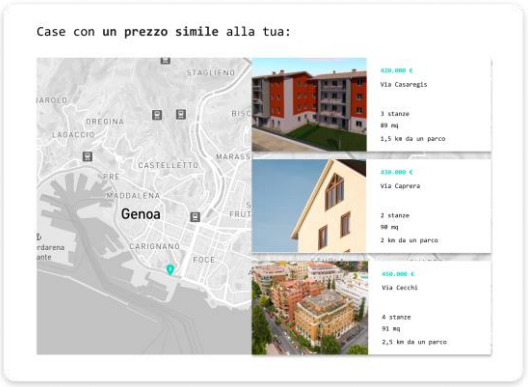
The feature interaction explanatory form results to answer to Seven out of ten question found in the question driven design approach for explainable AI, but the ones most mentioned by experiments participants barely reached half of the total respondents: ‘How’ local and ‘How to be that’. The smaller number of mentions resulted by the experiment are probably due to the difficulties in understanding the form mentioned in the previous analysis. Following the number of mentions ranking the experiment results suggest the form as able to provide some hints to answer even to What if, why, why not, what outputs and what data questions.

Typical example

The typical or prototypical example explanatory form describes a representative instance, for a precise prediction. Thus, the example communicates to the user the typical features that are used to get the outcome they got from the AI based decision support system. By providing similar examples with the same record as the prediction, the typical example explanatory form helps to reinforce and clarify the prediction made.

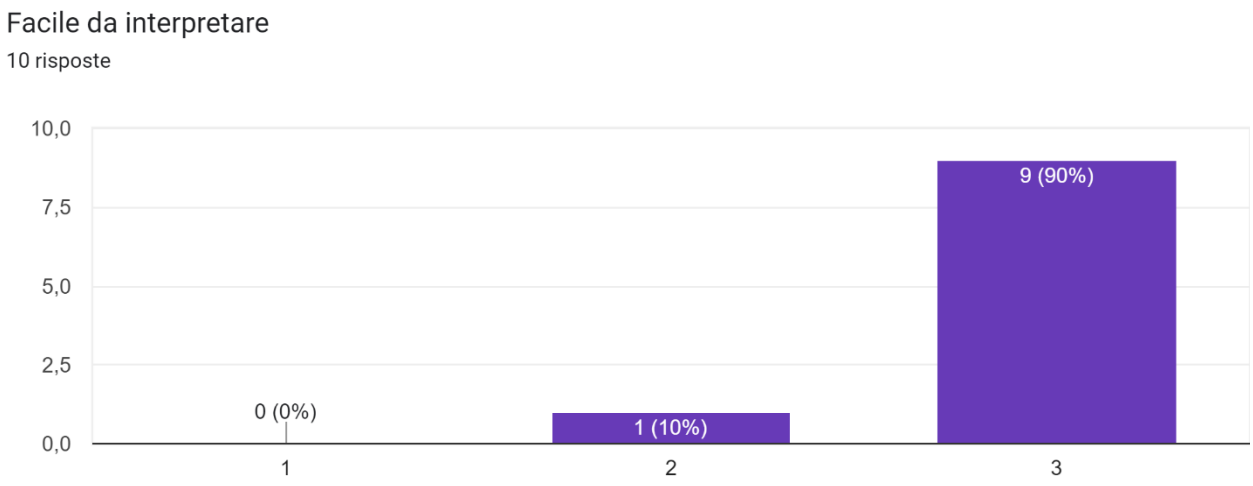
In our experiment the prototypical example cards presented one patient with the same diagnosis of the input in the diabetes risk scenario and a sample of three houses with a similar predicted price in the house cost estimation scenario. In both cases, the card displayed the list of the feature of the samples and the corresponding values.





Understandability

User friendliness and perceived usefulness



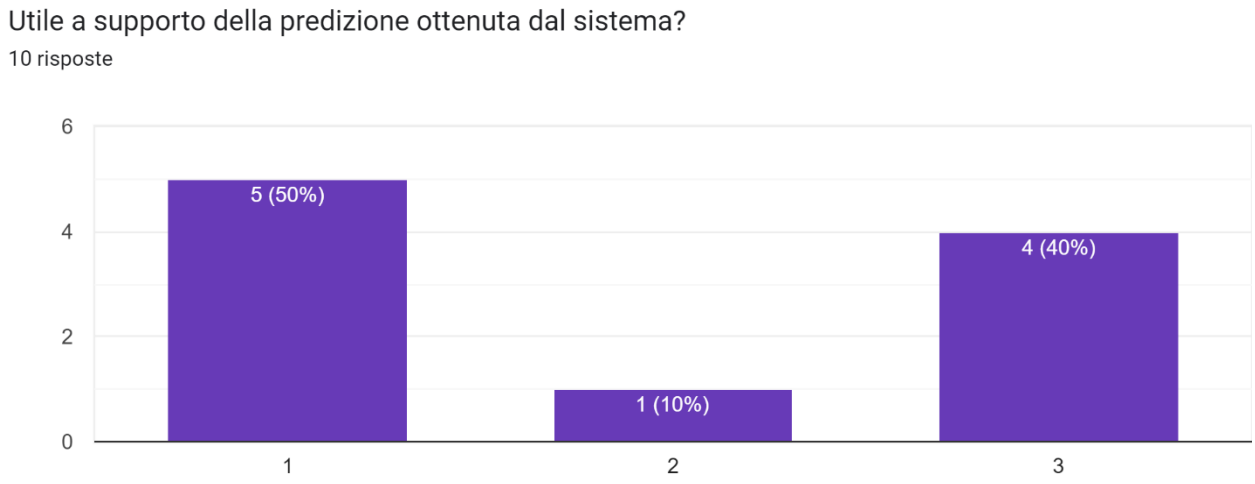
Overall, the prototypical example explanatory form has been evaluated as easy to be interpreted (9/10) by the experiment participants.

The participants agreed that the information extracted from the explanatory form allow them to compare the given example with their own feature and thus the prototypical example has been recognized as able to put the participant prediction in context. The

comparison performed has been said as able to let them understand the correlation between different features and the outcome (‘ it makes me realise that even if the features are different I could get the same price’) and how small differences in features can impact it (‘I can therefore understand how much the parameters can vary to get the range’) as well as the typical values and the distribution of the features in the context of their precise prediction.

Even if the explanatory form has been considered easy to be interpreted the analysis of the interviews led us to underline some criticalities that can lead to a misapplication of the information extracted. First of all not all the participants understood the concept of the prototypical example confusing it with a similar example. One participant claimed: ‘this is the perfect example with which the machine has been trained on’.
For the diabetes risk scenario one participant claimed that the ‘lacking in labelling the features as 'good and bad'' in terms of affecting the outcome make for him impossible do the comparison with their own values and this led them to consider the explanatory form as useless.

Applicable context of use:



Even if the typical example explanatory form has been considered as easy to be interpreted and let participants to easily compare their own feature with them, the information extracted by interacting with it wasn’t enough to perceive it as useful for half the participants; at the same time 4 out of 10 claimed on the contrary and here there’s a summarisation on the context of use in which can help to support the prediction given:

-‘Only if the characteristics of the example are the same of the mine ones, it can help me to understand better the prediction’

- ‘The explanatory for can be useful in hypothesizing how the AI system take decisions’ thanks to the comparison and thus let the user agree or disagree based on shared basis. In this sense, has been mentioned even that can help to perceive the system accuracy.
- As mentioned in the previous analysis by comparing the example feature with yours can be useful to understand how slightly different characteristics (may not) affect the predicted outcome and for explaining the importance of them in the system computation.

Explanation Usability

Literature claims

In the literature about question driven design approaches the typical example explanatory form has been mentioned firstly, in the work by Liao at al., as able to answer mainy to a Why question, but even able to provide some hints for an ’How to still be this’ question. Lately, the UXAI toolkit has added ‘why not’ to the list of questions answered by the explanatory form.

User study results

Seven out of ten participants found this explanatory form able to answer to at least one of the protypical questions analised. In this case ther reason behind this result has to be found not in the hardness to interpret the form but in the lack of perceived usefulness.

How it works ● **4/7** ● **Partial** ● **Indirect**

Four out of seven respondents recognized the prototypical example explanatory form as able to partially answer the how it works question by providing information about the parameters it is based on. Said that has been mentioned that does not explain the general logic behind it, only offering comparisons.

How ● **3/7** ● **Partial** ● **Indirect**

Three out of seven participants claimed that the prototypical example provide partial hints to answer to a how (under what condition) question: the partiality of the answer is due to the possibility to understand the conditions only of the given outcome and not for all the ones that may be predicted by the system. In particular one participant explained that the how it works question in answered letting him understand that the system predicts outcomes by correlating different parameters and that different values can provide the same prediction; one another underlined the possibility to use the given examples to approximate the reason behind the calculation.

Why 4/7 Partial Indirect

Four up to seven participants declared that the prototypical example helped them to answer partially to the 'why' question because provides the value on which the predicted outcome is based on. The key elements that help users to have an answer for the question include the ability to relate different parameters, compare, and have coherence between the input features and the other cases. Anyway, this process slower the answer achievement only letting it given as indirectly.

How to still be this 4/7 Complete Direct

Four out of seven respondents recognized the prototypical example as directly able to provide an answer to the how to still be this question: this can be done by analysing the samples provided and compare their feature values with the input ones, thus understanding ranges to maintain the predicted outcome.

What if 5/7 Partial Indirect

Half of the respondents (5/10) considered the prototypical example as able to answer to a what if question. Of course, only providing a restricted amount of samples that has to be analysed in terms of feature and compared to the input ones, that answer can be given only partially and indirectly.

What data 4/7 Partial Indirect

The ability of the prototypical example explanatory form to answer the question "What data?" is limited, as a matter of fact, has been recognized only but four out of ten

participants and described as partial and indirect. Anyway, the elements that suggest them answers has been the list of parameters and the fact that they considered the sample provided as part of the database.

Conclusions

Question type	Mentions	Efficiency	Effectiveness
What if	5/7	Partial	Indirect
How it works	4/7	Partial	Indirect
Why	4/7	Partial	Indirect
How to still be this	4/7	Complete	Direct
What data	4/7	Partial	Indirect
How	3/7	Partial	Indirect

Overall, the typical example explanatory form has resulted as able to answer to 6 out of the 10 question proposed by the question driven design approach even if mostly partially and no one of them has been mentioned by more than half of the experiment participants. This result validates the insights about the perceived usefulness of the form discussed before. Anyway, the question answered by the explanatory from in order of mentions are: What if, How it works, Why, How to still be this, What data and How local. The only two mentioned as mostly completely answered by the respondent were the How to still be this and the what data ones.

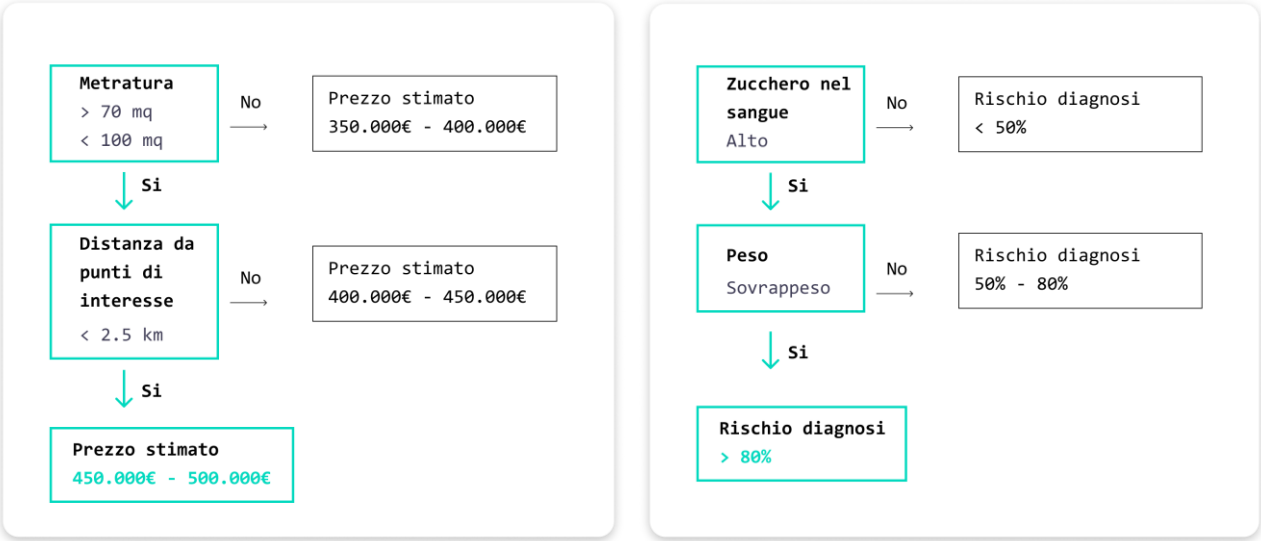
Compared to the literature findings our experiment confirms the ability of the typical example explanatory form to answer to the Why and How to still be this question, as the work from Liao et al suggested. Not the same confirm for the UXAI addition: only two of the experiment participants mentioned the ability of the form to answer partially and indirectly to why not question, a number not statistically significant and thus not considered in the experiment results. As a contribution to the literature, we can notice

that the list of question answered by the form, according to the experiment results counts new ones such as: what if, How it works, what data and How (under what condition)

Rule flowchart

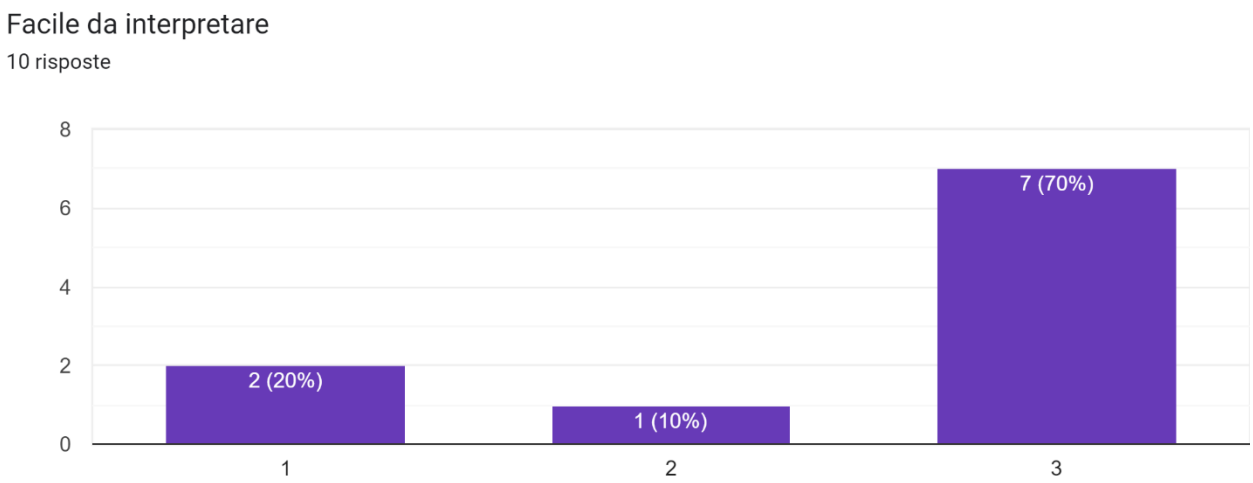
A rule flowchart explanatory form is a graphical representation of the decision-making process that an instance fits through in order to arrive at a prediction. The tree structure of the flowchart illustrates the branches of decisions that must be taken, with each branch leading to a specific outcome represented by a leaf. This form serves to clearly explain the rules and decision-tree path that the instance follows to guarantee its prediction.

In our experiment the rule flowchart explanatory form presented, as a sample, one floe composed by the analysis of two features value ranges. Each feature card presented two arrows (yes/no) and the corresponding prediction ranges.



Understandability

User friendliness and perceived usefulness



The participants to the experiment found the decision flowchart explanatory form mostly easy to be interpreted (7/10) but two out of ten had the opposite opinion.

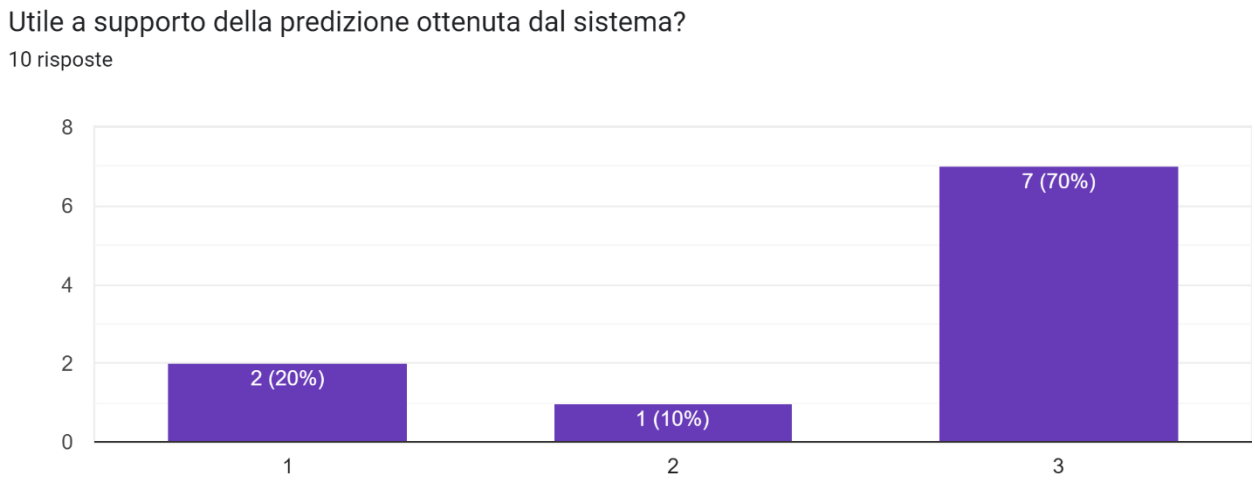
The participants in the interviews mostly agreed that all the information they were able to extract from the decision flowchart explanatory form was the same of the decision rules and decision trees. At this regard one participant expressed a preference in terms of straightforwardness: ‘is like the tree but it only takes one path and excluding the other, is more immediate. However, some participants found the flowchart difficult to interpret and preferred a decision tree visualisation because it was more intuitive for them. The comparison with the decision rules explanatory form has been deepened by another participant: ‘It is like the other one (decision rules) with a different readability because it is a diagram and not a text, for me better’

Anyway participants interpreted the explanatory form as a way to convey feature hierarchy and the possible prediction results according to ranges related to the input data. This helped them even to extract information about how the system works.

Lastly it’s worth to be mentioned on participant opinion about the accuracy of the information provided ‘seems to me a super simplification because some correlations are lost’

How already explained analysing the interpretation of the explanatory form by participants, the flowchart may not match user preferences in terms of textual/visual information compared to decision rules or decision tree explanatory form.

Applicable context of use



Most of participants (8/10) found the explanatory form useful to support the system outcome: here the main context of use mentioned:
• Correlating the various input data help in understanding the system logic, and in comparison to the decision tree, the rule flowchart has been recognized better in making the decision chain clearer.
• As previously mentioned the form explain the hierarchy among parameters allowing the user to make they're own consideration about possible changes
. •Lastly as one participant claimed 'even just the word "overweight" helps, because it gives more information on how the various parameters have been evaluated as good or bad based on your input values'.

Explanation Usability

Literature claims

In the literature about question driven design approaches the rule flowchart explanatory

form has been mentioned firstly, in the work by Liao at al., as able to answer mainly to a Why question, but even able to provide some hints for an 'How to still be this' question. Lately, the UXAI toolkit has added 'why not' to the list of questions answered by the explanatory form.

User study results

Eight out of 10 participants found this explanatory form able to answer to at least one of the protypical questions analised. This data directly correlate with the found difficulties in interpret it.

How it works 8/8 Complete Direct

All of the respondants believed that the rule flowchart explanatory form can provide answers to the question of "how it works" directly by taking parameters and understanding correlations, and by clearly showing the step-by-step decision-making process through all the input features. One participant claimed that the fact that the explanatory form doesn't deepen all the possibility but follows the flow according to the input feature is not enough to comprehend the overall logic, that's why they considered the how it works question only partially given.

How 8/8 Complete Direct

The analysis of the interview results suggests that the rule flowchart explanatory form helps most of users (8/8) to understand how (under what condition) predictions are made by providing information on the values that parameters must have, the conditions that must be met step-by-step, thus providing various outcome ranges.

Why 8/8 Complete Direct

All of participants found the rule flowchart explanatory form able to answer mostly completely and directly to the "why" question. The explanation is given highlighting the values of the relevant features that are evaluated in each steps allowing the user to compare with the one of them.

Why not 8/8 Complete Direct

All the respondents suggested that the rule flowchart explanatory form helped them understand completely why they did not receive a certain result by providing information about the feature values in each decision step, thus explaining the conditions that lead to the different results, allowing the user to directly check the correspondence with their own.

How to be that 8/8 Complete Direct

For what concern the how to be that question again all participants claimed that the rule flowchart explanatory form helps in understanding the steps to achieve a different result by following different paths. They appreciate the ability to understand the characteristics and potential changes, but prefer the hierarchical tree because it provides a more complete overview,

How to still be this 7/8 Complete Direct

Seven out of eight participants found the explanatory form able to answer directly to the 'how to still be this' question mainly following the arrow sequence.

What if 7/8 Complete Direct

Most of participants (7/8) found the explanatory form able to directly answer to the "what if" question mainly following different path in the flowchart.

What outputs 6/8 Complete Direct

Six out of ten participants suggested that the rule flowchart explanatory form can help to answer "what outputs" questions. Only one participant claimed it is able to only by providing partial results because of the value ranges.

Conclusions

Question type	Mentions	Efficiency	Effectiveness
How it works	8/8	Complete	Direct
How	8/8	Complete	Direct
Why	8/8	Complete	Direct
Why not	8/8	Complete	Direct
How to be that	8/8	Complete	Direct
How to still be this	7/8	Complete	Direct
What if	7/8	Complete	Direct
What outputs	6/8	Complete	Direct

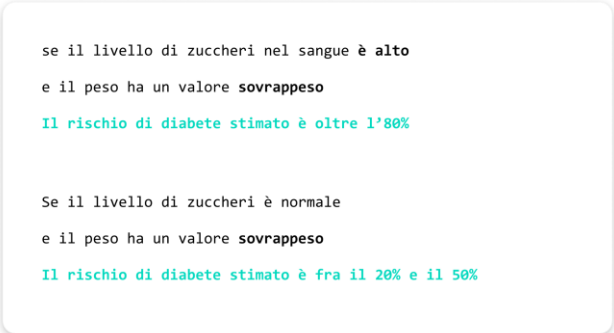
The rule flowchart explanatory form has been mentioned as able to answer to eight out of ten of the question proposed by the question driven design approach for explainable AI by most of the experiment participants in an almost complete way: How it works, How, Why, Why not, How to be that, but even how to still be this and what if, and lastly what outputs.

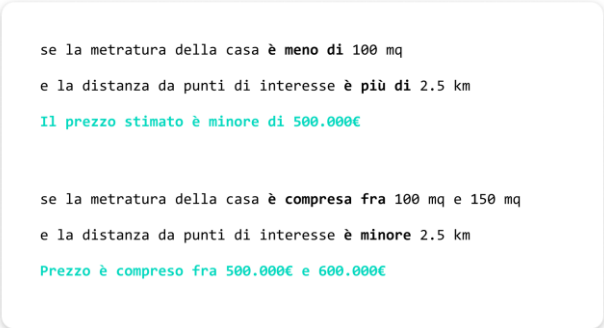
From the experiment results is clear that the form is able to answer the How to still be this Why and Why not question has the literature about question driven design approach for explainable AI has suggested. Additionally, new question answered resulted by our work: the rule flowchart explanatory form is able to answer properly event to How it works, How (under what conditions), How to be that, What if and what outputs questions.

Decision rules

The decision rules explanatory form is a representation of a predictive model that approximates it into a set of simple IF-THEN statements, known as decision rules, to explain the decision-making process that leads to a prediction.

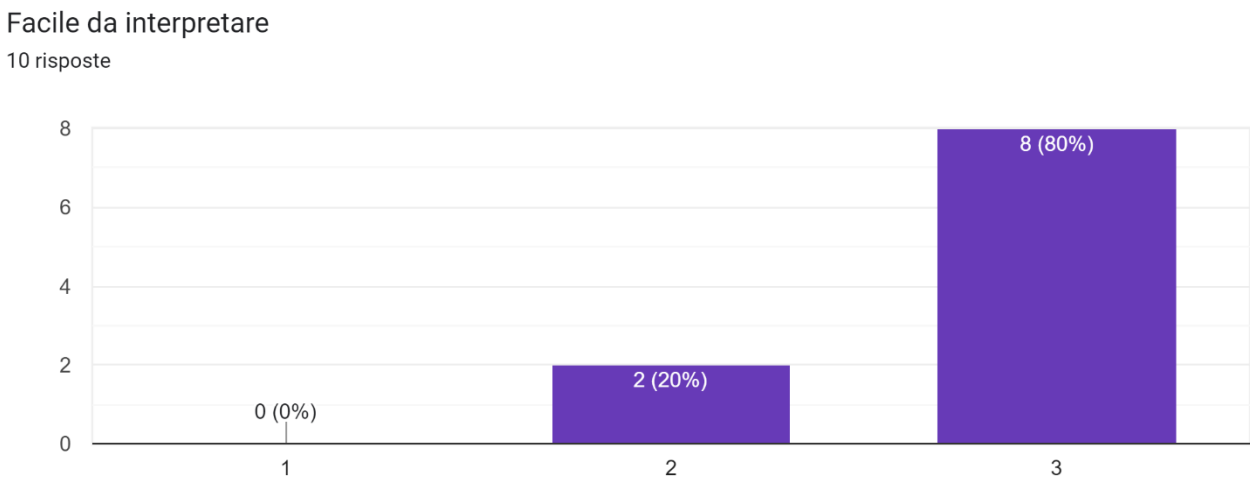
The decision rules specify the conditions that must be met for a certain prediction to be made, such as in the example of "IF blood sugar is high AND body weight is over-weighted, THEN the estimated diabetes risk is over 80%." This form allows for a clear explanation of how the model arrived at its predictions and provides insight into the underlying logic behind the model's decision-making process





Understandability

User friendliness and perceived usefulness

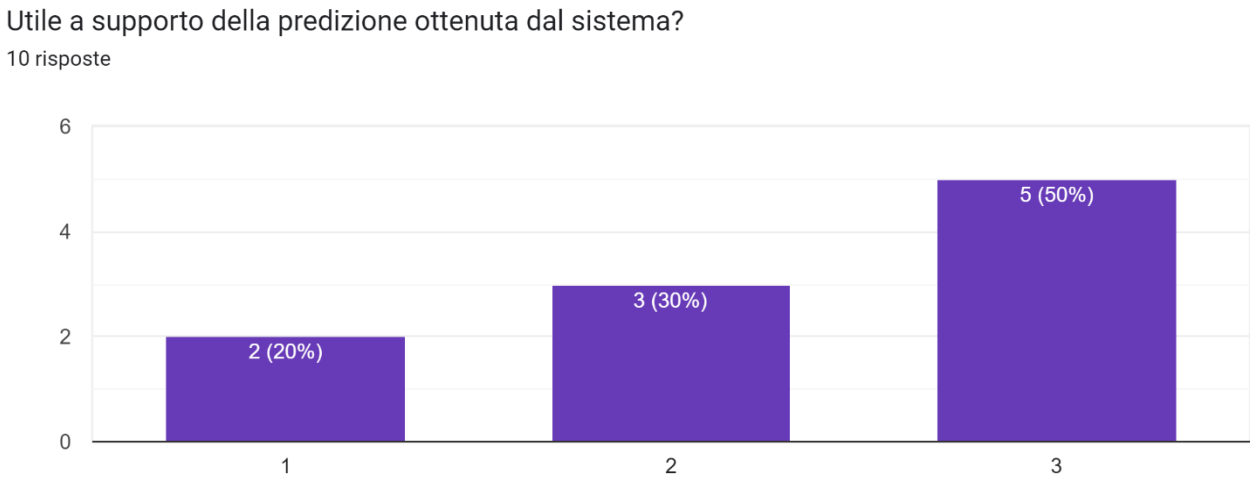


All of participants to the experiment found the explanatory form easy to be interpreted even if two of them claimed that the textual nature of the information provided was less effective than a graph form.

The participants generally found the decision rules explanatory form to be easy to interpret and extracted from it the rules laying behind the system functioning. They appreciated the list of parameters and corresponding values, evaluated as the main factor affecting the result. Easily, they recognized them as elements that could be changed to impact the given estimation. Anyway, some found the explanatory form incomplete and too long if all characteristics and possible scenarios were included.

Overall, the participants found the form to be informative and easy to interpret, However, some participants found the form to be too text-heavy and some suggested a more concise or schematic representation especially because, with much more parameters, the form would become difficult to understand. At this regard, one participant suggested a combination between this explanatory form and the decision three.

Applicable context of use:



Overall participants has different opinions about the usefulness of the decision rules explanatory form but only 2 out of ten directly declared his usefulness in any context. One participant found the form to be not so functional but explanatory even if he found difficult to correlate it with the question they may have in mind, some other mentioned

the difficulty to relate it to their own prediction because of the amount of different possible cases.

Anyway, overcoming these issues, the main context of use mentioned for the decision rules explanatory form has been to understand better the prediction and to have hints about what to change to have a different outcome.

Explanation Usability

Literature claims

In the literature about question driven design approaches the model the decision rules explanatory form has been mentioned both by the first work by Liao at Al. and then in the UXAI work as able to answer to **a why, why not, what if but mainly how it works** questions

User study results

Nine out of ten participants found this explanatory form able to answer to at least one of the protypical questions analised.

How it works ● **8/9** ● **Partial** ● **Direct**

The participants (8/9) indicated that the decision rules explanatory form is partially able to answer the "how it works" question. It provides information on the predictions coupling condition and results, and explains the logic behind the decisions letting users to understand that is based on the value of the different parameters. However, to get this information, participants underlined again, that the textual nature of the form longer the process and make it more cognitive demanding.

How ● **8/9** ● **Complete** ● **Direct**

Eight out of nine of the interviewers suggested that the decision rules explanatory form hele them to have a complete and direct answer to the "how (under what conditions)" question by providing clear explanation, telling the precise range of values, conditions

under which the result is obtained, thus indicating the parameters and their values that has been considered to provide the result.

Why 6/9 Partial Direct

The decision rules explanatory form provides users with an answer to the "why" question by linking the results to general characteristics and specific values considered thus enabling users to understand the connection between their outcomes and the relevant factors that have been considered. That reasoning has been mentioned by six out of nine participants during the experiment.

Why not 5/9 Complete Direct

In this case the why not question has been mentioned as answered as a consequence of what answer the ‘why’ question: because input feature is not in the ranges that would lead you to another result. The answer has been considered as mainly complete and directly given by half of participants.

How to be that 9/9 Complete Indirect

The decision rules explanatory form has been recognized all of the participants as helping to answer completely and directly to the "how to be that" question by providing all the values ranges thus allowing participants understanding what to change to have another result. However, participants considered the process to be potentially long and tedious, as well as time-consuming.

How to still be this 4/9 Partial Indirect

The decision rules explanatory form might not provide a straightforward answer to the question "how to still be this." For most of users, as a matter of fact only 4 out of 9 participants mentioned that question in their analysis and described the given answer as mainly partial and indirect. Even in this case the main cons are related to the potentially long and tedious, as well as time-consuming process to get the answer due to the amount of thing rule to consider and merge together.

What if 8/9 Complete Direct

The decision rules explanatory form has been recognised able to answer completely and directly to a what if question by 8 out of 9 respondents ‘because by reading all the sentences we have all the options’.

What outputs 6/9 Complete Direct

By providing all the possible system predictions following the textual formula IF/THEN the decision rules explanatory form has been recognised able to answer completely and directly to ‘what output’ question by 6 out of 10 participants.

What data 4/9 Complete Direct

Four out of nine participants considered the ‘what data’ question answered by the decision rules explanatory form due to the conveying of the different feature considered, thus able to explain what data are used to provide prediction.

Conclusion

Question type	Mentions	Efficiency	Effectiveness
How to be that	9/9	Complete	Indirect
How it works	8/9	Partial	Direct
How	8/9	Complete	Direct
What if	8/9	Complete	Direct
What outputs	6/9	Complete	Direct
Why	6/9	Partial	Direct
Why not	5/9	Complete	Direct
How to still be this	4/9	Partial	Indirect
What data	4/9	Complete	Direct

The decision rules explanatory form has been mentioned as able to answer to nine out of ten of the question proposed by the question driven design approach for explainable AI. most of the experiment participants mentioned it able to answer in approximatively complete way to the How to be that, How it works, How (under what conditions), What if and questions. Less mentioned in terms of frequency the What outputs, Why, Why not, How to still be this and the What data questions.

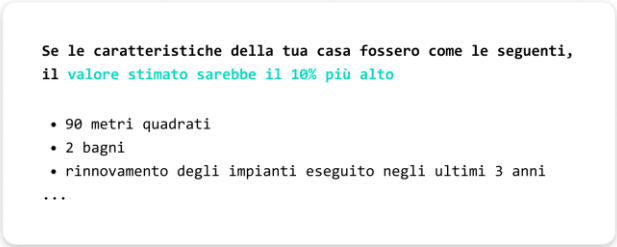
From the experiment results is clear that the form is greatly able to answer to the what if and how it works questions has the literature about question driven design approach for explainable AI has suggested. Surprisingly, the literature doesn't mention our most mentioned question: the How to be that and How (under what condition answers). The why and why not questions, even if confirmed by the experiment resulted not as much mentioned as we expected, they were claimed to be answered only by the half of respondent. Lastly, new question answered by the decision rules explanatory form

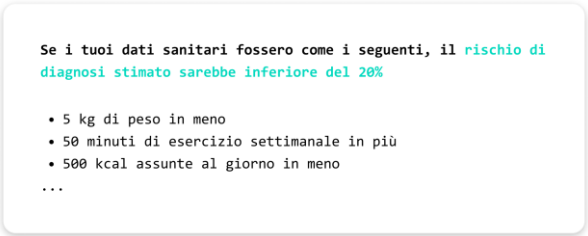
resulted by our work even if with not so many mentions: What outputs 6/10, Hot to still be this 4/10 and what data 4/10.

Counterfactual example

A counterfactual example is a variation of an input instance with minimal changes in its features that results in a distinct prediction outcome. These changes, whether they are absent or present, can be used to describe the feature(s) that would alter the original prediction if perturbed. For example, if an input instance (I) is predicted to have diabetes based on its high blood sugar level, a counterfactual example (C) would have all the same features as I, except for a lower blood sugar level, resulting in a prediction of good health.

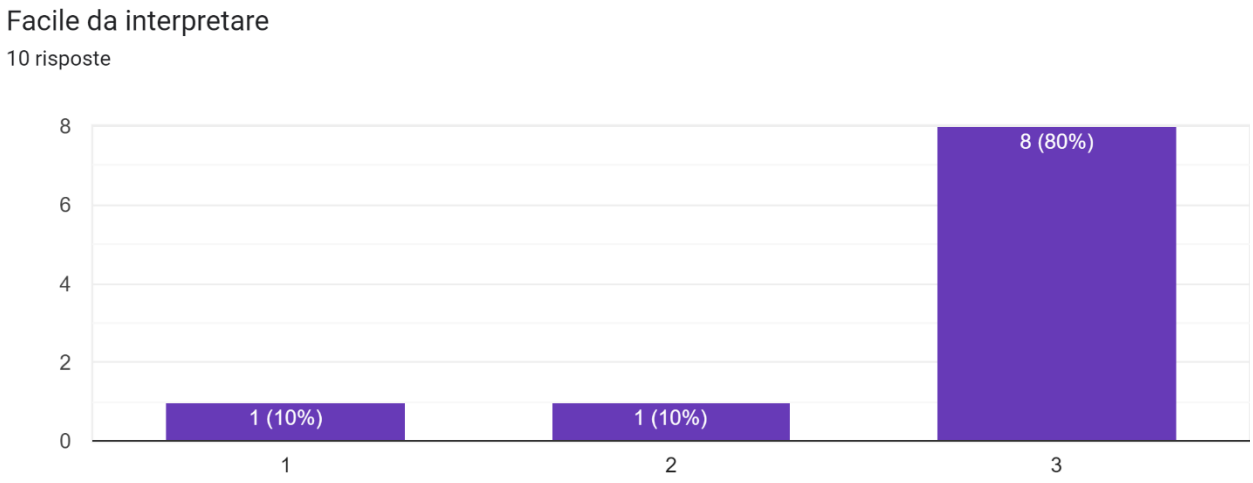
In our experiment the counterfactual example card contained a sentence describing what prediction could have been get by the system if the features and the corresponding values which followed it would have been used as input instance.





Understandability

User friendliness and perceived usefulness

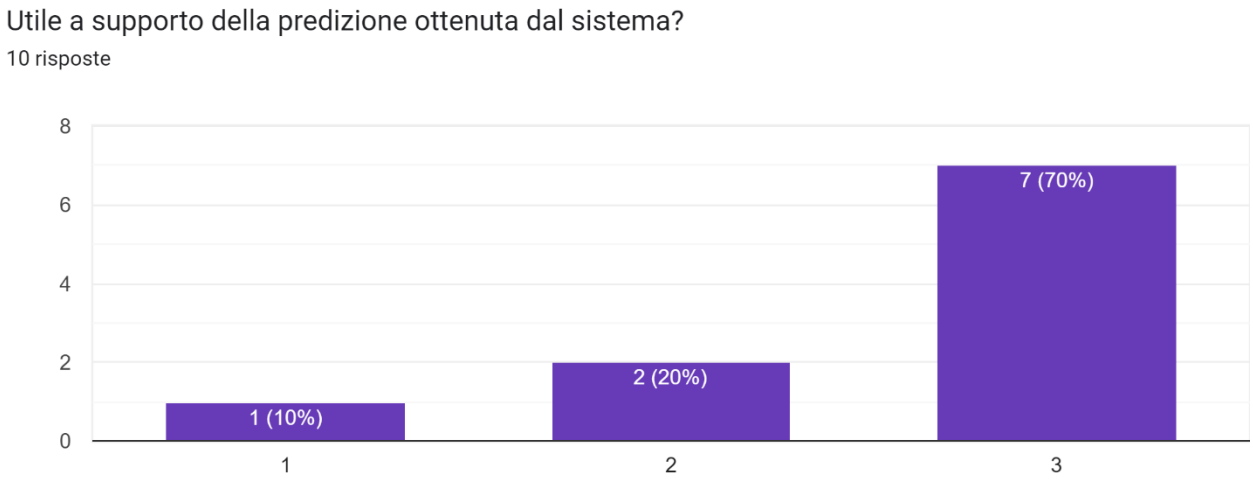


The counterfactual example explanatory form has been evaluated as easy to be interpreted by almost all the participants to the experiment (8/10)

Overall, the participants were able to extract from the counterfactual example information about how to modify certain features to improve the given prediction and get some clues about the logic behind why certain changes in the input features would be effective, Some participants appreciated that what the explanatory form provided seemed to be like a ‘recipe’ for how to improve the result, underlying only the features that can be effectively changed (e.g the age was not mentioned in the diabetes risk scenario) in comparison with the other explanatory form that conveyed similar information; Other participants expressed the preference to have the row data that can help them to understand that ‘recipe’ on their own.

For the counterfactual example explanatory form, the experiment conducted hasn’t highlighted any factor of misapplying.

Applicable context of use



The counterfactual example explanatory form has been recognized as useful to support the system outcome by most of the participants

Here the main context of use mentioned:
• To change the given prediction in order to improve it
• To do comparison with the input features and to understand what influence positively or negatively the outcome thus, get to know the logic behind the prediction

Explanation Usability

Literature claims

In the literature about question driven design approaches the model the counterfactual example explanatory form has been mentioned both by the first work by Liao at Al. and then in the UXAI work as greatly able to answer to **Why, why not, how to be that** questions

User study results

All of the participants found this explanatory form able to answer to at least one of the protypical questions analised. This data directly correlate with the found difficulties in interpret it.

How it works 6/10 Partial Indirect

Six out of ten participants recognized the counterfactual example explanatory form as able to provide some clues about how the system works thanks to the list of features provided and the hints about what values influence positively or negatively the outcome.

How 3/10 Partial Indirect

A partial answer to the how question has been retrieved from the counterfactual example explanatory form by three out of ten participants, but only comparing with their input features, that's why the process has been recognized as indirect and cognitive effortful.

How to be that 10/10 Complete Direct

All of the participants claimed that the counterfactual example explanatory form is able to provide an complete and direct answer to the 'how to be that' question. Anyway one participant asked: what if I would change to get another prediction [compared to the given one]?'.

What if 7/10 Partial Direct

A partial answer to a 'What if question' has been considered as given by the counterfactual example explanatory form in 7 out of ten cases during our experiment. The reason behind this answer partiality is obviously due to the fact that the explanatory form only provides one possibility to achieve one possible outcome.

What data 4/10 Partial Direct

The counterfactual example explanatory form provided an answer to a 'what data' question according to 4 out of 10 participants thanks of the list of features provided: this has been recognized only as partial to understand on what data the system rely to provide predictions.

Conclusions

Question type	Mentions	Efficiency	Effectiveness
How to be that	10/10	Complete	Direct
What if	7/10	Partial	Direct
How it works	6/10	Partial	Indirect
What data	4/10	Partial	Direct
How	3/10	Partial	Indirect

The counterfactual example explanatory form resulted able to answer completely for all the experiment participants to the How to be that question. All the other questions resulted, listed in order of mentions were only partial answered: What if [7/10], How it works [6/10], What data [4/10], How (under what conditions) [3/10]

The experiment results confirmed the ability of the counterfactual example explanatory form to answer to the How to be that question as the previous literature about question driven design approaches for explainable AI were suggesting. Not the same confirm has resulted from the analysis for the Why and Why not questions, mentioned only from 1 participant out of 10 thus not reaching a statistically relevant number. Lastly, our experiment results highlight new – partial – answers that the counterfactual example may provide to What if, How it works, What data and How (under what conditions) questions.

Discussion

Questions

6.3.1 Questions as user needs

In the moment of familiarisation with the questions participants has been asked to comment their impression aloud, as well as the scenario/context of use in which they thought those question type would be helpful. The analysis allows us to highlight correlation between question and explanation needs from the user perspective and, in some cases, the information they would search to get an answer to them:

How it works?

According to participants impressions. having an answer to the how it works question can help to solve the disagreement with the AI (P01: 'if I expect another output I can understand if I were wrong or if the data on which the systems rely are wrong) or to learn from the AI (P01: I can predict in the future without AI helping'). Some respondents claimed that this information is important to be given at the beginning of the interaction (P02: 'I want to know it before starting to use the system and to learn how to use it.'; P05: 'I'm interested to get to know how it works to grab an overall understanding of what's behind, and to what data relies on'). Generally, an 'How it works' answer is important, especially for AI novices, to build the trust in the system and start to build a general understanding of the functioning behind AI based systems (P10: 'since I don't know anything of course I have this question in mind')

How (under what conditions)?

The 'How (under what conditions)' answer is considered helpful to know ideal conditions to get a desired outcome and to change input parameters accordingly accomplish the goal to improve the predicted outcome. In this direction participants would search for information about the most important features considered to calculate the output (P01: I would like to know the most weighted thing for the risk to correct it'). For some respondents the How question is directly linked to the How it works, because the same information are able to provide an answer: what are the features and value considered that lead the system to give a precise output.

Why?

According to participants first sight, to answer to a why question is needed to communicate the most weighted features used in the computation as well as the list of

characteristic and values that made the output, again with the goal to improve the predicted outcome. It has been mentioned the need to get explanation in an understandable fashion with concrete and tangible data which can be actively changed (P06:'I need to understand the concomitant factors that lead to that result there and it can only be useful if they are tangible, concrete and variable, so I understand them and then I can change them')

Why not?

Get an answer to a why not question has been recognised useful in case of a disagreement with the AI (P01:'When I disagree and wonder if there is something wrong'), P05 mentioned that answering this question may provide awareness on the system reliability. P06 claimed that knowing why directly provide an answer to a why not question.

How confident?

The how confident question has been considered as the most important: most of participants claimed that should be mandatory and included in every phase of the interaction. Of course, the main explanatory goal mentioned achieved answering this question has been calibrate the trust in the system. It's worth to be mentioned that, from an AI novice perspective As Information able to provide the answer has been mentioned who developed the system, with what scientific consultancy, on what scientific knowledge (especially for the healthcare scenario) and who is using it. This align with the literature on social interactions as explanations in building trust in Ai based systems.

What data?

A 'what data' answer has been recognised as important in case of disagreement with the AI and related with the system reliability thus to build calibrated trust (P01:" When I disagree I ask myself if all the relevant data is there, it is related to reliability however if I trust it I don't ask that" or P02:" is like equations in mathematics that you learn knowing the result already, yes these data serve to build confidence"). Get to know the data is generally considered important, either to answer to the How it works question

and, additionally, answer a what data question has been recognised as well as related to the explanatory goal to detect bias.

What output(s)?

The ‘what output’ question has been subjected of tricky impressions, participants has not clear why should be interesting to know. After some reasonings participants considered it slightly useful to cope with curiosity and suggest to presented it in a side page of the system.

How to still be this?

The How to still be this question has been recognized as relevant to learn from AI (P02: “However, it makes sense to ask these questions, even to teach”) and to build trust toward the system. (P05: “Here it is related to the conditions for achieving a specific result and is perhaps related to reliability”). Additionally, has been recognised as important if you get a desirable result. To answer it participants mentioned to need information about the weight of feature used to compute the output. This question has been recognised as linked to the How (under what conditions) question.

How to be that?

Similar to the how to still be this question the ‘how to be that question’ has been recognized as relevant to learn from AI (P02: “However, it makes sense to ask these questions, even to teach”). Additionally, has been recognized as useful in case of disagreement with AI (P10: “I would not ask if I’m satisfied with the result, and if I am not satisfied probably yes”). Directly link in participants mind to the how to still be this question, considered as the contrary.

What if?

The what if question has been recognised as a broader view of the ‘How to be that’ and ‘How to still be this’ one. To answer it participants assumed they would get to know, with trials and error the weight of features took into consideration by the AI system (P05:” This one gives me more weight than the features, with trial and error”).

6.3.2 State of art in meeting user's needs

Turning the experiment results into a question point of view we can evaluate the state of art of explainable AI techniques to meet user needs (aka answer prototypical questions). The following table summarizes it for each of them:

Question type	Absolute value	Relative value	Maximum value	Minimum value
How it works	0.66	0.67	1.00	0.00
How	0.49	0.59	1.00	0.00
Why	0.32	0.52	1.00	0.00
Why not	0.31	0.38	1.00	0.00
How to be that	0.54	0.74	1.00	0.00
How to still be this	0.37	0.47	0.89	0.00
What if	0.47	0.55	1.00	0.00
What data	0.34	0.50	0.75	0.00
How confident	0.30	0.40	1.00	0.00
What outputs	0.38	0.61	1.00	0.00

State of art of questions answered by the state of art of explainable AI visualisations

From an absolute point of view, so considering all the 14 explanatory forms to calculate this score the how confident question is the one less answered, followed from the why not and why ones and then the what data. On the contrary the most answered is the how it works one.

From a relative perspective, thus calculating this score only considering the explanatory form resulted as able to answer the question the "how to be that" results as the one most probable to be answered for most of AI novices followed by the why not and again the How it works. Poor results for the how confident and the how to still be this and what data.

The max and min row has been added to make clear the span of effectiveness among the state of art of explainable AI techniques to answer those questions: when the max is 1 there's at least one explanatory form in literature certainly able to provide an answer to this, according to our study results. The minimum level, resulted as 0 for each question, inform practitioners about the possibility to not provide hints at all in answering the question thus warn them to take care of the explanatory form decision phase in the design process for XAI system in order to not make the system unable to meet all the possible user's needs in explanation seeking.

More importantly, we can notice that the max score of 1 for both the what data and the how to still be this question has been not achieved by any of the explanatory form in literature. This result may drive the XAI community to focus their effort in developing new XAI techniques to cover this gap.

6.3.3 Fine-grained question analisys:

As for the explanatory forms thanks to the qualitative analysis of the data resulted from our experiment, we were able to elaborate a fine-grained analysis of each question to guide practitioners which apply the question driven design approach for explainable AI to take informed decision in deciding the best explanatory form to meet user's need to design effective XUI. The following pages provide the detailed analysis of each of the 10 prototypical questions structured as follows. The first section is dedicated to a description resulted from the literature review: definition and sub questions belonging to the question type according to the literature review as well as general properties that this kind of explanation may have. The second section covers their analysis from our experiment results to orient the choice of the most usable explanatory form in the context of use of designing XUI. The paradigm followed to inform this decision is based on the explanatory form ability to provide usable answers to the prototypical question from an AI novices perspective. Thus, through the analysis, is differentiated between the explanatory form which answer completely or partially and in through a direct or indirect interaction highlighting what elements conveyed by them can provide AI novices hints to answer the questions.

How it works

Answering to an ‘how it works’ question helps users to get hints about How does the system make predictions, aiming to explain what is the overall model of how the system works.

To provide an how it works explanation the work on Question Driven Design Approach by Liao et al suggested the following main question and the relative sub questions:

How does the system make predictions?

- What features does the system consider?
- Is [feature X] used or not used for the predictions?
- What is the system’s overall logic?
- How does it weigh different features?
- What kind of rules does it follow?
- How does [feature X] impact its predictions?
- What are the top rules/features that determine its predictions?
- What kind of algorithm is used?
- How were the parameters set?

In our experiment context of use the how it works question were framed as follows:

- What is the overall logic the system follows to predict the diabetes risk?’ (Diabetes risk scenario)
- What is the overall logic the system follows to predict house prices?’(house selling scenario)

According to Question-Drive Design Process for XAI UX, generally, an ‘how it works’ explanation describe the general model logic as feature impact, rules or decision-trees**;** sometimes it’s needed to explain with a surrogate simple model and, If a user is only interested in a high-level view, describe what are the top features or rules considered.

Usable ‘How it works’ explanations

Literature claims

According to literature, bot UXAI and QDDA state that feature importance, decision tree approximation and rule extraction explanatory forms are the best candidates to provide an ‘how it works’ explanation.

User study results

Our analysis deepened what explanatory forms are able to provide an ‘How it works’ explanation, completely or partially and directly or indirectly.

Explanatory form	Satisfaction	Efficiency	Effectiveness
Global feature importance	10/10	Complete	Direct
Decision tree	9/9	Complete	Direct
Rule flowchart	8/8	Complete	Direct
Feature importance	9/10	Complete	Direct
Decision rules	8/9	Partial	Direct
Feature relevance	8/10	Partial	Indirect
Counterfactual example	6/10	Partial	Indirect
Typical example	4/7	Partial	Indirect
Feature shape	4/8	Partial	Indirect
Similar example	3/8	Partial	Indirect

Efficient ‘How it works’ explanations

The how it works question has been evaluated as completely answered by the following explanatory form, listed according to the number of participants mentions: Global feature importance (10/10), decision tree (9/9), rule flowchart (8/8), local feature importance (9/10).

The elements in the explanatory forms that helped participants to the research to get a efficient answer to the ‘how it works’ question were:

•Clearly highlight the input influential features

•The average weight of each feature and their precise value for each prediction

•Show the Step-by-step decision-making process

•Helping in understanding correlations between parameters and values

On the contrary the following explanatory forms, again listed according to the number of participants mentions, has been recognized as only able to answer the ‘how it works’ question partially: Decision rules (8/9), Feature relevance (8/10), Counterfactual example (6/10), Typical example (4/7), Feature shape (4/8), Similar example (3/8)

The elements in the explanatory forms that helped participants to the research to get a partial answer to the ‘how it works’ question were:

•Showing parameters.

•Offering comparisons through examples.

Effective ‘How it works’ explanations

The explanatory forms able to answer directly to the ‘how it works’ question are Global feature importance, Rule flowchart, Decision tree. Local Feature importance, Decision rules, Feature interaction.

On the contrary, the remaining explanatory forms even though able to answer to the question provided hints just indirectly slowing the process of informing user mental models: Feature relevance, Counterfactual example, Typical example, Feature shape, Similar example, Dataset, Output accuracy.

Conclusion:

The global feature importance is the most usable explanatory form to answer to an ‘how it works’ question:

Explanatory form	Satisfaction	Efficiency	Effectiveness
Global feature importance	10/10	Complete	Direct

How (under what conditions)

Answering to an ‘how (under what condition)’ question helps users to get hints about How does the system decide for a particular prediction, aiming to explain are the conditions that lead to a precise outcome.

The work on Question Driven Design Approach by Liao et al hasn’t divided between how it works and how (under what condition) question. Since their nature is the same, to pursue the aim of this work, the How works question and related subquestion has been thus reframed to refer only to a precise prediction:

- **How does the system decide for a particular prediction?**
  - What features does the system consider?
  - Is [feature X] used or not used for the get [the prediction]?
  - What is the system’s overall logic?
  - How does it weigh different features?
  - What kind of rules does it follow?
  - How does [feature X] impact [this] prediction?
  - What are the top rules/features that determine [the prediction]?
  - What kind of algorithm is used?
  - How were the parameters set?

In our experiment context of use the how under what condition’ questions were framed as follows:

- ‘Under what conditions the system predicts a diabetes risk of 80%?’ (Diabetes risk scenario)

- ‘Under what conditions the system predicts a house price between 420k and 470k?’ (House selling scenario)

According to the Question-Drive Design Process for XAI UX discussion, an ‘how (under what condition)’ explanation may describe the general model logic as feature impact, rules or decision-trees**;** sometimes it’s needed to explain with a surrogate simple model and, If a user is only interested in a high-level view, describe what are the top features or rules considered to get a precise prediction

Usable ‘How (under what conditions)’ explanations

Literature claims

As said, in literature, only the UXAI work has divided the how (under what condition) from the how it works question. In their analysis, the suggested explanatory form able to helps users to answer to the question is the feature importance one.

User study results

Our analysis deepened what explanatory forms are able to provide an How (under what condition) explanation, completely or partially and directly or indirectly.

Explanatory form	Satisfaction	Efficiency	Effectiveness
Decision tree	9/9	Complete	Direct
Rule flowchart	8/8	Complete	Direct
Decision rules	8/9	Complete	Direct
Feature relevance	8/10	Partial	Direct
Feature shape	6/8	Partial	Indirect
Feature interaction	5/7	Partial	Direct
Typical example	3/7	Partial	Indirect
Global feature importance	4/10	Partial	Direct
Similar example	3/8	Partial	Indirect
Feature importance	3/10	Partial	Direct
Counterfactual example	3/10	Partial	Indirect

Efficient ‘How’ explanations

The how (under what conditions) question has been evaluated as completely answered by the following explanatory form, listed according to the number of participants mentions: decision tree (9/9), rule flowchart (8/8), Decision rules (8/9).

The elements in the explanatory forms that helped participants to the research to get a complete answer to the ‘how (under what conditions)’ questions were:

•Information on the values that parameters must have recognized as the conditions that must be met, to get the various outcome ranges.

•Justify the step-by-step prediction process with the relative outcome ranges

On the contrary the following explanatory forms, again listed according to the number of participants mentions, has been recognized as only able to answer the ‘how (under what condition)’ question partially: Feature relevance (8/10), Feature shape (6/8), Feature interaction (5/7), Typical example (3/7), Global feature importance (4/10), Similar example (3/8), Feature importance (3/10), Counterfactual example (3/10).

The elements in the explanatory forms that resulted not enough to provide a complete ‘how (under what conditions)’ answer for participants are:

•Showing parameters value to get a prediction only for one or two features at the time.
•Offering comparisons through examples or data points.
•Clarify the weight of different parameters to get the predictions

Effective ‘How’ explanations

The explanatory forms able to answer directly to the ‘how’ question are the Decision tree, Rule flowchart, Decision rules, Feature relevance, Feature interaction, Global and local feature importance.

On the contrary, the remaining explanatory forms even though able to answer to the question, provided hints just indirectly slowing the process of informing user mental models: Feature shape, Typical example, Similar example and the Counterfactual example.

Conclusion:

The decision tree is the most usable explanatory form to answer to an ‘how’ question:

Explanatory form	Satisfaction	Efficiency	Effectiveness
Decision tree	9/9	Complete	Direct

Why

Answering to an ‘Why’ question helps users to get hints about Why did the system decide for a particular prediction.

To provide a ‘why’ explanation the work on Question Driven Design Approach by Liao et al suggests to answer to the following main question and the relative subquestions:

- Why/how is this instance given this prediction?
    - What feature(s) of this instance determine the system’s prediction of it?
    - Why are [instance A and B] given the same prediction

In our experiment context of use the ‘why’ questions were framed as follows:

- Why did the system evaluate my risk as 80%?’ (Diabetes risk scenario)
- Why did the system estimate my house price between 420k and 470k?’ (House selling scenario)

According to the Question-Drive Design Process for XAI UX a ‘Why’ explanation may describe what key features of the instance determine the model’s prediction of it, describe rules that the instance fits to guarantee the prediction, show similar examples with the same predicted outcome to justify the model’s prediction.

Usable ‘Why’ explanations

Literature claims

The QDDA work from Liao et al. identified 7 explanatory forms as able to provide hints to answer a why question: Rules or trees as approximation or extractions, contrastive or counterfactual feature, feature importance and saliency, prototypical or representative and counterfactual examples. Later the UXAI contribution added the feature influence or relevance and the model confidence ones to the list.

User study results

Our analysis deepened what explanatory forms are able to provide why explanation, completely or partially and directly or indirectly.

Explanatory form	Satisfaction	Efficiency	Effectiveness
Rule flowchart	8/8	Complete	Direct
Decision tree	9/9	Complete	Direct
Decision rules	8/9	Partial	Direct
Feature importance	9/10	Complete	Direct
Typical example	4/7	Partial	Indirect
Feature relevance	8/10	Complete	Indirect
Feature shape	4/8	Partial	Indirect
Feature interaction	4/7	Partial	Direct
Similar example	3/8	Partial	Indirect

Efficient ‘Why’ explanations

The ‘Why’ question has been evaluated as completely answered by the following explanatory form, listed according to the number of participants mentions: Rule flowchart (8/8), Decision tree (8/9), Feature importance (6/10), Feature relevance (5/10)

The elements in the explanatory forms that helped participants to the research to get a complete answer to the ‘why’ question are:

- the values of the relevant features that are evaluated in each steps of the decision making process allowing the user to compare with the input ones
- Provide which factors are considered most important and how they contribute to the final result.

- Showing the values of the features and following the parameters

On the contrary the following explanatory forms, again listed according to the number of participants mentions, has been recognized as only able to answer the ‘why’ question partially: Decision rules (6/9), Typical example (4/7), Feature shape (4/8), Feature interaction (3/7), Similar example (3/8).

The elements in the explanatory forms were considered not enough to get a complete ‘why’ answer from participants were:

- Showing parameters value to get a prediction only for one or two features at the time.
- Offering comparisons through examples or data points.
- Clarify the weight of different parametres to get the predictions

Effective ‘How’ explanations

The explanatory forms able to answer directly to the ‘why’ question are: Rule flowchart, Decision rules, Decision tree, Feature importance, feature interaction.

On the contrary, the remaining explanatory forms even though able to answer to the question, provided hints just indirectly slowing the process of informing user mental models: Typical example, Feature relevance, Feature shape, Similar example, Global feature importance and Counterfactual example.

Conclusion:

The rule flowchart is the most usable explanatory form to answer to a ‘why’ question:

Explanatory form	Satisfaction	Efficiency	Effectiveness
Rule flowchart	8/8	Complete	Direct

Why not

Answering to an ‘Why not’ question helps users to get hints about Why did not the system decide for a particular prediction.

To provide a ‘Why not’ explanation the work on Question Driven Design Approach by Liao et al suggests to answer to the following main question and the relative subquestions:

- **Why is this instance NOT predicted to be [a different outcome Q]?**
    - Why is this instance NOT predicted to be [a different outcome Q]?
    - Why are [instance A and B] given different predictions

In our experiment context of use the ‘why’ question were framed as follows:

- ‘Why not the system has predicted my diabetes risk as 50%?’ (diabetes risk scenario)
- ‘Why not the system has evaluated my house price between 300k and 350k?’ (house selling scenario)

According to the Question-Drive Design Process for XAI UX a ‘Why not’ explanation may describe what changes are required for the instance to get the alternative prediction and/or what features of the instance guarantee the current prediction; show prototypical examples that had the alternative outcome

Usable ‘Why’ explanations

Literature claims

The QDDA work from Liao et al. identified 4 explanatory forms as able to provide hints to answer a why not question: constrastive or counterfactual features, counterfactual example, decision tree approximation, rule extraction. The UXAI work contributed adding five more explanatory forms: local rules or trees, feature importance and saliency method, prototypical or representative examples, feature influence and relevance and model confidence.

User study results

Our analysis deepened what explanatory forms are able to provide ‘why not’ explanation, completely or partially and directly or indirectly.

Explanatory form	Satisfaction	Efficiency	Effectiveness
Rule flowchart	8/8	Complete	Direct
Decision tree	7/9	Complete	Direct
Decision rules	5/9	Complete	Direct
Feature shape	4/8	Partial	Indirect
Feature interaction	3/7	Partial	Direct
Feature relevance	3/10	Complete	Indirect

Efficient 'Why not' explanations

The 'Why not' question has been evaluated as completely answered by the following explanatory form, listed according to the number of participants mentions: Rule flowchart (8/8), Decision tree (7/9), Decision rules (5/9), Feature relevance (3/10).

The elements in the explanatory forms that helps to get a complete answer to the 'why not' question are:

- feature values in each decision-making process step, to highlight the conditions that may lead to different outcomes and allow the user to directly check the correspondence with their own features.

On the contrary the following explanatory forms, again listed according to the number of participants mentions, has been recognized as only able to answer the 'why' question partially: Feature shape (4/8), Feature interaction (3/7), Typical example (2/7), Similar example (3/8)

The elements in the explanatory forms considered as providing some 'hints' but not enough to get a complete 'why not' answer are:

•Showing parameters value to get a prediction only for one or two features at the time.

•Offering comparisons through examples or data points.

Effective ‘Why not’ explanations

The explanatory forms able to answer directly to the ‘why not’ question resulted to be: Rule flowchart, Decision tree, Decision rules, feature interaction, Feature importance.

On the contrary, the remaining explanatory forms even though able to answer to the question, provided hints just indirectly slowing the process of informing user mental models: Feature shape, Feature relevance, Typical example, Similar example, Global feature importance, Counterfactual example.

Conclusion:

The rule flowchart is the most usable explanatory form to answer to a ‘why not’ question:

Explanatory form	Satisfaction	Efficiency	Effectiveness
Rule flowchart	8/8	Complete	Direct

How to be that

Answering to an ‘how to be that’ question helps users to get hints about what would need to be changed for their instance to get a precise prediction, different from the given one.

To provide a ‘How to be that explanation the work on Question Driven Design Approach by Liao et al suggests to answer to the following main question and the relative subquestions:

- **How should this instance change to get a different prediction Q? •**
    - What is the minimum change required for this instance to get a different prediction Q?

- How should a given feature change for this instance to get a different prediction Q?
- What kind of instance is predicted of [a different outcome Q]?

In our experiment context of use the ‘How to be that’ questions were framed as follows:

- ‘How to get a diabetes risk as 50%?’ (Diabetes risk scenario)
- ‘How to get an estimated price as 500k?’ (House selling scenario)

According to the Question-Drive Design Process for XAI UX a ‘How to be that’ explanation should highlight features that, if changed (increased, decreased, absent, or present) could alter the prediction or Show examples with minimum differences but had a different outcome than the prediction

Usable ‘How to be that’ explanations

Literature claims

The QDDA work from Liao et al. identified 3 explanatory forms as able to provide hints to answer an how to be that question: constrastive or counterfactual features, counterfactual example, Feature influence or relevance method. Later on the UXAI work contributed to the list adding the feature importance and saliency one.

User study results

Our analysis deepened what explanatory forms are able to provide ‘how to be that’ explanation, completely or partially and directly or indirectly.

Explanatory form	Satisfaction	Efficiency	Effectiveness
Counterfactual example	10/10	Complete	Direct
Decision tree	9/9	Complete	Direct
Rule flowchart	8/8	Complete	Direct
Decision rules	9/9	Complete	Indirect
Feature relevance	8/10	Complete	Indirect
Feature shape	6/8	Partial	Indirect
Feature interaction	5/7	Partial	Direct
Feature importance	5/10	Partial	Indirect
Similar example	3/8	Partial	Indirect
Global feature importance	3/10	Partial	Indirect

Efficient ‘How to be that’ explanations

The ‘How to be that’ question has been evaluated as completely answered by the following explanatory form, listed according to the number of participants mentions: Counterfactual example (10/10), Decision tree (9/9), Rule flowchart (8/8), Decision rules (9/9), Feature relevance (8/10)

The elements in the explanatory forms that helps to get a complete answer to the ‘How to be that’ question are:

- A list or a path for feature values needed to get a specific prediction.

On the contrary the following explanatory forms, again listed according to the number of participants mentions, has been recognized as only able to answer the ‘how to be that’

question partially: Feature shape (6/8), Feature interaction (5/7), Feature importance (5/10), Similar example (3/8), Global feature importance (3/10).

The elements in the explanatory forms considered as providing some ‘hints’ but not enough to get a complete ‘How to be that’ answer from participants are:

- Showing parameters value to get a prediction only for one or two features at the time.
- Offering comparisons through similar examples
- Providing the list of the most weighted features to get a prediction

Effective ‘How to be that’ explanations

The explanatory forms able to answer directly to the ‘How to be that’ question resulted to be: Counterfactual example, Decision tree, Rule flowchart, feature interaction,

On the contrary, the remaining explanatory forms even though able to answer to the question, provided hints just indirectly slowing the process of informing user mental models: Decision rules, Feature relevance, Feature shape, Feature importance, Similar example, and Global feature importance.

Conclusion:

The counterfactual example is the most usable explanatory form to answer to a ‘How to be that’ question:

Explanatory form	Satisfaction	Efficiency	Effectiveness
Counterfactual example	10/10	Complete	Direct

How to still be this

Answering to an ‘how to still be this’ question helps users to get hints about what the scope of change is permitted to still get the same prediction they get from the decision making support system.

To provide a ‘How to still be this’ explanation the work on Question Driven Design Approach by Liao et al suggests to answer to the following main question and the relative subquestions:

- **What is the scope of change permitted for this instance to still get the same prediction?**
    - What is the range of value permitted for a given feature for this prediction to stay the same?
    - What is the necessary feature(s)/feature-value(s) present or absent to guarantee this prediction?
    - What kind of instance gets the same prediction?

In our experiment context of use the ‘How to still be this’ question, even if highlighting inconsistences with scenarios task, were framed as follows:

- ‘What can I change, in my lifestyle, to still get a diabetes risk as 80%?’ (Diabetes risk scenario)
- ‘What can I change, of my house, to still get an estimated price between 420k-470k?’ (House selling scenario)

According to the Question-Drive Design Process for XAI UX a ‘How to still be this’ explanation should describe features/feature ranges or rules that could guarantee the same prediction or show examples that are different from the particular instance but still had the same outcome.

Usable ‘How to still be this’ explanations

Literature claims

Both The QDDA work from Liao et al. and the UXAI work agreed on the ability of Feature influence and relevance method, local rules or tree or prototypical and representative examples to provide hints to answer an how to still be this question.

User study results

Our analysis deepened what explanatory forms are able to provide an ‘How to still be this’ explanation, completely or partially and directly or indirectly.

Explanatory form	Satisfaction	Efficiency	Effectiveness
Decision tree	8/9	Complete	Direct
Rule flowchart	7/8	Complete	Direct
Feature relevance	7/10	Complete	Indirect
Typical example	4/7	Complete	Direct
Feature shape	4/8	Partial	Indirect
Decision rules	4/9	Partial	Indirect
Feature importance	4/10	Partial	Indirect

Efficient ‘How to still be this’ explanations

The ‘How to still be this’ question has been evaluated as completely answered by the following explanatory form, listed according to the number of participants mentions: Decision tree (8/9), Rule flowchart (7/8), Feature relevance (7/10), Typical example (4/7)

The elements in the explanatory forms that helps to get a complete answer to the ‘How to be that’ question are:

- A list or a path for feature values for instances who would get the same given prediction.

On the contrary the following explanatory forms, again listed according to the number of participants mentions, has been recognized as only able to answer the ‘how to be that’ question partially: Feature shape (4/8), Decision rules (4/9), Feature importance (4/10).

The elements in the explanatory forms providing some ‘hints’ but not enough to get a complete ‘How to still be this’ answer are:

- Showing parameters value to get the prediction only for one or two features at the time.
- Offering comparisons through examples with similar features but different predictions
- Providing the list of the most weighted features to get predictions

Effective 'How to still be this' explanations

The explanatory forms able to answer directly to the 'how to still be this' question resulted to be: Decision tree, Rule flowchart, Typical example, Feature interaction.

On the contrary, the remaining explanatory forms even though able to answer to the question, provided hints just indirectly slowing the process of informing user mental models: Feature relevance, Feature shape, Decision rules, Feature importance.

Conclusions

The decision tree is the most usable explanatory form to answer to a 'How to still be this question:

Explanatory form	Satisfaction	Efficiency	Effectiveness
Decision tree	8/9	Complete	Direct

What if

Answering to a 'What if' question helps users to get hints about what whould be the system prediction if something different occur in the input instance, aiming to explain what the prediction in case of different input features could be.

To provide a 'What if' explanation the work on Question Driven Design Approach by Liao et al suggests to answer to the following main question and the relative sub questions:

- **What would the system predict if this instance changes to...?**
    - What would the system predict if a given feature changes to...?
    - What would the system predict for [a different instance]?

In our experiment context of use the What if' question were framed as follows:

- 'What would be the diabetes risk estimation if I would have different clinical data?' (Diabetes risk scenario)
- 'What would be my price estimation if my house would have different characteristics?' (House selling scenario)

According to the Question-Drive Design Process for XAI UX a 'What if' explanation should show how the prediction changes corresponding to the inquired change.

Usable 'What if' explanations

Literature claims

Both The QDDA work from Liao et al. and the UXAI work agreed on the ability of Feature influence and relevance method, decision tree approximation and rule extraction explanatory forms to provide hints to answer to a 'What if' question.

User study results

Our analysis deepened what explanatory forms are able to provide a 'What if' explanation, completely or partially and directly or indirectly.

Explanatory form	Satisfaction	Efficiency	Effectiveness
Feature relevance	10/10	Complete	Direct
Decision tree	9/9	Complete	Direct
Decision rules	8/9	Complete	Direct
Rule flowchart	7/8	Complete	Direct
Typical example	5/7	Partial	Indirect
Counterfactual example	7/10	Partial	Direct
Feature interaction	4/7	Complete	Direct
Feature shape	4/8	Partial	Indirect

Efficient ‘What if’ explanations

The ‘What if’ question has been evaluated as completely answered by the following explanatory form, listed according to the number of participants mentions: Feature relevance (10/10), Decision tree (9/9), Decision rules (8/9), Rule flowchart (7/8), Feature interaction (4/7)

The elements in the explanatory forms that helps to get a complete answer to the ‘What if’ question are:

- Allowing the user to control the instance features and get corresponding prediction at any change.

- Providing a list of instances directly manipulated or providing graphs, textual rules or tree visualisation and corresponding predictions ranges/values.

On the contrary the following explanatory forms, again listed according to the number of participants mentions, has been recognized as only able to answer the ‘how to be that’

question partially: Typical example (4/7), Counterfactual example (7/10), Feature shape (4/8)

The elements in the explanatory forms considered as providing some ‘hints’ but not enough to get a complete ‘What if’ answer are:

- Showing parameters value to get predictions only for one feature at the time.
- Offering comparisons through examples or dataset instances

Effective ‘What if’ explanations

The explanatory forms able to answer directly to the ‘how to still be this’ question resulted to be: Feature relevance, Decision tree, Decision rules, Rule flowchart, Counterfactual example, Feature interaction.

On the contrary, the remaining explanatory forms even though able to answer to the question, provided hints just indirectly slowing the process of informing user mental models: Typical example and Feature shape.

Conclusions

The feature relevance is the most usable explanatory form to answer to a ‘What if’ question:

Explanatory form	Satisfaction	Efficiency	Effectiveness
Feature relevance	10/10	Complete	Direct

What data

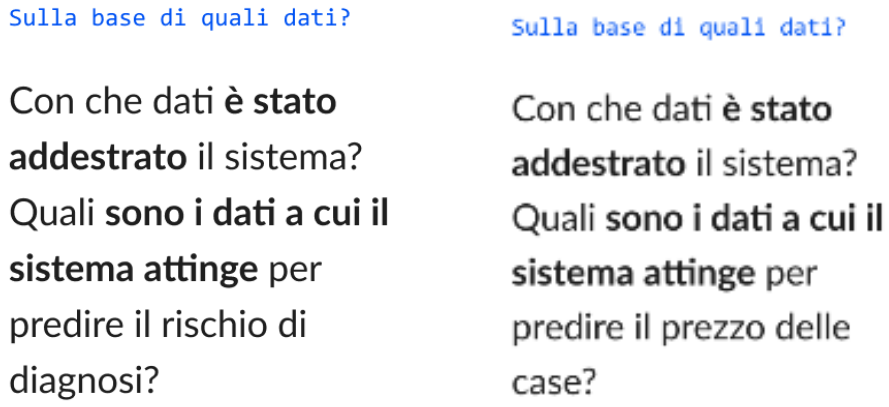
Answering to an ‘What data’ question helps users to get hints about what data does the system use to learn/has learned from’ aiming to explain what are the information used by the system to take decisions.

To provide a ‘What data’ explanation the work on Question Driven Design Approach by Liao et al suggests answering to the following main question and the relative sub questions:

- **What kind of data was the system trained on?**
    - What is the source of the training data?
    - How were the labels/ground-truth produced?
    - What is the sample size of the training data?
    - What dataset(s) is the system NOT using?
    - What are the potential limitations/biases of the data?
    - What is the size, proportion, or distribution of the training data with given feature(s)/feature-value(s)?

In our experiment context of use the 'What data' question were framed as follows:

- 'What are the data on which the system has been trained on? What are the data used by the system to predict diabetes risk?' (Diabetes risk scenario)
- 'What are the data on which the system has been trained on? What are the data used by the system to estimate house price?' (House selling scenario)



According to the Question-Drive Design Process for XAI UX a 'What data' explanation
•Document comprehensive information about the training data, including the source, provenance, type, size, coverage of population, potential biases, etc.

Usable 'What data' explanations

Literature claims

The QDDA work from Liao et al. doesn't provide any suggestion or example about explanatory forms able to answer a what data question. The UXAI mentions a generical

data source, as a list of features took into consideration by the system to get predictions.

User study results

Our analysis deepened what explanatory forms are able to provide a ‘What data’ explanation, completely or partially and directly or indirectly.

Explanatory form	Satisfaction	Efficiency	Effectiveness
Feature shape	6/8	Partial	Indirect
Dataset	4/7	Complete	Direct
Typical example	4/7	Partial	Indirect
Similar example	4/8	Partial	Indirect
Decision rules	4/9	Complete	Direct
Feature interaction	3/7	Complete	Direct
Counterfactual example	4/10	Partial	Direct
Decision tree	3/9	Complete	Direct
Feature relevance	3/10	Complete	Direct
Global feature importance	3/10	Partial	Indirect

Efficient ‘What data’ explanations

The ‘What data’ question has been evaluated as completely answered by the following explanatory form, listed according to the number of participants mentions: Dataset (4/7), Decision rules (4/9), Feature interaction (3/7), Decision tree (3/9), Feature relevance (3/10)

The elements in the explanatory forms that helps to get a complete answer to the ‘What data’ question are:

- Showing samples of data points in the train set allowing the user to directly interact with them
- Listing of the features considered by the system to give the predictions

On the contrary the following explanatory forms, again listed according to the number of participants mentions, has been recognized as only able to answer the ‘What data’ question partially: Feature shape (6/8), Typical example (4/7), Similar example (4/8) , Counterfactual example (4/10), Global feature importance (3/10).

The elements in the explanatory forms considered as providing some ‘hints’ but not enough to get a complete ‘What data’ answer are:

- Showing parameters value to get the predictions only for one feature at the time.
- Offering comparisons through examples
- Providing the list of the most weighted features to get predictions

Effective ‘What data’ explanations

The explanatory forms able to answer directly to the ‘What data’ question resulted to be: Dataset, Decision rules, Feature interaction, Counterfactual example, decision tree, feature relevance.

On the contrary, the remaining explanatory forms even though able to answer to the question, provided hints just indirectly slowing the process of informing user mental models: Feature shape, Typical example, Similar example, Global and feature importance

Conclusions

There’s not a clear result for the most usable explanatory form to answer to a ‘What data’ question. In terms of serving the bigger pool of possible users the feature shape seems to be more satisfactorily but providing only partial answers and with a consistent cognitive effort. On the other hand, the dataset explanatory form, has less relevance in

terms of being recognised as able to answer the question but, for the one which make this effort provide complete and direct information. Anyway, compared to the other user needs, the one underlying the what data question is the less served by the current state of the explainable AI visualisation, at least for AI novices; that's why we claim that should be the primary focus for further developments.

Explanatory form	Satisfaction	Efficiency	Effectiveness
Feature shape	6/8	Partial	Indirect
Dataset	4/7	Complete	Direct

How confident

Answering to an 'How confident' question helps users to get hints about how certain the system in a prediction or outcome is.

To provide an 'How confident' explanation the work on Question Driven Design Approach by Liao et al suggests to answer to the following main question and the relative subquestions:

How accurate/precise/reliable are the predictions?

- How often does the system make mistakes?
- In what situations is the system likely to be correct/ incorrect?
- What are the limitations of the system?
- What kind of mistakes is the system likely to make?
- Is the system's performance good enough for...?

In our experiment context of use the 'How confident' question were framed as follows:

- 'How reliable is the system diabetes risk estimated by the system?' (Diabetes risk scenario)
- 'How reliable is the evaluation estimated by the system?' (House selling scenario)
-

According to the Question-Drive Design Process for XAI UX a 'What data' explanation should provide performance metrics of the model or Show uncertainty information for each prediction or Describe potential strengths and limitations of the model

Usable 'How confident' explanations

Literature claims

The QDDA work from Liao et al. doesn't provide any suggestion or example about explanatory forms able to answer an how confident question. The UXAI mentions generic model confidence examples as categories, numbers, N-best alternatives, or visualisations.

User study results

Our analysis deepened what explanatory forms are able to provide a 'How confident' explanation, completely or partially and directly or indirectly.

Explanatory form	Satisfaction	Efficiency	Effectiveness
Performance	10/10	Complete	Direct
Output accuracy	9/9	Complete	Direct
Feature shape	3/8	Partial	Indirect
Feature relevance	3/10	Partial	Indirect

Efficient 'How confident' explanations

The 'How confident' question has been evaluated as completely answered by the following explanatory form, listed according to the number of participants mentions: Performance (10/10) and Output accuracy (9/9).

The elements in the explanatory forms that helps to get a complete answer to the 'What data' question were:

- Showing the overall performance how the system with the error range
- Providing the accuracy score calculated for each prediction

On the contrary the following explanatory forms, again listed according to the number of participants mentions, has been recognized as only able to answer the ‘How confident’ question partially: Feature shape (3/8) and Feature relevance (3/10)

The elements in the explanatory forms considered as providing some ‘hints’ but not enough to get a complete ‘How confident’ answer are:

- Provide the prediction as a range of values, suggesting the system capabilities and the uncertainty level
- Showing parameters value that led to different outcomes
- Offering examples or data samples making users evaluate prediction based on their own opinions.
- Providing the list of the most weighted features to get predictions

Effective ‘How confident’ explanations

The explanatory forms able to answer directly to the ‘What data’ question resulted to be: Performance and Output accuracy.
On the contrary, the remaining explanatory forms even though able to answer to the question, provided hints just indirectly slowing the process of informing user mental models: Feature shape and Feature relevance.

Conclusions

System performance is the most usable explanatory form to answer to a ‘How confident’ question:

Explanatory form	Satisfaction	Efficiency	Effectiveness
Performance	10/10	Complete	Direct

What outputs

Answering to a ‘What output’ question helps users to understand what are the possible outcome that the system can produce, setting the proper expectation about its functioning and preventing misuse or misunderstandings.

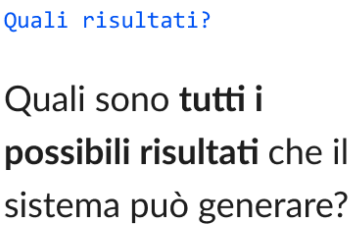
To provide a ‘What output’ explanation the work on Question Driven Design Approach by Liao et al suggests answering to the following main question and the relative sub questions:

• What kind of output does the system give?

- What does the system output mean?
- What is the scope of the system’s capability? Can it do...?
- How is the output used for other system component(s) ?
- How should I best utilize the output of the system
- How should the output fit in my workflow?

In our experiment context of use the ‘What output’ question were framed as follows:

- ‘What are all the possible outcome that the system can give?’ (Diabetes risk and House selling scenario)



According to the Question-Drive Design Process for XAI UX a ‘What output’ explanation should Describe the scope of output or system functions or Suggest how the output should be used for downstream tasks or user workflow.

Usable ‘What outputs’ explanations

Literature claims

The QDDA work from Liao et al. doesn’t provide any suggestion or example about explanatory forms able to answer an how confident question. The UXAI mentions generic system capability explanation aimed to show what the system can do.

User study results

Our analysis deepened what explanatory forms are able to provide a ‘What outputs’ explanation, completely or partially and directly or indirectly.

Explanatory form	Satisfaction	Efficiency	Effectiveness
Feature shape	8/8	Complete	Direct
Decision tree	8/9	Complete	Direct
Rule flowchart	8/8	Complete	Direct
Dataset	5/7	Partial	Indirect
Decision rules	6/9	Complete	Direct
Feature relevance	8/10	Complete	Indirect
Feature interaction	3/7	Complete	Indirect

Efficient ‘What outputs’ explanations

The ‘What output’ question has been evaluated as completely answered by the following explanatory form, listed according to the number of participants mentions: Feature shape (8/8), Decision tree (8/8), Rule flowchart (6/8), Decision rules (6/9), Feature relevance (6/10), Feature interaction (3/7).

The elements in the explanatory forms that helps to get a complete answer to the ‘What data’ question were:

- Listing/showing all the different results user could get based on different instance features

On the contrary the dataset (5/7) explanatory form has been recognized as only able to answer the ‘What if’ question partially.
The elements in the explanatory form considered as providing some ‘hints’ but not enough to get a complete ‘What outputs answer are:

- Offering a bunch of similar examples or data samples from the database
- Provide outcome as a range of possible values

Effective ‘What outputs’ explanations

The explanatory forms able to answer directly to the ‘What output’ question resulted to be: Feature shape, Decision tree, Rule flowchart, Decision rules.
On the contrary, the remaining explanatory forms even though able to answer to the question, provided hints just indirectly slowing the process of informing user mental models: Dataset, Feature relevance, Feature interaction.

Conclusions

Feature shape is the most usable explanatory form to answer to a ‘What outputs’ question:

Explanatory form	Satisfaction	Efficiency	Effectiveness
Feature shape	8/8	Complete	Direct

7. Limitations and further developments

The analysis performed to extract the state of art of XAI techniques took as a reference the most promising frameworks found in literature to design explanatory interfaces, for that reason we may have inherited their limitation. Since conducting a systematic literature review on the topic was out of the scope of this thesis the results achieved by this work may not take into consideration the last developments of this everyday growing field.

In addition, our experiment was constrained into the scenarios boundaries and the methodology that has followed, this means that there's a possibility that the data we gained was biased by the fact that the explanatory forms was not truly interactive or related to a real world application and context of use as well as some of them or the questions, proposed with the aim to serve general consideration may not be so relatable in the scenario provided (e.g. in the diabetes risk, since the given prediction was a bad score, the question how to still be this may have been evaluated as useless by most of the participants.) affecting their results in terms of perceived usefulness or completeness of the given answer.

Additionally, in relation to the experiment, even if designed ro follow a random sequence in showing the explanatory forms to participants to avoid bias, may not be enough to guarantee that already grasped information about the system functioning by the first ones showed has not biased the following ones during the activity.

This work represents a first attempt to evaluate explanatory forms usability with a restricted pool of AI novices, next studies may took our results as a reference to conduct a focused analysis in terms of explanation usability and ability to answer questions with an higher pool of participants.

As introduced in the methodology section, the quantitative analysis performed to aggregate experiment results must be taken only as a reference in supporting the qualitative one since, again, the pool of participants in the experiment has been not enough to conduct exhaustive qualitative research.

This last consideration can be the starting point for the further development of this work in order to combine the new results with the quantitative analysis performed by Jin et al. for the EUCA framework which resulted in a granular analysis in pairing explanatory forms and explanatory goals. We claim that restricting the knowledge on the current state of XAI techniques and their visualisation (explanatory form) is still limited thus performing an additional analysis to correlate explanatory goals and question seeking

may truly generates actionable knowledge for practitioners to inform the development of effective explanatory narratives with end to end user flows which, triggered by the user need characterized by the explanatory goals and following the principle of progressive disclosure may overcome all the limitation found in the literature review we performed to finally provide interactive explanations targeted on user needs able to answer follow up question until the level of understanding needed for the intention of use [the explanatory goal] is fulfilled.

8. Conclusions

This thesis covered the topic of Understandable AI and Explanation Usability from AI novices perspective and contributed to the literature on designing explainable AI user experiences in the context of explainable interfaces providing actionable insights for practitioners to design explanatory narratives serving AI novices user needs, expressed, according to the question driven design approach for explainable AI, as question users may have in mind while they are seeking for explanations.

The discussion has covered the first research question of this thesis (What is the AI novice reasoning at the first interaction with explanation types? What information are easily caught, what mental model they inform, what is their perceived usefulness and their intention of use?) analysing in detail the level of understandability of the 14 explanatory forms, visualisations of the state of art of explainable AI techniques outputs, providing insights in terms of user friendliness and perceived usefulness and highlighting user's opinion in terms of applicable context of use and possible strategies to improve their visualisation in order to overcome possible misleading factors.

From the analysis of the AI novices' experiences with the 14 explanatory forms we were able to extract actionable insights to characterize AI novices - XAI interaction patterns, to inform practitioners on how to enhance explanation effectiveness for this disregarded from literature target group.

For what concern the second research question (Explanatory forms/explanation type can convey the information needed to AI novices to answers the prototypical questions given by the question driven design approach?), this work contributes to advance the field of the question driven design approach for explainable AI with the first user study focused on the capability of explanation type to serve user explanatory needs (namely, answer the 10 prototypical questions) directly involving end users themselves.

This resulted into a general analysis of the state of art of explainable AI visualisation to serve AI novices needs and, within the fine-grained explanatory forms analysis, an actionable resource for XUI designers which list, for each of them, what question are the best answered following the usability evaluation differentiating the one given in an efficient manner i.e., providing complete or partial answer and if they are given with effectiveness aka if the information grasped to answer them are directly or indirectly given, thus, more or less cognitive demanding.

In this direction our contribution to the field of usable explanations adopt the definition of usability as the combination of effectiveness, efficiency, and satisfaction, as per

ISO9241-210, 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” and postulates that the most usable explanation is the one which is most probable to serve user needs/goals (answer the prototypical question user have in mind) resulting in a higher level of satisfaction, secondly is needed to convey the right amount of information to fulfill their goal (thus, providing a complete answer) while requiring the minimum amount of cognitive effort (thus, providing a direct answer).

Following the same reasoning, we have reversed the explanatory forms’ results providing a fine-grained question analysis to provide, for each of the 10 prototypical questions representing user needs while seeking for explanations what the most usable explanatory form are to answer them in order to fulfil user needs.

Additionally, for each question the elements extracted from the form to answer them are summarised differentiating what ones can provide complete or partial hints in order to serve the whole XAI community in the development of new XAI techniques able to provide the most effective explainable experiences even for AI novices, in the near future.

9. Appendix

Tables:

explanatory form user friendliness scale

Explanatory form	User friendliness
Feature relevance	3
Similar example	3
Decision tree	2,9
Feature importance	2,9
Typical example	2,9
Output accuracy	2,9
Decision rules	2,8
Counterfactual example	2,7
Performance	2,7
Global feature importance	2,6
Rule flowchart	2,5
Feature shape	2,1
Dataset	2
Feature interaction	1,7

Explanatory form perceived usefulness scale

Explanatory form	Perceived usefulness
Feature importance	2,9
Decision tree	2,8
Performance	2,8
Global feature importance	2,8
Feature relevance	2,7
Output accuracy	2,6
Counterfactual example	2,6
Rule flowchart	2,5
Decision rules	2,3
Similar example	2,2
Feature shape	2,1
Feature interaction	2
Typical example	1,9
Dataset	1,7

Fine grained explanatory forms analysis

explanatory form			Questions answered	question asweres mentioned
Decision tree			How it works [9/9] completely [7], partially [1], directly [9]: How (Local) [9/9] completely [7], partially [2], directly [9] How to be that [9/9] completely [7], partially [2], directly [7], indirectly [2] What if [9/9] completely [9], directly [8] Why [8/9] completely [6], partially [1], directly [6], indirectly [1] How to still be this [8/9] completely [6], partially [2], directly [6], indirectly [2] What outputs [8/9] completely [8], directly [8] Why not [7/9] completely [7] directly [7] What data [3/9] completely [3] directly [3]	70
Rule flowchart			How it works [8/8] completely [6], partially [1], directly [6] How (Local) [8/8] completely [6], partially [1], directly [8] Why [8/8] completely [6], partially [2], directly [6], indirectly [2] Why not [8/8] completely [8], directly [8] How to be that [8/8] completely [6], partially [1], directly [7] How to still be this [7/8] completely [6], directly [5], indirectly [1] What if [7/8] completely [5], partially [2], directly [7] What outputs [6/8] completely [4], partially [1], directly [4] what data [2/8] completely [1], partially [1], directly [1], indirectly [1] How confident [1/8], partially [1], indirectly [1]	63
Feature relevance			What if [10/10] completely , directly How it works [8/10] partially, indirectly How (Local) [8/10] partially [3], directly [6] How to be that [8/10] completely [5], indirectly [5] How to still be this [7/10] completely [3] indirectly [5] What outputs [6/10] completely [4],, indirectly [4] Why [5/10] completely [3], indirectly [3] Why not [3/10] partially [1], indirectly [2] How confident [3/10] partially [2], indirectly [2] What data [3/10] completely [2], directly [2]	61

			How confident [3/10] completely [1], partially [2], directly [1], indirectly [2] What data [3/10] completely [2], partially [1], directly [2]	
Decision rules			How to be that [9/9] completely [6/8], partially [2/8], directly [4], indirectly [3] How it works [8/9] completely [4], partially [4], directly [5], indirectly [3] How (Local) [8/9] completely [5], partially [3], directly [7], indirectly [1] What if [8/9] completely [6], partially [2], directly [6], What outputs [6/9] completely [6], directly [5], indirectly [1] Why [6/9] completely [3], partially [2], directly [3], indirectly [1] Why not [5/9] completely [3], partially [1], directly [3], indirectly [1] How to still be this [4/9] completely [2], partially [2], directly [1], indirectly [2] What data [4/9] completely [3], partially [1], directly [3], indirectly [1] How confident [2/9] partially [2], indirectly [1]	60
Feature shape			What outputs [8/10] completely [6], partially [2], directly [5], indirectly [1] How (Local) [6/10] completely [1], partially [5], indirectly [5] How to be that [6/10] completely [1], partially [5], indirectly [4] What data [6/10] completely [1], partially [5], directly [1], indirectly [3] How it works [4/10] completely [1], partially [2], directly [1] Why [4/10] completely [1], partially [3], directly [1], indirectly [3] Why not [4/10] completely [2], partially [2], directly [1], indirectly [3] How to still be this [4/10] partially [4], indirectly [2] What if [4/10] completely [1], partially [3], directly [1], indirectly [2] How confident [3/10] partially [3], indirectly [2]	45
Counterfactual example			How to be that [10/10] completely [9], partially [1], directly [10] What if [7/10] partially [7], directly [5], indirectly [2] How it works [6/10], partially [6], directly [3], indirectly [3] What data [4/10] partially [4], directly [3] How (Local) [3/10] partially [2], indirectly [2] How to still be this [2/10], partially [2], indirectlyt [1] Why [1/10], indirectly [1] Why not [1/10], indirectly [1]	34

Feature importance			How it works [9/10] completely [4], partially [2], directly [6], indirectly [1] Why [6/10] completely [5], partially [1], directly [6], How to be that [5/10] completely [2], partially [2], indirectly [4] How to still be this [4/10] completely [2], partially [1], indirectly [3] How (Local) [3/10] completely [1], partially [2], directly [2] Why not [2/10] completely [1], partially [1], directly [1] What data [2/10] completely [1], partially [1], directly [1]	31
Feature interaction			How (Local) [5/7] partially [5], directly [1], indirectly [3] How to be that [5/7] partially [4], indirectly [4] What if [4/7] completely [2], partially [2], directly [1], indirectly [3] Why [3/7] partially [3], indirectly [1] Why not [3/7] partially [2], directly [1], indirectly [2] What outputs [3/7] completely [2], partially [1], directly [1], indirectly [1] What data [3/7] completely [3], directly [2] How it works [2/7] partially [2, indirectly [1] How to still be this [2/7] partially [2], indirectly [2]	30
Typical example			What if [5/7] partially [5], directly [1], indirectly [3] How it works [4/7] partially [4], directly [2], indirectly [1] Why [4/7] partially [4], directly [1], indirectly [3] How to still be this [4/7] completely [3], partially [1], directly [4] What data [4/7] completely [1], partially [3], directly [1], indirectly [3] How (Local) [3/7] partially [3], directly [1], indirectly [2] Why not [2/7] partially [2], directly [1], indirectly [1] How confident [2/7 partially [2], indirectly [2]	28
Similar example			What data [4/8] completely [2], partially [2], directly [1], indirectly [1] How it works [3/8] partially [2], indirectly [3] How (Local) [3/8] partially [3], indirectly [2] Why [3/8] partially [3], directly [1], indirectly [1] How to be that [3/8] completely [1], partially [2], directly [2], indirectly [1] Why not [2/8] partially [1], directly [1], indirectly [2] How to still be this [2/8, partially [2], directly [1], indirectly [1] What if [2/8] , partially [2], directly [1], indirectly [1] How confident [2/8] partially [2], directly [1], indirectly [1]	26

		What outputs [2/8] completely [1], partially [1], directly [1], indirectly [1]	
Global feature importance		How it works [10/10] completely [5], partially [3], directly [9] How (Local) [4/10] completely [2], partially [2], directly [2] How to be that [3/10] partially [3] indirectly [2] What data [3/10] completely [1], partially [2], directly [1], indirectly [1] How confident [2/10] 2 partially, 1 indirectly How to still be this [1/10, partially [1] Why [1/10] partially [1], indirectly [1] Why not [1/10] partially [1], indirectly [1]	25
Dataset		What outputs [5/7] 1 completely 3 partially / 2 directly,3 indirectly What data [4/7] 2 completely 2 partially 2 directly,1 indirectly How confident [2/7] 2 partially / 1 directly,, 1 indirectly How it works [1/7] partially [1], indirectly [1] How [1/7] What if [1/7] partially [1] indirectly [1]	14
Output accuracy		How confident [9/9] 5 completamente 2 parzialmente 8 direttamente (?) What ouptputs [2/9] 2 parzialmente How it works [1/9] 1 parzialmente	12
Performance		How confident [10/10] completely [8], partially [1], directly [10]	10

explanatory form	**how it works**	**How local**	**why**	**why not**	**how to be that**	**how to stilll be this**	**what if**	**What data**	**How confident**	**What outputs**
Decision tree	1,00	1,00	0,89	0,78	1,00	0,89	1,00	0,33	0,00	0,89
Feature relevance	0,80	0,80	0,50	0,30	0,80	0,70	1,00	0,30	0,30	0,60
Rule flowchart	1,00	1,00	1,00	1,00	1,00	0,88	0,88	0,25	0,13	0,75
Decision rules	0,89	0,89	0,67	0,56	1,00	0,44	0,89	0,44	0,22	0,67
Feature shape	0,50	0,75	0,50	0,50	0,75	0,50	0,50	0,75	0,38	1,00
Counterfactual example	0,60	0,30	0,10	0,10	1,00	0,20	0,70	0,40	0,00	0,00
Feature importance	0,90	0,30	0,60	0,20	0,50	0,40	0,00	0,20	0,00	0,00

Feature interaction	0,29	0,71	0,43	0,43	0,71	0,29	0,57	0,43	0,00	0,43
Typical example	0,57	0,43	0,57	0,29	0,00	0,57	0,71	0,57	0,29	0,00
Global feature importance	1,00	0,40	0,10	0,10	0,30	0,10	0,00	0,30	0,20	0,00
Similar example	0,38	0,38	0,38	0,25	0,38	0,25	0,25	0,50	0,25	0,25
Dataset	0,14	0,14	0,00	0,00	0,00	0,00	0,14	0,57	0,29	0,71
Performance	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00
Output accuracy	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,22

explanatory form	**how it works**	**How local**	**why**	**why not**	**how to be that**	**how to stilll be this**	**what if**	**What data**	**How confident**	**What outputs**
Decision tree	9/9	9/9	8/9	7/9	9/9	8/9	9/9	3/9		8/9
Feature relevance	8/10	8/10	5/10	3/10	8/10	7/10	10/10	3/10	3/10	6/10
Rule flowchart	8/8	8/8	8/8	8/8	8/8	7/8	7/8	2/8	1/8	6/8
Decision rules	8/9	8/9	6/9	5/9	9/9	4/9	8/9	4/9	2/9	6/9
Feature shape	4/8	6/8	4/8	4/8	6/8	4/8	4/8	6/8	3/8	8/8
Counterfactual example	6/10	3/10	1/10	1/10	10/10	2/10	7/10	4/10		
Feature importance	9/10	3/10	6/10	2/10	5/10	4/10		2/10		
Feature interaction	2/7	5/7	3/7	3/7	5/7	2/7	4/7	3/7		3/7
Typical example	4/7	3/7	4/7	2/7		4/7	5/7	4/7	2/7	
Global feature importance	10/10	4/10	1/10	1/10	3/10	1/10		3/10	2/10	
Similar example	3/8	3/8	3/8	2/8	3/8	2/8	2/8	4/8	2/8	2/8
Dataset	1/7	1/7					1/7	4/7	02/07	5/7
Performance									10/10	
Output accuracy	1/9								09/09	02/09

explanatory form	how it works	How local	why	why not	how to be that	how to stilll be this	what if	What data	How confident	What outputs
Decision tree	9/9	9/9	8/9	7/9	9/9	8/9	9/9	3/9	x	8/9
Feature relevance	8/10	8/10	5/10	3/10	8/10	7/10	10/10	3/10	3/10	6/10
Rule flowchart	8/8	8/8	8/8	8/8	8/8	7/8	7/8	2/8	1/8	6/8
Decision rules	8/9	8/9	6/9	5/9	9/9	4/9	8/9	4/9	2/9	6/9
Feature shape	4/8	6/8	4/8	4/8	6/8	4/8	4/8	6/8	3/8	8/8
Counterfactual example	6/10	3/10	1/10	1/10	10/10	2/10	7/10	4/10		
Feature importance	9/10	3/10	6/10	2/10	5/10	4/10		2/10		
Feature interaction	2/7	5/7	3/7	3/7	5/7	2/7	4/7	3/7		3/7
Typical example	4/7	3/7	4/7	2/7		4/7	5/7	4/7	2/7	
Global feature importance	10/10	4/10	1/10	1/10	3/10	1/10		3/10	2/10	
Similar example	3/8	3/8	3/8	2/8	3/8	2/8	2/8	4/8	2/8	2/8
Dataset	1/7	1/7					1/7	4/7	2/7	5/7
Performance									10/10	
Output accuracy	1/9								9/9	2/9

Fine grained question analysis

Explanatory forms which answer ‘how it works questions’ (completely, partially, directly, indirectly)

Global feature importance	Decision tree	Rule flowchart	Feature importance	Decision rules	Feature relevance	Counterfactual example	Typical example	Feature shape	Similar example	Feature interaction	Dataset	Output accuracy
1,00	1,00	1,00	0,90	0,89	0,80	0,60	0,57	0,50	0,38	0,29	0,14	0,11
10/10	9/9	8/8	9/10	8/9	8/10	6/10	4/7	4/8	3/8	2/7	1/7	1/9

10/10	9/9	8/8	9/10	8/9	8/10	6/10	4/7	4/8	3/8	2/7	1/7	1/9

Explanatory forms which answer 'how it works questions' (completely, partially, directly, indirectly)

Decision tree	Rule flowchart	Decision rules	Feature relevance	Feature shape	Feature interaction	Typical example	Global feature importance	Similar example	Feature importance	Counterfactual example	Dataset
1,00	1,00	0,89	0,80	0,75	0,71	0,43	0,40	0,38	0,30	0,30	0,14
9/9	8/8	8/9	8/10	6/8	5/7	3/7	4/10	3/8	3/10	3/10	1/7
9/9	8/8	8/9	8/10	6/8	5/7	3/7	4/10	3/8	3/10	3/10	1/7

Explanatory forms which answer 'why' question (completely, partially, directly, indirectly)

Rule flowchart	Decision tree	Decision rules	Feature importance	Typical example	Feature relevance	Feature shape	Feature interaction	Similar example	Global feature importance	Counterfactual example
1,00	0,89	0,67	0,60	0,57	0,50	0,50	0,43	0,38	0,10	0,10
8/8	8/9	6/9	6/10	4/7	5/10	4/8	3/7	3/8	1/10	1/10
8/8	8/9	6/9	6/10	4/7	5/10	4/8	3/7	3/8	1/10	1/10

Explanatory forms which answer 'why not' question (completely, partially, directly, indirectly)

Rule flowchart	Decision tree	Decision rules	Feature shape	Feature interaction	Feature relevance	Typical example	Similar example	Feature importance	Global feature importance	Counterfactual example
1,00	0,78	0,56	0,50	0,43	0,30	0,29	0,25	0,20	0,10	0,10

8/8	7/9	5/9	4/8	3/7	3/10	2/7	2/8	2/10	1/10	1/10
8/8	7/9	5/9	4/8	3/7	3/10	2/7	2/8	2/10	1/10	1/10

Explanatory forms which answer 'why not' question (completely, partially, directly, indirectly)

Counterfactual example	Decision tree	Rule flowchart	Decision rules	Feature relevance	Feature shape	Feature interaction	Feature importance	Similar example	Global feature importance
1,00	1,00	1,00	1,00	0,80	0,75	0,71	0,50	0,38	0,30
10/10	9/9	8/8	9/9	8/10	6/8	5/7	5/10	3/8	3/10
10/10	9/9	8/8	9/9	8/10	6/8	5/7	5/10	3/8	3/10

Explanatory forms which answer 'How to still be this' question (completely, partially, directly, indirectly)

Decision tree	Rule flowchart	Feature relevance	Typical example	Feature shape	Decision rules	Feature importance	Feature interaction	Similar example	Counterfactual example	Global feature importance
0,89	0,88	0,70	0,57	0,50	0,44	0,40	0,29	0,25	0,20	0,10
8/9	7/8	7/10	4/7	4/8	4/9	4/10	2/7	2/8	2/10	1/10
8/9	7/8	7/10	4/7	4/8	4/9	4/10	2/7	2/8	2/10	1/10

Explanatory forms which answer 'What if' question (completely, partially, directly, indirectly)

Feature relevance	Decision tree	Decision rules	Rule flowchart	Typical example	Counterfactual example	Feature interaction	Feature shape	Similar example	Dataset
1,00	1,00	0,89	0,88	0,71	0,70	0,57	0,50	0,25	0,14
10/10	9/9	8/9	7/8	5/7	7/10	4/7	4/8	2/8	1/7
10/10	9/9	8/9	7/8	5/7	7/10	4/7	4/8	2/8	1/7

Explanatory forms which answer 'What data' question (completely, partially, directly, indirectly)

Feature shape	Dataset	Typical example	Similar example	Decision rules	Feature interaction	Counterfactual example	Decision tree	Feature relevance	Global feature importance	Rule flowchart	Feature importance
0,75	0,57	0,57	0,50	0,44	0,43	0,40	0,33	0,30	0,30	0,25	0,20
6/8	4/7	4/7	4/8	4/9	3/7	4/10	3/9	3/10	3/10	2/8	2/10
6/8	4/7	4/7	4/8	4/9	3/7	4/10	3/9	3/10	3/10	2/8	2/10

Explanatory forms which answer 'How confident' question (completely, partially, directly, indirectly)

Performance	Output accuracy	Feature shape	Feature relevance	Typical example	Dataset	Similar example	Decision rules	Global feature importance	Rule flowchart
1,00	1,00	0,38	0,30	0,29	0,29	0,25	0,22	0,20	0,13
10/10	9/9	3/8	3/10	2/7	2/07	2/8	2/9	2/10	1/8
10/10	9/9	3/8	3/10	2/7	2/07	2/8	2/9	2/10	1/8

Explanatory forms which answer 'What output' question (completely, partially, directly, indirectly)

Feature shape	Decision tree	Rule flowchart	Dataset	Decision rules	Feature relevance	Feature interaction	Similar example
1,00	0,89	0,75	0,71	0,67	0,60	0,43	0,25
8/8	8/9	6/8	5/7	6/9	6/10	3/7	2/8
8/8	8/9	6/8	5/7	6/9	6/10	3/7	2/8

10. References

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M., “Trends and trajectories for explainable, accountable and intelligible systems”, In: CHI 2018 (2018)

Adadi A. and Berrada M., "Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)", in IEEE Access, vol. 6, pp. 52138-52160, (2018), doi: 10.1109/ACCESS.2018.2870052.

Arrieta A. B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion”, Volume 58, (2020), Pages 82-115, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2019.12.012.

Arya, V., et al., “One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques”, arXiv (2019)

Berkovsky S., Ronnie Taib, and Dan Conway, “How to Recommend?: User Trust Factors in Movie Recommender Systems. In Proceedings of the 22nd International Conference on Intelligent User Interfaces”, (2017), (IUI ’17). ACM, New York, NY, USA, 287–300. https://doi.org/10.1145/3025171.3025209

Binns R., Kleek M.V., Veale M., Lyngs U., Zhao J., and Shadbolt N., (2018), “It’s Reducing a Human Being to a Percentage’: Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems”, ACM, 377.

Brennen A., 2020, “What Do People Really Want When They Say They Want “Explainable AI?” We Asked 60 Stakeholders”. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–7. https://doi.org/10.1145/3334480.3383047

Bussone A., Stumpf S., and O’Sullivan D., 2015, “The role of explanations on trust and reliance inclinical decision support systems”. In International Conference on Healthcare Informatics (ICHI). IEEE, 160–169.

Cai C.J., Jongejan J., and Holbrook J., 2019, “The effects of example-based explanations in a machine learning interface”. In Proceedings of the 24th International Conference on Intelligent User Interfaces. ACM, 258–262.

Caruana R., Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noémie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Vol. 2015-Augus. Association for Computing Machinery, New York, New York, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

Charles Antaki and Ivan Leudar. 1992. Explaining in conversation: Towards an argument model. European Journal of Social Psychology 22, 2 (1992), 181–194.

Chen J., Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 339–348.

Cheng H.F., Ruotong Wang, Zheng Zhang, Fiona O’Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 559.

Chromik M., Butz, A. (2021). Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. In: , et al. Human-Computer Interaction – INTERACT 2021. INTERACT 2021. Lecture Notes in Computer Science(), vol 12933. Springer, Cham. https://doi.org/10.1007/978-3-030-85616-8_36

Clark H.H. and Susan E Brennan. 1991. Grounding in communication. (1991).

Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K. and Dugan, C. (2019), “Explaining models: an empirical study of how explanations impact fairness judgment”, in Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 275-285, doi: 10.1145/3301275.3302310.

Doshi-Velez F. and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. (feb 2017). arXiv:1702.08608 http://arxiv.org/abs/1702.08608

Eiband M., Schneider H., Bilandzic M., Fazekas-Con J., Haug M. and Hussmann H. (2018), “Bringing transparency design into practice”, in Proceedings of the 23rd International Conference on Intelligent User Interfaces, pp. 211-223, doi: 10.1145/3172944.3172961

Erickson T., Catalina M Danis, Wendy A Kellogg, and Mary E Helander. 2008. Assistance: the work practices of human administrative assistants and their implications for it and organizations. In Proceedings of the 2008 ACM conference on Computer supported cooperative work. ACM, 609–618.

EUCA: End-User-Centered Explainable AI Framework by weinajin. (n.d.). EUCA: End-User-Centered Explainable AI Framework by Weinajin. https://weina.me/end-user-xai/

Ferreira, J.J., Monteiro, M.S. (2020). What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In: Marcus, A., Rosenzweig, E. (eds) Design, User Experience, and Usability. Design for Contemporary Interactive Environments. HCII 2020. Lecture Notes in Computer Science(), vol 12201. Springer, Cham. https://doi.org/10.1007/978-3-030-49760-6_4

Goebel, R., et al.: Explainable AI: the new 42? In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2018. LNCS, vol. 11015, pp. 295–303. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99740-7_21

Goodman, B., & Flaxman, S. (2017, October 2). European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation.” AI Magazine, 38(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gregor S. and Izak Benbasat. 1999. Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. MIS Quarterly 23, 4 (dec 1999), 497. https://doi.org/10.2307/249487

Grice H.P. 1975. Logic and Conversation. In Syntax and semantics 3: Speech arts. 41–58.

Guidotti R., Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, 5, Article 93 (August 2018), 42 pages. https://doi.org/10.1145/3236009

Gunning, D., & Aha, D. (2019, June 24). DARPA’s Explainable Artificial Intelligence (XAI) Program. AI Magazine, 40(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

H2O.ai Machine Learning Interpretability. https://github.com/h2oai/mli-resources. 2017.

Hilton D.J. 1990. Conversational processes and causal explanation. Psychological Bulletin 107, 1 (1990), 65.

Hohman F., Head A., Caruana R., DeLine R., and Drucker S.M. 2019. “Gamut: A design probe to understand how data scientists understand machine learning models”. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 579.

Holzinger A., Bernd Malle, Peter Kieseberg, Peter M. Roth, Heimo Müller, Robert Reihs, and Kurt Zatloukal. 2017. Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology. (dec 2017). arXiv:1712.06657 http://arxiv.org/abs/1712.06657

Holzinger A., Langs G., Denk H., Zatloukal K., Müller H.: Causability and explainability of artificial intelligence in medicine. WIREs Data Mining Knowl Discov. 2019; 9:e1312. https://doi.org/10.1002/widm.1312

Holzinger, A., Carrington, A. & Müller, H. Measuring the Quality of Explanations: The System Causability Scale (SCS). Künstl Intell 34, 193–198 (2020). https://doi.org/10.1007/s13218-020-00636-z

IBM AIX 360. aix360.mybluemix.net/. 2019.

Jin W., Fan, J., Gromala, D., Pasquier, P., & Hamarneh, G. (2021). EUCA: the End-User-Centered Explainable AI Framework, doi:10.48550/arXiv.2102.02437

Jin W., Mostafa Fatehi, Kumar Abhishek, Mayur Mallya, Brian Toyota, and Ghassan Hamarneh. 2020. Artificial intelligence in glioma imaging: challenges and advances. Journal of neural engineering 17, 2 (2020), 021002. https://doi.org/10.1088/1741-2552/ab8131 arXiv:1911.12886

Kim B., Austin, U.T.: Examples are not enough, learn to criticize! criticism for interpretability. In: NIPS (2016)

Kleinberg J., Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan, David Abrams, Matt Alsdorf, Molly Cohen, Alexander Crohn, Gretchen Ruth Cusick, Tim Dierks, John Donohue, Mark Dupont, Meg Egan, Elizabeth Glazer, Joan Gottschall, Nathan Hess, Karen Kane, Leslie Kellam, Angela LascalaGruenewald, Charles Loeffler, Anne Milgram, Lauren Raphael, Chris Rohlfs, Dan Rosenbaum, Terry Salo, Andrei Shleifer, Aaron Sojourner, James Sowerby, Cass Sunstein, Michele Sviridoff, Emily Turner, and Judge John. 2017. Human Decisions and Machine Predictions. September (2017), 1–53. https://doi.org/10.1093/qje/qjx032/4095198/Human-Decisions-and-Machine-Predictions

Kocielnik R., Saleema Amershi, and Paul N Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 411.

Krause J., Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, 5686–5697.

Kulesza T., Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debuggingto personalize interactive machine learning. In Proceedings of the 20th International Conference on Intelligent User Interfaces. ACM, 126–137.

Laato S., Tiainen M., Najmul, Islam A.K.M. and Mäntymäki M. (2022), "How to explain AI systems to end users: a systematic literature review and research agenda", Internet Research, Vol. 32 No. 7, pp. 1-31. https://doi.org/10.1108/INTR-08-2021-0600

Lai V. and Tan Chenhao, 2018. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. arXiv preprint arXiv:1811.07901 (2018).

Lee M.K., Jain A., Cha H.J., Ojha S., and Kusbit D., 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 182.

Liao Q.V., & Varshney, K.R. (2021). Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. ArXiv, abs/2110.10790.

Liao Q.V., Gruen D., and Miller S., 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

Liao Q.V., Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-Driven Design Process for Explainable AI User Experiences. arXiv preprint arXiv:2104.03483 (2021).

Lim B. Y, Qian Yang, Ashraf Abdul and Danding Wang. 2019. Why these Explanations? Selecting Intelligibility Types for Explanation Goals. In Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019, 7 pages. https://doi.org/10.1145/1234567890

Lim B.Y, 2011. Improving Understanding, Trust, and Control with Intelligibility in Context-Aware Applications. Human-Computer Interaction (2011).

Lim B.Y and Anind K Dey. 2009. Assessing demand for intelligibility in context-aware applications. In Proceedings of the 11th international conference on Ubiquitous computing. 195–204.

Lim B.Y, Anind K. Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Boston, MA, USA) (CHI '09). Association for Computing Machinery, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023

Lundberg S.M., Lee, S.: A Unified Approach to Interpreting Model Predictions (2017)

Madumal P., Miller T., Sonenberg L., and Vetere F. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019, IFAAMAS, 9 pages. https://doi.org/10.48550/arXiv.1903.02409

Mehrabi N., Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019).

Meske C., Abedin B., Klier M. et al., Explainable and responsible artificial intelligence. Electron Markets 32, 2103–2106 (2022). https://doi.org/10.1007/s12525-022-00607-2

Microsoft InterpretML. hhttps://github.com/interpretml/interpret. 2019.

Miller T. 2019. Explanation in artificial intelligence : Insights from the social sciences. Artificial Intelligence 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt B., Russell C. and Wachter S. 2019. Explaining Explanations in AI. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 279–288. https://doi.org/10.1145/3287560.3287574

Model Interpretation with Skater. https://oracle.github.io/Skater/. 2018.

Mohseni S., Zarei N., and Ragan E.D. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans. Interact. Intell. Syst. 1, 1, Article 1 (January 2020), 46 pages. https://doi.org/10.1145/3387166

Montavon G., Samek w., Müller K.-R., Methods for interpreting and understanding deep neural networks, Digital Signal Processing, Volume 73, 2018, Pages 1-15, ISSN 1051-2004, https://doi.org/10.1016/j.dsp.2017.10.011.

Moore, J.D., Paris, C.: Requirements for an expert system explanation facility. Comput. Intell. 7, 367–370 (1991)

Mueller S.T., Hoffman, R.R., Clancey, W., Emrey, A., Klein Macrocognition, G.: Explanation in human-AI systems. arXiv (2019)

Naiseh M., D. Cemiloglu, D. Al Thani, N. Jiang and R. Ali, "Explainable Recommendations and Calibrated Trust: Two Systematic User Errors," in Computer, vol. 54, no. 10, pp. 28-37, Oct. 2021, doi: 10.1109/MC.2021.3076131.

Neerincx MA, van Vught W, Blanson Henkemans O, Oleari E, Broekens J, Peters R, Kaptein F, Demiris Y, Kiefer B, Fumagalli D and Bierman B (2019) Socio-Cognitive Engineering of a Robotic Partner for Child's Diabetes Self-Management. Front. Robot. AI 6:118. doi: 10.3389/frobt.2019.00118

Norris, S.P., Guilbert, S.M., Smith, M.L., Hakimelahi, S., Phillips, L.M.: A theoretical framework for narrative explanation in science. Sci. Educ. 89(4), 535–563 (2005)

Nunes I. and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. User Modeling and User-Adapted Interaction 27, 3-5 (2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0

Oh C., Song J., Choi J., Kim, S., Lee, S. and Suh, B. (2018), "I lead, you help, but only with enough details: understanding user experience of co-creation with artificial intelligence", in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1-13, doi: 10.1145/3173574.3174223.

Oulasvirta A., Hornbaek, K.: HCI Research as Problem-Solving. CHI '16 (2016)

Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S. and Mohammadian, A.K. (2020), "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis", Accident Analysis and Prevention, Vol. 136, doi: 10.1016/j.aap.2019.105405.

Poursabzi-Sangdeh F., Goldstein D.G., Hofman J.M., Vaughan J.M., and Wallach H., 2018. Manipulating and measuring model interpretability. arXiv preprint arXiv:1802.07810 (2018).

Rader E., Cotter K., and Cho J. 2018. Explanations as mechanisms for supporting algorithmic transparency. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 103.

Ribeiro M.T., Singh, S., Guestrin, C.: Why Should I Trust You?: Explaining the Predictions of Any Classifier (2016)

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016), ""Why should I trust you?" Explaining the predictions of any classifier", in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144, doi: 10.1145/2939672.2939778.

Ribera M. and Lapedriza A., 2019. Can we do better explanations? A proposal of User-Centered Explainable AI. In Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019, 7 pages. ACM, New York, NY, USA, 7 pages. http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf

Santos, T.I., Abel, A.: Using feature visualisation for explaining deep learning models in visual speech. In: 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), pp. 231–235, March 2019. https://doi.org/10.1109/ICBDA.2019.8713256

Schoonderwoerd T.A.J., Jorritsma W., Neerincx M.A., Karel van den Bosch, Human-centered XAI: Developing design patterns for explanations of clinical decision support systems, International Journal of Human-Computer Studies, Volume 154, 2021,102684, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2021.102684.

Schrills, T. and Franke, T. (2020), “Color for characters-effects of visual explanations of AI on trust and observability”, in Proceedings of the International Conference on Human-Computer Interaction, pp. 121-135, doi: 10.1007/978-3-030-50334-5_8.

Shneiderman B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., Diakopoulos, N.: Confessions: grand challenges for HCI researchers. Interactions (2016)

Shneiderman, B.: Bridging the gap between ethics and practice. ACM Trans. Interact. Intell. Syst. 10, 1–31 (2020)

Sovrano, F., Vitali, F., Palmirani, M. (2020). Modelling GDPR-Compliant Explanations for Trustworthy AI. In: Kő, A., Francesconi, E., Kotsis, G., Tjoa, A., Khalil, I. (eds) Electronic Government and the Information Systems Perspective. EGOVIS 2020. Lecture Notes in Computer Science(), vol 12394. Springer, Cham. https://doi.org/10.1007/978-3-030-58957-8_16 [17]

Tsai, C.H., Brusilovsky, P.: Evaluating Visual Explanations for Similarity-Based Recommendations: User Perception and Performance. UMAP ’19 (2019)

UXAI: Home. (n.d.). UXAI: Home. https://www.uxai.design/

van der Waa, J., Schoonderwoerd T., van Diggelen J. and Neerincx M. (2020), “Interpretable confidence measures for decision support systems”, International Journal of Human–Computer Studies, Vol. 144, doi: 10.1016/j.ijhcs.2020.102493.

van der Waa J., Nieuwburg E., Cremers A. and Neerincx M. (2021), Evaluating XAI: a comparison of rule-based and example-based explanations”, Artificial Intelligence, Vol. 291, doi: 10.1016/j.artint.2020.103404.

Vilone G., Longo, L.: Explainable artificial intelligence: a systematic review. arXiv (2020)

Walton D. 2004. A new dialectical theory of explanation. Philosophical Explorations 7, 1 (2004), 71–89.

Wang D., Yang, Q., Abdul, A., Lim, B.Y., (2019), “Designing Theory-Driven User-Centric Explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems” (Glasgow, Scotland Uk) (CHI ’19). Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

Weitz, Schiller, Schlagowski, Huber, & André. (2019b). “Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design”. Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, 7–9. https://doi.org/10.1145/3308532.3329441

Weller, A., 2017. Challenges for transparency. arXiv preprint arXiv:1708.01870 (2017).

West D. M. (2018, July 30). The Future of Work. In Robots, AI, and Automation. Brookings Institution Press.

Wiegand G., Matthias Schmidmaier, Thomas Weber, Yuanting Liu, and Heinrich Hussmann. 2019. I drive-you trust: Explaining driving behavior of autonomous cars. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, LBW0163.

Xu, W.: Toward human-centered AI: A perspective from human-computer interaction. Interactions (2019)

Yang, L., Wang, H., Deleris, L.A. (2021). What Does It Mean to Explain? A User-Centered Study on AI Explainability. In: Degen, H., Ntoa, S. (eds) Artificial Intelligence in HCI. HCII 2021. Lecture Notes in Computer Science(), vol 12797. Springer, Cham. https://doi.org/10.1007/978-3-030-77772-2_8

Yin, M., Wortman Vaughan, J. and Wallach, H. (2019), “Understanding the effect of accuracy on trust in machine learning models”, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-12, doi: 10.1145/3290605.3300509.