frontiers | Frontiers in Computer Science

TYPE Technology and Code
PUBLISHED 08 January 2025
DOI 10.3389/fcomp.2024.1412341

Check for updates

OPEN ACCESS

EDITED BY
Maria Chiara Caschera,
National Research Council (CNR), Italy

REVIEWED BY
Pantelis Pergantis,
University of the Aegean, Greece
Jaesik Choi,
Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea

*CORRESPONDENCE
Marco Matarese
✉ marco.matarese@iit.it

RECEIVED 30 October 2024
ACCEPTED 18 December 2024
PUBLISHED 08 January 2025

CITATION
Matarese M, Rea F, Rohlfing KJ and Sciutti A (2025) How informative is your XAI? Assessing the quality of explanations through information power.
Front. Comput. Sci. 6:1412341.
doi: 10.3389/fcomp.2024.1412341

COPYRIGHT
© 2025 Matarese, Rea, Rohlfing and Sciutti. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

How informative is your XAI? Assessing the quality of explanations through information power

Marco Matarese[1]*, Francesco Rea[1], Katharina J. Rohlfing[2] and Alessandra Sciutti[1]

[1]CONTACT Unit, Italian Institute of Technology, Genoa, Italy, [2]Faculty of Arts and Humanities, Paderborn University, Paderborn, Germany

A growing consensus emphasizes the efficacy of user-centered and personalized approaches within the field of explainable artificial intelligence (XAI). The proliferation of diverse explanation strategies in recent years promises to improve the interaction between humans and explainable agents. This poses the challenge of assessing the goodness and efficacy of the proposed explanation, which so far has primarily relied on indirect measures, such as the user's task performance. We introduce an assessment task designed to objectively and quantitatively measure the goodness of XAI systems, specifically in terms of their "information power." This metric aims to evaluate the amount of information the system provides to non-expert users during the interaction. This work has a three-fold objective: to propose the Information Power assessment task, provide a comparison between our proposal and other XAI goodness measures with respect to eight characteristics, and provide detailed instructions to implement it based on researchers' needs.

KEYWORDS

explainable artificial intelligence, XAI objective assessment, human-in-the-loop, information power, qualitative explanations' quality

1 Introduction

The widespread adoption of complex machine learning (ML) models across diverse applications has accentuated the need for explainable artificial intelligence (XAI) methods to address the inherent opacity of these systems. As artificial intelligence (AI) technologies continue to play pivotal roles in critical decision-making processes across multiple domains, the demand for transparency and interpretability has become increasingly pronounced. Consequently, there is a growing emphasis on researching and evaluating the effectiveness and reliability of XAI methods.

Recent developments in the XAI research field have shifted the focus toward user-centered approaches to explainability, fostering the proliferation of personalized XAI methods with increased complexity (Williams, 2021). These user-centric approaches find application in diverse contexts, ranging from computer applications designed for personalized teaching (Embarak, 2022; Cohausz, 2022) to Human–Robot Interaction (HRI) scenarios where agents maintain user-specific models to deliver tailored explanations (Matarese et al., 2021; Stange et al., 2022). These efforts show a promising trend: user-centered XAI positively influences the interaction between users and systems and increases trust in AI solutions (Ali et al., 2022). This positive impact brings to higher users' willingness to reuse the system [e.g., with recommendations (Conati et al., 2021)],

agents' persuasiveness during human–agent decision-making tasks and human–AI teams performance (Schemmer et al., 2022b).

Alongside producing personalized XAI, recent works also aim to evaluate the goodness of the explanation produced. However, the term "XAI goodness" remains vague, with consensus acknowledging its dependency on the application contexts and intended users. A main disadvantage is that existing evaluations often rely on subjective or indirect measurements. Most assessment approaches do not involve users directly; when they do, they usually use indirect measures such as preference or performance. Self-reported questionnaires are commonly employed to assess users' explainability preferences, while more objective measurements focus on human–AI performance or users' ability to simulate the AI behavior.

Miller (2019) argued that research in XAI can draw inspiration from humanities studies. In those fields, an explanation is considered good when it results in understanding (Lombrozo, 2016). However, the definition of understanding in response to explanation is unclear: Buschmeier et al. (2023) propose a range from shallow to deep comprehension and enabledness. While there are many approaches to assessing understanding, they underline that understanding serves specific tasks and depends on the purpose and goal of explanations. In essence, the challenge is not only in providing explanations that result in understanding but also in tailoring the level and type of understanding to the user's needs and the context of the task at hand. This aligns with the broader notion that the effectiveness of an explanation depends on its relevance to the user's goals and the specific requirements of the situation (Booshehri et al., 2024).

In the field of human–AI collaboration, assessments of XAI systems' goodness have often been based on indirect measures, such as team performance, system persuasiveness, or users' predictive abilities regarding AI decisions (Vilone and Longo, 2021). This emphasizes the lack of objective and quantitative measures to gauge the inherent quality of XAI systems to compare across different contexts. This gap contributes to the difficulty of rigorously comparing different XAI strategies—and also comparing them across contexts—, as the results of existing studies are deeply tied to the application under consideration. The contextual dependence of XAI models may contribute to inconsistent findings in the field (Islam et al., 2022).

Following Hoffman et al. (2018), we assume that the accuracy of the users' mental models about an AI system's functioning reflects the goodness of the explanations received. Therefore, when dealing with expert AI systems, users' mental models regarding the task at hand can be considered reflective of the quality of the explainable agent they interacted with. Building on these assumptions, we are interested in introducing an objective and quantitative assessment task that allows the measurement of XAI systems' goodness without solely relying on indirect measures.

Drawing from the motivation described above, the primary objective of XAI systems is to provide information about the functioning of the underlying AI model. Recent surveys (Mohseni et al., 2018) and systematic reviews (Nauta et al., 2022) have highlighted the need for more objective and quantitative measures to assess the goodness of XAI techniques. In this context, we propose merging two objectives: providing information and objectively evaluating its efficacy by defining a measure for the goodness of XAI systems based on the information they offer to users.

Although measuring how much information a system can generate could be challenging, focusing on the amount of *new knowledge* that the explanatory exchange creates in the users' mental models offers a viable approach. By allowing only non-expert users to interact with XAI systems, the acquired knowledge can be attributed to the system interaction. If quantifiable, this knowledge becomes a basis for objectively assessing how informative the system is. This paper proposes an assessment task to objectively and quantitatively measure the goodness of XAI systems during human–AI decision-making, specifically evaluating their informativeness to non-expert users. Our proposal includes theoretical foundations, a mathematical model for the assessment framework, a running example, and an implementation of the assessment task.

2 Related works

Several works have conducted assessments of the properties of Explainable Artificial Intelligence (XAI) systems through user studies, with a notable focus on non-expert users who can benefit from these technologies (Janssen et al., 2022). These latter used tasks that require some skills to be performed correctly. For instance, Lage et al. (2019) exploited aliens' food preferences and clinical diagnosis, while Wang and Yin (2021) delved into recidivism prediction and a forest cover prediction task. Additionally, van der Waa et al. (2021) used a diabetes self-management use-case, where naive users and the system collaborated to find the optimal insulin doses for meals. Furthermore, image classification tasks, such as recognizing bird species, have been investigated by both Goyal et al. (2019) and Wang and Vasconcelos (2020).

Most user studies in the XAI field regard decision-making (Wang and Yin, 2021) or classification tasks (Goyal et al., 2019). Our focus aligns with decision-making, driven by the promise of improved performance when coupling human users with expert AI systems (Wang and Yin, 2022). While this promise is generally kept, some studies show decreased team performance when using certain forms of XAI (Schemmer et al., 2022b). Notably, issues arise from people's difficulty ignoring incorrect AI advice (Schemmer et al., 2022a; Ferreira and Monteiro, 2021; Janssen et al., 2022).

To ensure that AI's advice benefits human users, we need to focus on rigorously assessing the quality of explanations. Furthermore, robust assessment strategies allow researchers and practitioners to compare different XAI strategies. Several works in the XAI field concerning the assessment or comparison of XAI methods tend to define their own measure of goodness (van der Waa et al., 2021; Lage et al., 2019). A recent method proposed to objectively measure the *degree of explainability* of the information provided by an XAI system through an algorithm quantifying the system's ability to answer *archetypical* questions (Sovrano and Vitali, 2022). They assume that the information under study is "good" if it can address all such questions and form the basis of the explanation provided by the XAI system. Another approach

introduced the *System Causability Scale* (Holzinger et al., 2020) to measure the quality of the explanations based on their notion of causability (Holzinger et al., 2019). This scale seeks to account for the completeness of an explanation in the sense that it is accurate on its own (Gilpin et al., 2018). Wang and Yin (2022) adopted a different perspective, comparing various XAI types in different contexts concerning three desiderata: to improve people's understanding of the AI model, help people recognize the model uncertainty, and support people's calibrated trust in the model. They reveal that most current XAI methods fail to satisfy these properties when applied to decision-making with non-expert users.

A comprehensive review by Vilone and Longo (2021) categorizes evaluation methods for XAI into two main groups: objective and human-centered evaluations. The first group includes objective metrics and automated approaches to evaluate explainability, while the second contains human-in-the-loop strategies that exploit users' feedback and judgments. Similar categorizations appear in the works of Bibal and Frénay (2016) and Preece (2018).

Chander and Srinivasan (2018) introduced the notion of "cognitive value" of explanations to objectively describe the effectiveness of different types of XAI within a given context. Moreover, Sokol and Flach (2020) designed the *Explainability Fact Sheets* to systematically classify explainable approaches, facilitating the grasp of XAI methods' capabilities and limitations. More technical approaches have been taken in the realm of image classification. Rio-Torto et al. (2020) proposed an explanation assessment metric for convolution neural networks. Their assessment, called Percentage of Meaningful Pixels Outside the Mask (POMPOM), has been designed for image classification and measures the number of meaningful pixels outside the region of interest in proportion to its total number of pixels. Moreover, Samek et al. (2017) proposed a general objective quality methodology for explanations based on perturbation; they focused on evaluating ordered collections of pixels, such as heatmaps.

Zhang and Zhu (2018) presented two quantitative evaluation metrics to assess the interpretability of visual explanation methods for neural networks, which use semantics annotated by humans on testing images and check if the network locates the relevant part of the same object in different images, respectively. Two alternative metrics for visual explanations have been developed by Yeh et al. (2019), which operate with the network's inputs and outputs perturbations. To evaluate textual explanations, Barratt (2017) proposed three automated quantitative metrics. The quality of rule-based explanations has been assessed in Ignatiev (2021) by focusing on the input space and the percentage of invalid or redundant rules. In addition, Laugel et al. (2019) presented general evaluation metrics for *post-hoc* XAI, focusing on the risk of generating unjustified counterfactual explanations. While these assessment methods claim objectivity, they heavily depend on technologies and application contexts, and deviating from our proposal, they do not involve human input in their functioning.

In addition to the above-mentioned objective methods, Vilone and Longo (2021) reviewed several assessment methods with a human-in-the-loop approach. A significant number of those works used self-reported questionnaires to collect participants' impressions of different XAI techniques with learning systems (Aleven and Koedinger, 2002), intelligent agents (Hepenstal and McNeish, 2020), and neural networks (Weitz et al., 2021). Other scholars preferred to let users interact with the system and collect the explanations' impact on reliability, trust, and reliance of users (Dzindolet et al., 2003), or the system's degree of persuasiveness (Dragoni et al., 2020).

Similarly to our approach, several methods assess end-users' capability to develop mental models of AI internal processes (Harbers et al., 2010a,b; Poursabzi-Sangdeh et al., 2021) and measuring the systems' degree of interpretability by considering the ease with which users replicated the AI functioning. Kaur et al. (2020) performed a longitudinal study to investigate how explainability methods help data scientists understand machine learning models. They qualitatively assessed the data scientists' capability to describe the visualization output of those interpretability tools accurately and discussed the implications for researchers and tool designers.

Moreover, Sanneman and Shah (2020) measured users' situation awareness while interacting with autonomous agents, evaluating whether the explanations contained enough information to allow the users to perform their tasks. Hence, the authors defined different levels to classify users' situation awareness. In particular, Level 1 regards "what an AI system did," Level 2 answers "why an AI system acted in a certain way," and Level 3 addresses "what an AI system will do next or in similar scenarios." Although this proposal is solidly grounded in human factors literature, it does not account for fine-grained measurements and comparisons within different explainability approaches since it provides a method to classify XAI techniques rather than measuring their goodness with respect to each other.

In addition to the methods mentioned above, in the following sections, we present a novel perspective on assessing explanations' goodness through an objective and quantitative task: XAI *Information Power* (IP). IP refers to the amount of information an XAI system provides about the general functioning underlying the AI model, the reasons behind a particular model's choice, or the system's potential actions in different circumstances. The broad definition of IP accommodates different scenarios, emphasizing not just the quantity but also the correctness and timing of the information provided.

3 Methods

3.1 Measuring explanations' quality via information power

The evaluation of an XAI model's Information Power (IP) necessitates the interaction between non-expert users and the system. First, the experimenter needs to design an environment governed by rules and a task that is used during the assessment (for details about the task, see Section 4.1 and subsequent).

During this assessment, the users aim to learn as many system rules as possible. Thus, we need to collect measures for each rule and combine them to obtain the model's information power. The general assessment steps are the following:

1. Quantify the number of rules related to the task (e.g., each feature) and define a method for measuring the number of learned rules.
2. Quantify the informative weight of each rule. For simplicity, we can assume equal weights ($\frac{1}{k}$, where k is the number of rules). The informative weights must satisfy the following constraint $\forall j \in \{1, ..., k\}, \sum_j \gamma_j = 1$. This weight describes the difficulty of understanding the rules regarding different aspects of the task.
3. Measure the model's IP for each user during a training phase, where non-expert users aim to learn the task with the help of the (X)AI system. Further, a set of secondary descriptive measures (e.g., users' satisfaction and their perception about the explanations' usefulness and intrusiveness) could be collected as well.
4. Assess the non-expert users' performance and behavior in a post-training phase without assistance from the (X)AI system.
5. Evaluate non-expert users' knowledge about the task using a set of tests.
6. Average these measures to obtain the final results. Secondary descriptive measures may also hold valuable insights.

Hence, if $k \in \mathbb{N}$ is the number of the environment's features, $\gamma_j \in [0, 1]$ is the informative weight of the feature j, $n_j^r \in \mathbb{N}$ is the number of rules regarding the feature j, $n_j^{lr(i)} \in \mathbb{N}$ is the number of rules regarding the feature j learned by the user i. We also included the AI model's accuracy to consider the goodness of the AI suggestions during the IP measurement: *i.e.*, better models cause higher IP. Thus, $a_m \in [0, 1]$ represents the accuracy of the AI model m, then the informative power of the model m for the user i is computed as follows:

$$IP_i(m) = a_m \sum_{j=1}^{k} \gamma_j \left(\frac{n_j^{lr(i)}}{n_j^r} \right) \in [0, 1]. \quad (1)$$

If n^p is the number of users who took part in the assessment, the IP of the model m is

$$IP(m) = \frac{1}{n^p} \sum_{i=1}^{n^p} IP_i(m) \in [0, 1]. \quad (2)$$

Apart from the number of rules regarding each feature, a delicate aspect of the assessment regards the definition of the information weights. It may be the case that some rules are explicit and easy to learn, while others are more hidden and hard to get. In such cases, it may be relevant to mathematically consider the difficulty of learning those and reach a more precise measure of informativeness of the system by weighing more the features referring to the implicit rules than those referring to explicit ones. We suggest at least two ways to set them: making them equal or defining the weights using experimental data. The former is more straightforward, and we followed this approach in our task (Section 4.1). Alternatively, features' information weight can be set by normalizing the number of user interactions with the system to understand those features. To this end, an *a priori* method is needed to understand which features each user interaction referred to. For example, if the interaction between users and the system stopped at their suggestion request, one can assume that such an interaction regarded the feature affecting the suggestion. Alternatively, suppose such an interaction continued with a user request for an explanation. In that case, one can assume the interaction regarded the feature mentioned in the explanation.

3.1.1 Experimental measures

During the third step of the assessment, several quantitative measures need to be performed to compute the model's information power:

- Performance measures, such as the users' final score.
- Rules understanding measures, such as the number of task rules learned, the number of requests and interactions users needed to learn such rules, and the number of correct answers to the post-experiment test.
- Generalization measures, such as the number of correct answers to *what-if* questions about the agent's decisions in particular environment states.

The first two measures relate to the particular rules explained by the XAI system and their effects on users' ability to perform the task. Higher scores in these measures correspond to higher levels of IP. In contrast, the third measure regards the users' ability to generalize the system's behavior to different situations, answering whether interacting with the explainable agent enabled users to learn the task rules and apply them to unmet situations or similar contexts.

Subjective measures that may also be collected are:

- Satisfaction measures, such as users' satisfaction level with the explanations and the interaction.
- AI perception measures, such as users' feelings toward the system and perception of it.

The first indicates users' willingness to reuse the system and satisfaction with the explanations received. The second regards users' feelings that might arise during the interaction with the system.

4 Running example

To exemplify our method, we provide a running example and a functioning implementation of the assessment task. The NPP environment and models, alongside the multiple-choice test in the appendix, represent a possible implementation of the IP assessment method. The code that refers to its Python implementation is available on GitLab[1]: research can replace the proposed AI/XAI models with their own to test their informativeness.

4.1 The task

The assessment consists of a decision-making task where users can interact with a control panel to perform actions in a simulated environment (see Section 4.2). During the task, users can interact

1 https://gitlab.iit.it/mmatarese/npp-gym

with an expert AI agent by asking *what* it would do, and its XAI system by asking *why* it would do that. Users start the task without knowledge about it: within a fixed time or number of steps, they must perform actions and interact with the agent to discover the task's goals, rules used by the AI model for actions, and rules governing the simulated environment.

4.1.1 Interaction modalities

The agent cannot perform actions: its role is limited to assisting users during decision-making. The agent can not take the initiative to give suggestions either, but it always answers users' questions. Thus, only the users can interact with the control panel and act in the simulated environment.

4.1.2 Characteristics of the task

We considered non-expert users who passed through the interaction and obtained new information. We considered only participants without knowledge about the task and its underlying rules. For this reason, we implemented a simulation of a nuclear power plant (NPP) management task because it is challenging and engaging for non-expert users, governed by relatively simple rules, and generally, people know nothing about the functioning of NPPs.

The main objectives of the task are to generate as much energy as possible and maintain the system in equilibrium. The environment's features are subject to rules and constraints, which we can summarize as follows:

- Each action corresponds to an effect on the environment, thus changing its features' value.
- Specific preconditions must be satisfied to start and continue nuclear fission and produce energy.
- Some conditions irremediably damage the plant.
- The task is divided into steps in which the users can interact with either the agent or the control panel.

4.2 The environment

We modeled the nuclear power plant as a reinforcement learning (RL) environment using the OpenAI Gym API (Brockman et al., 2016).

4.2.1 Features of the environment

The simulated power plant is composed of four continuous features:

- Pressure in the reactor's core.
- Temperature of the water in the reactor.
- Amount of water in the steam generator.
- Reactor's power.

Furthermore, the power plant has four other discrete features that regard the reactor's rods: security rods, fuel rods, sustain rods, and regulatory rods. The first two have two levels: up and down. Instead, the latter two have three levels: up, medium, and down.

The reactor power linearly decreases over time for the effect of the de-potentiation of the fuel rods. Hence, the reactor's power depends on the values of the environment's features and whether nuclear fission is taking place. Moreover, the energy produced at each step is computed by dividing the reactor's power by 360, which is the power that a 1,000MW reactor without power dispersion produces in 10 seconds (the expected time duration of a step).

4.2.2 Actions to perform on the environment

Users can perform twelve actions, including changing rod positions, adding water to the steam generator, or skipping to the next step. The actions alter three parameters representing the water's temperature in the core, the core's pressure, and the water level in the steam generator. The setting of the rods determines the entity of feature updates. Such updates are performed at the end of each step, right after the users' action.

For example, if the safety rods are lowered in the reactor's core, the nuclear fission stops; thus, the temperature and pressure decrease until they reach their initial values, and the water in the steam generator remains still. On the other hand, if nuclear fission occurs and the user lowers the regulatory rods, the fission accelerates. This acceleration consumes more water in the steam generator, raising the core's temperature and pressure more quickly, but also increasing the reactor's power and the electricity produced. If the users do not act within the time provided for each step, the application automatically chooses a *skip* action, which applies the features' updates based on the setting of the rods at hand.

4.3 The agent's AI

We trained a deterministic decision tree (DT) using the Conservative Q-Improvement (CQI) learning algorithm (Roth et al., 2019), which allowed us to build the DT using an RL strategy. Instead of extracting the DT from a more complex ML model (Vasilev et al., 2020; Xiong et al., 2017), we used this learning strategy to simplify the translation from the AI to the XAI without losing performance. The agent uses this expert DT to choose its action: it can perform each of the 12 actions based on the eight environment's features.

CQI learns a policy as a DT by splitting its current nodes only if it represents a policy improvement. Leaf nodes correspond to abstract states and indicate the action to be taken, while branch nodes have two children and a splitting condition based on a feature of the state space. Over time, their algorithm creates branches by replacing existing leaf nodes if the final result represents an improved policy. In this sense, the algorithm is considered additive, while it is conservative in performing the splits (Roth et al., 2019).

Starting from its root node, the DT is queried on each of its internal nodes - representing binary splits - to decide in which of the two sub-trees continue the descent. Each internal node regards a feature x_i and a value for that feature v_i: the left sub-tree contains instances with values of $x_i \leq v_i$, while the right sub-tree contains instances with values of $x_i > v_i$ (Buhrman and de Wolf, 2002).

The DT's leaf nodes represent actions; in the implementation of Roth et al. (2019), they are defined with an array containing the

actions' expected Q-values: the greater Q-value is associated with the most valuable action. This way, the DT can be queried by users with both *what* and *why* questions. To answer a *what* question, we only need to navigate the DT using the current values of the environment's features and present the resulting action to the user.

4.4 The agent's XAI

As we already saw in Section 4.1, answering *why* questions regards the agent XAI. Since the AI model to explain is already transparent (Adadi and Berrada, 2018), it also constitutes the XAI model. We can use DTs to provide explanations simply by using one (or more) of the feature values we encounter during the descent.

As we have seen in Section 4.3, during the DT descent, we encounter a set of split nodes defined by a feature x_i and a value v_i; the direction of the descent tells us if the current scenario has a value of $x_i \leq v_i$ or $x_i > v_i$. Each of those inequalities can be used to provide an explanation that can help users relate actions with specific values of the environment's features. In our case, an explanation for the action "add water to the steam generator" could be "because the water level in the steam generator is ≤ 25," which is dangerously low.

Which of the features to use among the ones encountered during descent is a problem called "explanation selection." In our case, the selection of explanations n depends on the XAI strategy we want to test. For example, classical approaches use the most relevant features [in terms of the Gini index, information gain, or other well-established measures (Stoffel and Raileanu, 2001)].

The XAI strategy we provided in the code explains using only the AI outcomes and the environment's states. In particular, it justifies the agent's suggestions using the most relevant features, the first ones in the DT's structure (see Roth et al., 2019). The system always tries to give different explanations by keeping track of the DT's nodes already used and progressively choosing the others to decrease the level of relevance.

5 Discussion

The proposed assessment task satisfies properties unique to the explainable artificial intelligence (XAI) research field. Firstly, it focuses on the utility of explanations for the users. Then, it defines the goodness of the XAI system as the amount of information that an explainable agent can provide to them. This allows for an objective and quantitative analysis of such goodness, possibly making precise comparisons between different explainability strategies.

We grounded the assessment in a decision-making setup because it represents the most used context for human-AI collaboration and presents a high potential for the future. Throughout the paper, we stressed one of the fundamental characteristics of the Information Power (IP) assessment method: putting the human in the loop. In Section 2, we have already seen that the XAI community has recognized it as crucial to consider the end-user in the XAI evaluation process. However, most proposed approaches still rely on subjective measures (self-reported questionnaires) or consider only specific application contexts. In contrast, our approach offers an objective assessment that could be applied to a range of different application contexts.

The IP assessment task focuses on objective and quantitative measures of the explanations' goodness; however, it can be enriched by several secondary and quantitative measures. Such secondary measures could include users' perception of the explanations' usefulness and intrusiveness or their satisfaction and willingness to reuse them (Conati et al., 2021). We can consider this flexibility to provide for objective, subjective, quantitative, and qualitative measurements of the explanation's goodness, the last key element of our proposal. This can be reached through the post-training phase, in which users are asked to perform the task at their best and with the final test, which may contain both multiple-choice and open-ended questions. The training, post-training, and test phases, and the consequent variety of measures that can be collected, give the IP assessment task unique characteristics in the panorama of the assessment approaches for explainable systems.

In Table 1, we first present all the requirements that were decisive for our approach in the line "IP" and then compare our assessment framework with those mentioned in Section 2 with respect to these requirements. We consider this summary as important to push forward discussions about measurements and assessment in XAI.

The following list clarifies the meaning of the eight requirements mentioned in the table:

- Decision-making: the assessment is encapsulated in a decision-making task, where users are asked to make decisions that could be correct or wrong.
- Human–AI collaboration: the assessment provides for collaborative scenarios in which users can benefit from AI expertise.
- (X)AI model agnostic: the assessment applies to all (or a large majority) AI models and XAI techniques.
- Objective: the assessment includes objective variables.
- Subjective: the assessment includes subjective variables.
- Quantitative: the assessment provides quantitative measures for the objective and subjective variables.
- Qualitative: the assessment provides qualitative measures for the objective and subjective variables.
- Human-in-the-loop: the assessment follows a human-in-the-loop approach, strongly relying on user studies.

Interestingly, whereas we pointed out that it is difficult to find objective measures for quality in the XAI literature, the same is valid for human studies concerning the understanding they gain from an explanation. It is important to measure knowledge as an effect of an explanation, and the discussion of what to consider as different forms of understanding (varying from being enabled to do something to deeply comprehending a matter) just started (Buschmeier et al., 2023).

Concerning user interaction with an XAI, the above-presented assessment task with those characteristics is flexible enough to test different AI models and XAI techniques as long as they allow interaction between the user and the system and the AI

TABLE 1 Comparison between the *IP* assessment framework and other XAI goodness' assessment methods with respect to eight requirements.

Methods	Decision-making	Human-AI collab.	(X)AI model agnostic	Objective	Subjective	Quantit.	Qualit.	Human-in-the-loop
IP	✓	✓	✓	✓	✓	✓	✓	✓
van der Waa et al. (2021)	✓	✓	✗	✗	✓	✓	✓	✓
Lage et al. (2019)	✗	✗	✗	✗	✓	✓	✗	✓
Sovrano and Vitali (2022)	✓	✓	✓	✓	✗	✓	✗	✗
Holzinger et al. (2020)	✓	✓	✗	✗	✓	✗	✓	✓
Chander and Srinivasan (2018)	✗	✗	✓	✓	✗	✗	✗	✗
Rio-Torto et al. (2020)	✗	✗	✗	✓	✗	✓	✗	✗
Samek et al. (2017)	✗	✗	✗	✓	✗	✓	✗	✗
Zhang and Zhu (2018)	✗	✗	✗	✓	✗	✓	✗	✗
Yeh et al. (2019)	✗	✗	✗	✓	✗	✓	✗	✗
Barratt (2017)	✗	✗	✗	✓	✗	✓	✗	✗
Ignatiev (2021)	✗	✗	✗	✓	✗	✓	✗	✗
Laugel et al. (2019)	✗	✗	✓	✓	✗	✓	✗	✗
Sanneman and Shah (2020)	✗	✗	✓	✓	✗	✗	✓	✓

All the requirements refer to the assessment phase. The meaning of the requirement is explained in the Section 5 (e.g., "decision-making" means that the assessment method provides for a decision-making task).

10.3389/fcomp.2024.1412341

model (implicitly or explicitly) aligns with the tasks' rules: we assume that as long as the task and environment reflect the characteristics presented in Section 4.1 and subsequent, and the AI can master the task, it cannot be that the tasks' rules are not reflected by the AI's actions. However, there are cases of imprecise mapping or explanation presentation techniques that do not explicitly take the form of rules (as in our example), e.g., using visual elements (LIME) or features importance (SHAP). In such cases, our framework can still be used to assess their capacity to implicitly refer to the environment's rules, e.g., by comparing them with more explicit techniques. Moreover, the IP framework can be used to objectively compare two or more HCI/HRI approaches to test whether interactive dynamics ease the users' understanding of the agent's suggestions and explanations. In particular, to reproduce the IP assessment in a different scenario, researchers need the following:

- A decision-making task with the same characteristics as the one presented in Section 4.1.
- An expert AI and several non-expert users to whom administer the assessment task.
- Two or more approaches to compare, such as XAI algorithms or HCI/HRI dynamics.
- At least one quantitative measure about users' understanding of the task; if two or more, also a method to compact them into a single measure.
- At least one quantitative measure about users' ability to generalize to unseen scenarios.

Therefore, the decision-making task needs to be submitted to non-expert users, giving them the objective of learning the functioning of the task. During the task, participants should interact with an expert (X)AI model (under investigation) and ask it for help during the learning. Right after the learning phase, there should be an assessment phase in which the users must prove their understanding of the task by performing it at their best. Finally, the experimenter should submit participants to the test to objectively measure their knowledge of the task. An assessment task with those characteristics is flexible enough to test different AI models and XAI techniques as long as they allow user and system interaction.

It has to be noted that non-expert users' prior capabilities and characteristics may differ and somehow bias the results. However, such asymmetries between users can be measured before the assessment and involved in the analysis. For example, users' prior knowledge about the overall functioning of NPPs and their need for cognition [the tendency to reflect on things before acting (Cacioppo and Petty, 1982)] may impact the final results. Those characteristics may be measured before the task administration and considered while computing the system's IP, e.g., by dividing users into groups according to their prior knowledge and need for cognition.

5.1 Limitations

The most critical requirement of our assessment task, which both uniquely characterizes and limits its possible use cases, is the interaction with the user. In particular, it allows the users to query the system by asking what it would do in a specific situation and why. Consequently, the XAI system should be able to answer both *what* and *why* questions to exploit the full potential of the assessment task. However, exploiting our assessment task is impossible if such an interaction is impossible (or very circumscribed). In these cases, we recommend using other metrics, such as the Degree of Explainability by Sovrano and Vitali (2022).

Moreover, the IP framework needs deterministic systems and non-expert users: both constraints limit the application of our assessment method. While the former constraint can be overcome by selecting participants according to their levels of prior knowledge about the task, it is impossible to assess non-deterministic algorithms with the IP framework because the tasks' rules need to be well-defined, as described in Section 4.1.

Another area for improvement of our approach is that, as a method of quantification, it disregards the fact that in the decision-making process, some actions might be more difficult than others. The same is true for the features that the XAI reveals; with some of them, users might already have experiences that yield more understanding than others, which might be new for specific cases. These contextual factors regarding the user's experience and knowledge should be considered in future research to develop adapted or adapting measurements for the quality of XAI that is relevant for the users. This is consistent with Miller (2019), who points out that explanatory power depends on personal relevance.

6 Conclusions

In this work, we proposed an assessment task to objectively and directly measure the goodness of XAI systems in terms of their informativeness. We designed the assessment as an XAI-assisted decision-making task with non-expert users and a final test to assess their understanding of the task itself. Starting from no knowledge about it, during the task users need to understand its rules and objectives in a learning-by-doing approach. The subsequent test aims at measuring their acquired knowledge.

In conclusion, the proposed Information Power assessment task provides a valuable contribution to evaluating XAI systems. Its emphasis on user utility, objectivity, and flexibility positions it as a comprehensive and practical approach for assessing the goodness of explanations in decision-making contexts. As the field of XAI evolves, ongoing research and refinements of assessment methodologies will contribute to advancing our understanding of how AI systems can best provide meaningful and effective explanations to users.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://gitlab.iit.it/mmatarese/npp-gym.

Author contributions

MM: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. FR: Supervision, Writing – review & editing. KR: Supervision, Writing – review & editing. AS: Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that generative AI was used in the creation of this manuscript. While preparing this work, the author(s) used Grammarly and GPT 3.5 to check the grammar and improve the manuscript's readability. After using these tools, the author(s) reviewed and edited the content as needed and took full responsibility for the publication's content.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Aleven, V. A., and Koedinger, K. R. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based cognitive tutor. *Cogn. Sci.* 26, 147–179. doi: 10.1207/s15516709cog2602_1

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., et al. (2022). What we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion* 99:101805. doi: 10.1016/j.inffus.2023.101805

Barratt, S. (2017). Interpnet: neural introspection for interpretable deep learning. *arXiv* [Preprint]. arXiv:1710.09511. doi: 10.48550/arXiv.1710.09511

Bibal, A., and Frénay, B. (2016). "Interpretability of machine learning models and representations: an introduction," in *24th European symposium on artificial neural networks, computational intelligence and machine learning* (CIACO) (The European Symposium on Artificial Neural Networks), 77–82. Available at: https://www.semanticscholar.org/paper/Interpretability-of-machine-learning-models-and-an-Bibal-Fr%C3%A9nay/464656fc6431f1db8b2e0b0b3093a5df1cb7958e

Booshehri, M., Buschmeier, H., Cimiano, P., Kopp, S., Kornowicz, J., Lammert, O., et al. (2024). "Towards a computational architecture for co-constructive explainable systems," in *Proceedings of the 2024 Workshop on Explainability Engineering (ExEn'24)* (New York, NY: ACM). doi: 10.1145/3648505.3648509

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., et al. (2016). Openai gym. *arXiv* [Preprint]. arXiv:1606.01540. doi: 10.48550/arXiv.1606.01540

Buhrman, H., and de Wolf, R. (2002). Complexity measures and decision tree complexity: a survey. *Theoret. Comput. Sci.* 288, 21–43. doi: 10.1016/S0304-3975(01)00144-X

Buschmeier, H., Buhl, H. M., Kern, F., Grimminger, A., Beierling, H., Fisher, J., et al. (2023). Forms of understanding of xAI-explanations. *arXiv* [Preprint]. arXiv:2311.08760. doi: 10.48550/arXiv.2311.08760

Cacioppo, J. T., and Petty, R. E. (1982). The need for cognition. *J. Pers. Soc. Psychol.* 42:116. doi: 10.1037/0022-3514.42.1.116

Chander, A., and Srinivasan, R. (2018). "Evaluating explanations by cognitive value," in *Machine Learning and Knowledge Extraction*, eds. A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl (Cham: Springer International Publishing), 314–328. doi: 10.1007/978-3-319-99740-7_23

Cohausz, L. (2022). "Towards real interpretability of student success prediction combining methods of xAI and social science," in *Proceedings of the 15th International Conference on Educational Data Mining*, 361–367.

Conati, C., Barral, O., Putnam, V., and Rieger, L. (2021). Toward personalized xAI: a case study in intelligent tutoring systems. *Artif. Intell.* 298:103503. doi: 10.1016/j.artint.2021.103503

Dragoni, M., Donadello, I., and Eccher, C. (2020). Explainable AI meets persuasiveness: translating reasoning results into behavioral change advice. *Artif. Intell. Med.* 105:101840. doi: 10.1016/j.artmed.2020.101840

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud.* 58, 697–718. doi: 10.1016/S1071-5819(03)00038-7

Embarak, O. H. (2022). Internet of behaviour (IOB)-based AI models for personalized smart education systems. *Procedia Comput. Sci.* 203, 103–110. doi: 10.1016/j.procs.2022.07.015

Ferreira, J. J., and Monteiro, M. (2021). The human-AI relationship in decision-making: AI explanation to support people on justifying their decisions. *arXiv* [Preprint]. arXiv:2102.05460. doi: 10.48550/arXiv.2102.05460

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., Kagal, L., et al. (2018). Explaining explanations: an overview of interpretability of machine learning. *arXiv* [Preprint]. arXiv:1806.00069. doi: 10.48550/arXiv.1806.00069

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S., et al. (2019). "Counterfactual visual explanations," in *Proceedings of the 36th International Conference on Machine Learning* (PMLR), 97, 2376–2384.

Harbers, M., Broekens, J., Van Den Bosch, K., and Meyer, J.-J. (2010a). "Guidelines for developing explainable cognitive models," in *Proceedings of ICCM* (Citeseer), 85–90. Available at: https://www.researchgate.net/publication/48320823_Guidelines_for_Developing_Explainable_Cognitive_Models

Harbers, M., van den Bosch, K., and Meyer, J.-J. (2010b). "Design and evaluation of explainable BDI agents," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol. 2* (New York, NY: ACM), 125–132. doi: 10.1109/WI-IAT.2010.115

Hepenstal, S., and McNeish, D. (2020). "Explainable artificial intelligence: what do you need to know?" in *Augmented Cognition. Theoretical and Technological Approaches*, eds. D. D. Schmorrow, and C. M. Fidopiastis (Cham: Springer International Publishing), 266–275. doi: 10.1007/978-3-030-50353-6_20

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv* [Preprint]. arXiv:1812.04608. doi: 10.48550/arXiv.1812.04608

Holzinger, A., Carrington, A., and Müller, H. (2020). Measuring the quality of explanations: the system causability scale (SCS). *Künstliche Intelligenz* 34, 193–198. doi: 10.1007/s13218-020-00636-z

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* 9:e1312. doi: 10.1002/widm.1312

Ignatiev, A. (2021). "Towards trustable explainable AI," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI'20* (International Joint Conferences on Artificial Intelligence Organization). doi: 10.24963/ijcai.2020/726

Islam, M. R., Ahmed, M. U., Barua, S., and Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Appl. Sci.* 12:1353. doi: 10.3390/app12031353

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., and Kuk, G. (2022). Will algorithms blind people? the effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. *Soc. Sci. Comput. Rev.* 40, 478–493. doi: 10.1177/0894439320980118

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman Vaughan, J., et al. (2020). "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20* (New York, NY: Association for Computing Machinery), 1–14. doi: 10.1145/3313831.3376219

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., et al. (2019). An evaluation of the human-interpretability of explanation. *arXiv* [Preprint]. arXiv:1902.00006. doi: 10.48550/arXiv.1902.00006

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., and Detyniecki, M. (2019). "The dangers of *post-hoc* interpretability: unjustified counterfactual explanations," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI'19* (Philadelphia, PA: AAAI Press), 2801–2807. doi: 10.24963/ijcai.2019/388

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends Cogn. Sci.* 20, 748–759. doi: 10.1016/j.tics.2016.08.001

Matarese, M., Rea, F., and Sciutti, A. (2021). A user-centred framework for explainable artificial intelligence in human-robot interaction. *arXiv* [Preprint]. arXiv:2109.12912. doi: 10.48550/arXiv.2109.12912

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Mohseni, S., Zarei, N., and Ragan, E. D. (2018). A survey of evaluation methods and measures for interpretable machine learning. *arXiv* [Preprint] arXiv:11839.18111. doi: 10.48550/arXiv.11839.18111

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., et al. (2022). From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *arXiv* [Preprint]. arXiv:2201.08164. doi: 10.48550/arXiv.2201.08164

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21* (New York, NY: Association for Computing Machinery). doi: 10.1145/3411764.3445315

Preece, A. (2018). Asking "why" in AI: explainability of intelligent systems-perspectives and challenges. *Intell. Sys. Acc. Fin. Mgmt.* 25, 63–72. doi: 10.1002/isaf.1422

Rio-Torto, I., Fernandes, K., and Teixeira, L. F. (2020). Understanding the decisions of CNNS: an in-model approach. *Pattern Recognit. Lett.* 133, 373–380. doi: 10.1016/j.patrec.2020.04.004

Roth, A. M., Topin, N., Jamshidi, P., and Veloso, M. (2019). Conservative q-improvement: reinforcement learning for an interpretable decision-tree policy. *arXiv* [Preprint]. arXiv:1907.01180. doi: 10.48550/arXiv.1907.01180

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 2660–2673. doi: 10.1109/TNNLS.2016.2599820

Sanneman, L., and Shah, J. A. (2020). "A situation awareness-based framework for design and evaluation of explainable AI," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, eds. D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling (Cham: Springer International Publishing), 94–110. doi: 10.1007/978-3-030-51924-7_6

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., and Satzger, G. (2022a). Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv* [Preprint]. arXiv:2204.06916. doi: 10.48550/arXiv.2204.06916

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., and Vössing, M. (2022b). A meta-analysis on the utility of explainable artificial intelligence in human-AI decision-making. *arXiv* [Preprint] arXiv:2205.05126. doi: 10.48550/arXiv.2205.05126

Sokol, K., and Flach, P. (2020). "Explainability fact sheets: a framework for systematic assessment of explainable approaches," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20* (New York, NY: Association for Computing Machinery), 56–67. doi: 10.1145/3351095.3372870

Sovrano, F., and Vitali, F. (2022). "How to quantify the degree of explainability: experiments and practical implications," in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (Padua: IEEE), 1–9. doi: 10.1109/FUZZ-IEEE55066.2022.9882574

Stange, S., Hassan, T., Schröder, F., Konkol, J., and Kopp, S. (2022). Self-explaining social robots: an explainable behavior generation architecture for human-robot interaction. *Front. Artif. Intell.* 87:866920. doi: 10.3389/frai.2022.866920

Stoffel, K., and Raileanu, L. E. (2001). "Selecting optimal split-functions for large datasets," in *Research and Development in Intelligent Systems XVII* (Cham: Springer), 62–72. doi: 10.1007/978-1-4471-0269-4_5

van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. (2021). Evaluating xAI: a comparison of rule-based and example-based explanations. *Artif. Intell.* 291:103404. doi: 10.1016/j.artint.2020.103404

Vasilev, N., Mincheva, Z., and Nikolov, V. (2020). "Decision tree extraction using trained neural network," in *Proceedings of the 9th International Conference on Smart Cities and Green ICT Systems - SMARTGREENS* (SciTePress), 194–200. doi: 10.5220/0009351801940200

Vilone, G., and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* 76, 89–106. doi: 10.1016/j.inffus.2021.05.009

Wang, P., and Vasconcelos, N. (2020). "Scout: self-aware discriminant counterfactual explanations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 8981–8990. doi: 10.1109/CVPR42600.2020.00900

Wang, X., and Yin, M. (2021). "Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making," in *International Conference on Intelligent User Interfaces* (New York, NY: ACM), 318–328. doi: 10.1145/3397481.3450650

Wang, X., and Yin, M. (2022). "Effects of explanations in AI-assisted decision making: principles and comparisons," in *ACM Transactions on Interactive Intelligent Systems (TiiS)* (New York, NY: ACM). doi: 10.1145/3519266

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., and André, E. (2021). "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *J. Multimodal User Interfaces* 15, 87–98. doi: 10.1007/s12193-020-00332-0

Williams, O. (2021). *Towards human-centred explainable AI: A systematic literature review*. doi: 10.13140/RG.2.2.27885.92645

Xiong, Z., Zhang, W., and Zhu, W. (2017). "Learning decision trees with reinforcement learning," in *31st Conference on Neural Information Processing Systems (NIPS 2017)* (NIPS Workshop on Meta-Learning).

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. (2019). "On the (in)fidelity and sensitivity of explanations," in *Advances in Neural Information Processing Systems, Vol. 32*, eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Red Hook, NY: Curran Associates, Inc).

Zhang, Q.-s., and Zhu, S.-C. (2018). Visual interpretability for deep learning: a survey. *Front. Inf. Technol. Electron. Eng.* 19, 27–39. doi: 10.1631/FITEE.1700808