Int. J. Human–Computer Studies 199 (2025) 103484

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

The role of user feedback in enhancing understanding and trust in counterfactual explanations for explainable AI

Muhammad Suffian [a,c,*], Ulrike Kuhl [b], Alessandro Bogliolo [c], Jose M. Alonso-Moral [d]

[a] Department of Information, Infrastructure and Sustainable Energy Engineering (DIIES), Università degli Studi Mediterranea di Reggio Calabria, Via dell'Università, 25, Reggio Calabria, 89124, Calabria, Italy
[b] Research Institute for Cognition and Robotics, Bielefeld University, Inspiration 1, Bielefeld, 33615, Bielefeld, Germany
[c] Department of Pure and Applied Sciences, Universita degli Studi di Urbino Carlo Bo, Piazza della Republica, 13, Urbino, 61029, PU, Italy
[d] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, s/n, Santiago de Compostela, 15782, Galicia, Spain

ARTICLE INFO

Keywords:
Explainable AI
Human-centered explanations
Counterfactual explanations
Human behavioral analytics
User study

ABSTRACT

Counterfactual explanations (CEs) have emerged as a viable solution for generating comprehensible explanations in the context of explainable artificial intelligence (XAI). A CE provides actionable information to users on how to achieve the desired outcome from a machine learning (ML) model with minimal modifications to the input. XAI is crucial for improving transparency and reliability in AI systems, especially for meeting regulations like the General Data Protection Regulation (GDPR) or the European AI Act. However, the integration of CEs into XAI frameworks and their effectiveness in enhancing user trust and cognitive learning remains uncertain and requires further research. We have developed a user study to face this challenge with two user input-driven counterfactual generation XAI approaches: (i) User Feedback-based Counterfactual Explanation (UFCE) and (ii) Diverse Counterfactual Explanation (DiCE). They are integrated within a game-inspired online platform that enables direct comparisons between them. We compared the task performance, understanding, satisfaction, and trust between control and experimental groups, with a total of 101 participants. After curating the collected data, we had 70 users (24 in the control group) who successfully completed the experiment. Participants in the experimental group received explanations generated by UFCE or DiCE. Findings show that explanations generated by UFCE improve users' learning experiences, resulting in better task performance, comprehension, satisfaction, and trust. Moreover, participants who interacted with UFCE exhibited significantly higher reliance on suggestions than those who interacted with DiCE, what was supported by statistical validation. These results highlight the significance of human-centered XAI methods and promote meaningful cognitive engagement for users. Furthermore, the game-inspired platform is implemented as open-source to promote Open Science, and it is made publicly available along with data collected in the user study to support further investigations and to ensure reproducibility of reported results.

1. Introduction

The field of explainable artificial intelligence (XAI) seeks to interpret and explain artificial intelligence (AI) models to increase their transparency, fairness, and trustworthiness (Ali et al., 2023). It is crucial to build trust in AI systems and to understand how they make their decisions. The compliance of AI automated decisions with the General Data Protection Regulation (GDPR) 2016/679 (Voigt and Von dem Bussche, 2017) and the European AI act (Pehlivan, 2024) is crucial. In particular, according to the GDPR, individuals subjected to automated decision-making are entitled to receive an explanation for the decisions made (Goodman and Flaxman, 2017). Consequently, there has been a rise in the development of XAI approaches (Adadi and Berrada, 2018; Ali et al., 2023; Chander et al., 2024).

In the pursuit of advancing XAI, human-centered XAI holds significant promise. Scholars in the XAI research field emphasize the collaborative nature of explanations (Mueller et al., 2021; Suffian et al.,

* Corresponding author at: Department of Information, Infrastructure and Sustainable Energy Engineering (DIIES), Università degli Studi Mediterranea di Reggio Calabria, Via dell'Università, 25, Reggio Calabria, 89124, Calabria, Italy.
E-mail addresses: m.suffian@unirc.it (M. Suffian), ukuhl@techfak.uni-bielefeld.de (U. Kuhl), alessandro.bogliolo@uniurb.it (A. Bogliolo), josemaria.alonso.moral@usc.es (J.M. Alonso-Moral).

[1] In the rest of the paper, we use terms explanations and counterfactual explanations (CE), interchangeably.

https://doi.org/10.1016/j.ijhcs.2025.103484
Received 23 October 2024; Received in revised form 26 February 2025; Accepted 27 February 2025
Available online 14 March 2025

2022). They view the explanation as a reciprocal process, wherein users contribute input and XAI systems offer explanatory details (Hoffman et al., 2023; Stepin et al., 2024; Suffian et al., 2022). It is human nature to engage in counterfactual thinking to envision alternative scenarios while still holding onto the original representation, a process of alternating the related facts (Byrne, 2016). A piece of "contrary-to-fact" information constitutes a counterfactual explanation (CE)[1] (Guidotti, 2022; Stepin et al., 2021). Being contrastive by nature, CE allows us to find a minimal set of features that would have led the AI system to make a different decision. Such CEs are claimed to enhance users' trust in decisions made by AI systems. Building on this understanding, users commonly view explanations framed as counterfactuals as naturally intuitive and readily understandable to humans (Artelt and Hammer, 2021; Dandl et al., 2020; Guidotti et al., 2018; Stepin et al., 2021). However, the smooth integration of these insights to CEs in XAI remains uncertain, highlighting the crucial need to validate technical approaches that offer explainability for AI systems at the user level (Doshi-Velez et al., 2017). Some researchers overlook the significance of the evaluation of their works by formally conducting human studies. A recent review reveals that merely one in three counterfactual XAI papers incorporates user-based evaluations, often lacking in statistical rigor and reproducibility (Keane et al., 2021).

There are a few studies that directly compare user behavior in response to XAI feedback. The user engagement in customizing explanations by providing feedback to explanation generation methods represents a novel and unexplored perspective aimed at enhancing the cognitive learning such as task performance, understanding, satisfaction, and trust of users. The subjective metrics are crucial for evaluating user-centric aspects such as trust, perceived utility, confidence, and ease of learning—elements that cannot be fully captured through quantitative measures alone. While user evaluations play a crucial role in assessing the effectiveness of explanations, crafting a successful user study for control and experimental settings, presents significant challenges. The design of the user study must carefully consider the individuals receiving explanations and the purpose behind providing them (Adadi and Berrada, 2018; Sokol and Flach, 2020), while also addressing potential confounding variables and factors (Doshi-Velez et al., 2017). Hence, there is a noticeable absence of interactive human-centered XAI methodologies tailored for user studies, as well as a lack of domain-general environments to ensure consistency across participants. Since counterfactual explanations are often termed as user-friendly, however, usually discussed theoretically (Stepin et al., 2021), there is limited empirical research exploring how users interact with these explanations in practice (Rong et al., 2023). Our study addresses this gap by providing both quantitative and subjective insights into the user experience. The current user study is motivated by a lack of engaging user study designs that enable direct comparisons between various implementations of CEs. Another motivation is to utilize a domain-independent abstract use case, enabling us to distinguish the findings from the confounding effects of users' knowledge. It is evident that human-centered and user input-driven XAI techniques are crucial for effective evaluations, this user study fills the highlighted gap by employing two user input-driven counterfactual generation XAI approaches, User Feedback-based Counterfactual Explanation (UFCE) (Suffian et al., 2024a) and Diverse Counterfactual Explanation (DiCE) (Mothilal et al., 2020). Both approaches provide users with an opportunity to customize CEs, hence, turning to be user-friendly, actionable, and an aid to solve the tasks in the given task domain. However, their utility in enhancing learning processes by providing explanations in between and comparing with control conditions necessitates thorough examination.

Taking this whole picture into consideration, we conducted a user study involving 101 participants, who played with a game-inspired web-based system (Kuhl et al., 2023b; Suffian et al., 2024b). Unfortunately, only 70 users successfully completed the experiment. Anyway, it is worth noting that this is good number of users in comparison with other studies in the XAI literature, because recruitment of participants for user studies is usually hard. This fact is recognized as one of the main bottlenecks in qualitative research (Archibald and Munce, 2015; Patel et al., 2003). Sometimes, expanding the pool of participants could provide richer insights, but at a high cost. Moreover, it is not only a matter of quantity but mainly a matter of quality. Notice that, a small pool of well-motivated and careful participants is much better than a big pool of careless participants.

Our aim was to explore participants' comprehension of the AI system's functionality, both in terms of their objective performance in tasks and their perceived importance of decision-making factors. Additionally, we analyzed in depth the subjective assessment of the explanations provided by the underlying system, gauging their perception of explanation helpfulness and trustworthiness. The participants were divided into two groups to interact with the predictive system. These groups were the control group and the experimental group. The control group did not receive any explanations while engaging in task-solving activities and the experimental group received explanations generated by the counterfactual methods based on the provided input. The experimental group participants were further divided into two groups, one received the explanations from the UFCE method and the other from the DiCE method. The task activities were game-inspired scenarios where participants provided input to generate the counterfactual actions to achieve improved results in an unfamiliar domain (see Section 3.1). The log of participants' input choices over multiple attempts is recorded to evaluate their learning progress.

Our proposed setup mirrors real-world scenarios where individuals encounter unfamiliar systems and must rely on automated explanations to make informed decisions. For example, in healthcare, patients receiving AI-driven diabetes risk assessments often lack prior medical knowledge, and the counterfactual explanation-based system helps them understand actionable steps for risk reduction. Similarly, in finance, individuals navigating investment decisions may be unfamiliar with risk factors, and automated explanations can guide them toward better financial choices. By designing a task where all participants are novices, we create a controlled environment to study how users interact with explanations, making our setup a valid proxy for real-life applications where explanation methods support learning and decision-making.

Thus, the main goal of the user study is to compare the task performance, understanding, satisfaction, and trust of participants in the system and whether participants find CEs of DiCE or UFCE helpful as compared to control group participants. A secondary goal is to compare DiCE and UFCE in terms of their explanations as actionable, feasible, satisfactory, and trustworthy. By analyzing participants' performance metrics and eliciting feedback on trust and satisfaction levels, we aim to provide insights into how user feedback-based CEs contribute to enriched cognitive learning experiences. Through this endeavor, we aim to inform the design of human-centered XAI systems that not only make decisions transparently but also facilitate meaningful cognitive engagement for users. In general, we are confident that the collaboration between human cognition and human-centered XAI has the potential to drive transformative advancements in cognitive learning.

In summary, the main contributions in this paper are:

1. We design an experiment to evaluate the performance of control and experimental groups (in terms of task performance, understanding, satisfaction, and trust) in a user study.
2. We analyze empirically how two user input-driven CE methods (i.e., UFCE and DiCE) can (or cannot) enhance the cognitive learning of participants in the user study when they are provided with automated explanations versus no explanations. UFCE standed out in terms of satisfaction and trust.
3. We implement the current study as open-source to promote Open Science, and it is made publicly available to support further investigations.

The rest of the paper is structured as follows. Section 2 describes the related work by highlighting the need and value of user evaluations. Section 3 outlines the Alien Nutri-Solver framework — elucidating its implementation, evaluation measures, and experimental setting. Section 4 illustrates the results obtained from the experiments. Section 5 discusses the different factors and limitations, and finally, in Section 6, we emphasize potential applications and the avenues for future research for XAI-driven cognitive learning.

2. Related work

With the increasing number of explanation methods being proposed, researchers are seeking systematic overviews of the growing field of XAI. Several studies, including those cited in Adadi and Berrada (2018), Arrieta et al. (2020), Burkart and Huber (2021), Carvalho et al. (2019), Gilpin et al. (2018), Samek and Müller (2019) cover various aspects of XAI technologies, such as problem definitions, goals, AI/ML model explanations, and evaluation measures. In addition, the study in Abdul et al. (2018) highlights research trends and challenges in Human–Computer Interaction (HCI) applications. Many XAI surveys concentrate on the interpretability of specific model families and their explanation techniques, but they tend to provide only brief coverage of evaluation measures. One significant challenge in XAI research is evaluating and comparing different explanation methods, given the multidisciplinary nature of interpretability and explainability (Lipton, 2018). Evaluation measures are typically divided into two categories: human-grounded measures, which depend on human subjects, and functionally-grounded metrics, which can be computed without human input (Doshi-Velez and Kim, 2017; Nauta et al., 2023). Many researchers are exploring automated solutions for evaluating explanations. A thorough literature review focusing on these functionally grounded evaluation methods can be found in Nauta et al. (2023). Since explainability is fundamentally a human-centric property, the research community is increasingly recognizing the importance of human-centered evaluations in XAI (Doshi-Velez and Kim, 2017; Lipton, 2018).

For example, Mohseni et al. (2021) categorize evaluation metrics related to humans into four areas: understanding the model, user trust, performance in human–AI tasks, and satisfaction with explanations. On the other hand, Hoffman (2019) emphasizes psychometric evaluations, proposing a model of the XAI process with four key aspects for evaluation: quality and satisfaction of explanations, users' mental models, curiosity, and trust and performance. Apart from evaluating methods, XAI applications aim to aid decision-making and benefit end users. In a recent review, Lai et al. (2023) explore research on collaborative Human–AI decision-making, which may involve AI systems providing explanations. Success in tasks involving human–AI interaction serves as one of several ways to assess explanations' impact. Schoeffer et al. (2024) observed that explanations affect fairness perceptions, which subsequently influence humans' inclination to follow AI recommendations. Chromik and Schuessler (2020) introduce a framework for assessing XAI systems involving human participants. Their framework concentrates on various dimensions such as the intended goal of explanation, type of task, level of evaluation, and so forth.

AI's success in critical sectors depends on a deeper understanding of trust (Del Ser et al., 2024). Counterfactual explanations play a key role in fostering trust by enabling users to explore how hypothetical input changes affect outcomes, mirroring human reasoning in unfamiliar processes (Del Ser et al., 2024). However, effective counterfactual explanations require more than just their counterfactual nature; they must also balance multiple key aspects such as actionability, plausibility, and suggested changes in the input.

Insights from the studies that do examine XAI approaches from a user's perspective reveal that not always explanations are effective as compared to no explanations. For example, participants engaged in prediction and diagnosis tasks exhibiting consistent performance trends across agricultural and abstract domains, irrespective of whether they received CEs, pre-factual explanations, or simple control descriptions (Dai et al., 2022). Surprisingly, user ratings regarding helpfulness were similar across all explanation conditions, contradicting the anticipated superiority of CEs in terms of usability. In another user study, the research conducted by Kuhl et al. (2023b) demonstrates the effectiveness of CEs in comparison to control variants without explanations. Their study involved user interaction with a system incorporating the AlienZoo Framework, which aids players in establishing plant diets for the growth of Alien Shubs and identifying plants which are crucial for the system's predictions. Interestingly, the analysis of the data they gathered, indicates that CEs contribute positively to users' perception of system comprehensibility when contrasted with scenarios lacking explanations.

Considering all these factors, the studies we have discussed, which conducted human evaluations with users, did not prioritize involving humans in customizing explanations and then assessing system performance — a crucial aspect for human-centered XAI. Our study aims to assess whether user feedback-based explanations enhance users' cognitive learning compared to no explanations. Additionally, we aim to explore potential differences between user feedback-based explanations in their impact on evaluation and learning, stemming from their unique internal mechanisms of explanation generation. Subsequently, we will identify these evaluation measures and the extent of these differences.

3. Materials and methods

The growing volume of user studies in XAI is accompanied by a rising number of suggestions and principles for designing such studies (Kuhl et al., 2023b; Mohseni et al., 2021; Rong et al., 2023; van der Waa et al., 2021). We adhere closely to the guidelines proposed by Kuhl et al. (2023b) and van der Waa et al. (2021). Section 3.1 introduces the Alien Nutri-Solver framework. Section 3.2 presents the implementation details. Section 3.3 goes with the evaluation measures. Section 3.4 describes the experimental design for the user study.

3.1. The Alien Nutri-Solver framework

This section describes the components of the Alien Nutri-Solver framework employed to execute the user study such as the use case (Section 3.1.1), the training data and ML model (Section 3.1.2), and the CE methods (Section 3.1.3).

3.1.1. Use case

The efficacy of an explanation is impacted by its purpose and the specific context (audience) it is intended for (Mohseni et al., 2021). These two factors play a crucial role in selecting an appropriate use case and shaping the experimental setting. Thus, in this study, the Alien Nutri-Solver framework was adapted as an abstract and gamified context from Alien Zoo (Kuhl et al., 2023b), a virtual space wherein the users do not have prior knowledge and can only gain knowledge through the explanations provided by the underlying system.

In the usage scenario, the user's task is to nurture an Alien called 'Shub' by feeding various combinations of plants. In this study, in contrast to the original implementation (Kuhl et al., 2023b), the well-being of one individual Shub is directly influenced by the choices users make in selecting plant combinations. The selection of leaves for different plants incurs a specific time cost to search and find them. Users encounter a combinatorial problem where the number of leaves per plant varies randomly, presenting an unhealthy diet for the Shub. Users must address this challenge by forming different plant combinations to create a nutritious diet within their time constraints. The selected plants are then fed into an Alien Nutri-Solver framework, which assesses whether this combination can enhance the Shub's health.

The choices made during feeding directly impact the Shub's health, resulting in either poor or improved. An interactive system, as detailed in Section 3.2, assists users in making optimal decisions. Periodically, users receive explanations alongside their previous selections, highlighting choices that could have led to better outcomes. These explanations also offer insights into the underlying data distribution associated with the AI prediction model and how the Shub's health can be enhanced. This guidance empowers users to make informed decisions and enhances their understanding of the AI system's workings.

To illustrate how the fictional setting is established in the game, consider the following example set of instructions provided to participants.

> "You are the new guardian of a fictional alien species called Shub. Different groups of Shubs live on different planets. On all planets, Shubs eat different types of leaves for their growth. But beware: each group has adapted to their unique diet suited to their home planet. As you are new to the job, you do not know yet what works best on each planet. However, you are assisted by the Alien Nutri-Solver tool, an advanced, intergalactic dietary analysis tool. This tool uses cutting-edge algorithms to process and analyze vast arrays of data, simulating the current environment and dietary needs of alien species across planets. To work with the Alien Nutri-Solver tool, for each planet, you have to set beforehand preliminary ranges of leaves. These ranges are taken as global constraints in the search for optimal solutions by the Alien Nutri-Solver tool. The tool will conduct a thorough exploration of potential combinations within those limits, to suggest a healthier diet for the current planet. Your task is to travel to different planets, experiment with leaves and the Alien Nutri-Solver tool, and find the healthiest diet for the Shub".

In the given scenario, we assess the performance of real users in a task that is situated in an abstract context. One significant advantage of this abstract task lies in its ability to mitigate any potential interference stemming from users' prior knowledge. In the context of feeding aliens, every user is effectively a novice, which ensures the absence of misconceptions or pre-existing beliefs.

3.1.2. Training data and ML model

As per the usage scenario, the underlying data follow this pattern: The growth rate (fitness) displays a linear scaling between values 3 to 5 for plant 2, contingent upon plant 4 having a value ≥ 3, and plant 5 not being smaller than 3. The values and conditions used for the growth rate are chosen to create a controlled and clear relationship between the plants' attributes, ensuring a manageable and interpretable task for participants. The conditions (e.g., the value of plant 4 being >= and plant 5 not being smaller than 3) are designed to introduce complexity into the task, which was necessary for evaluating the effectiveness of the counterfactual explanations. In the original implementation by Kuhl et al. (2023b), the dataset was generated for a regression task aimed at predicting growth rate. However, we converted it into a classification dataset by labeling data points as 1 if the growth rate is greater than 1.1, and 0 otherwise (with the original dataset containing growth rates from 0 to 1.9). Consequently, in our context, the outcome label relates to fitness (either 1 or 0 based on the input). The synthetic dataset encompasses all feasible plant combinations, from which we select 100,000 data points with balanced classes. The dataset is transformed into a binary classification problem to align with our user study's goal of assessing users' ability to interpret and interact with counterfactual explanations. Binary classification is more intuitive, reducing cognitive load and enabling clearer counterfactual reasoning, which is crucial in controlled experiments. Unlike regression-based explanations, which can be complex and harder to interpret, binary classification provides straightforward decision boundaries that enhance user understanding. While this transformation may lead to some loss of nuance from the original regression problem, it ensures a more interpretable and focused evaluation. This tradeoff is necessary to maintain experimental clarity and provide meaningful insights into how users engage with counterfactual explanations.

To predict the outcome label and the subsequent fitness level from user input in each trial, we utilize a logistic regression (LR) model. This LR model has shown good accuracy when tested on the synthetic dataset (with accuracy: 0.97, recall: 0.91, and f1-score: 0.88), and it facilitates the efficient computation of CEs. To maintain consistent model outputs for all users throughout the experiment, we employ the same LR model constructed at the beginning.

3.1.3. Explanation methods

The field of XAI encompasses various technical approaches aimed at enhancing the transparency, interpretability, and explainability of AI systems (Ali et al., 2023). Human-centered XAI repositions human involvement as pivotal (Suffian et al., 2023). In light of the gap outlined in Section 1, our objective is to employ explanation methods enabling users to integrate their affordability constraints as feedback, thereby producing actionable explanations that are easily comprehensible and reliable. CEs dynamically adapt to individual cases, offering actionable insights into alternative outcomes. This enables users to understand not just what needs to change, but how to achieve a desired outcome. In this regard, we introduce below the two CE approaches employed to generate the explanations for the scenario under consideration.

UFCE. This method provides users with CEs to devise different strategies to solve the task at hand Suffian et al. (2024a). UFCE is not self-driven; rather, it is a human-in-the-loop explainer that needs user feedback to generate customized explanations. UFCE identifies influential factors (features) in the outcome by analyzing mutual information among them. Users can specify their desired features to modify and their allowable ranges of modifications. UFCE establishes a neighborhood based on these constraints, utilizing a kd-tree to partition the space and locate the closest feasible counterfactuals. It then employs a locally trained regressor model to predict outcomes for the nearest counterfactuals with minimal alterations in the subset of user-defined features. The internal mechanism of UFCE has three main components targeting one feature, two feature, and three feature changes in the actual input to find new valid and feasible CE. UFCE provides users with the opportunity to generate multiple CEs based on its functionality and a specific number of best CEs can be generated. We utilized all variations of UFCE and the CE that was the nearest to user input was selected from all. Since UFCE is user-feedback-driven, the integration of UFCE into this study was handled accordingly. In the configuration screen, depicted in Fig. 1, user feedback was taken from the learners by providing a drop-down menu for each plant at the bottom, prompting learners to select combinations from drop-down menus for each plant. The user selection of plants is fed to UFCE as a user-feedback for the generation of explanations. UFCE is available at Github.[2]

DiCE. This method offers a solution to the optimization problem of diversity and proximity, enabling the generation of any desired number of CE examples for a given input. DiCE (Mothilal et al., 2020) is adaptable to accommodate user-supplied inputs. The user can input a list of actionable features along with their allowable modifications. DiCE employs a loss function that minimizes the distance between the original instance and the counterfactual instances while maximizing diversity among the generated counterfactuals. This loss function penalizes deviations of counterfactual predictions from the original prediction. The integration of DiCE into this study was handled similarly like UFCE. We used the Python toolbox for DiCE openly available on GitHub.[3] The default hyperparameters were used which allows to setting of the ranges of input features.

[2] https://github.com/msnizami/UFCE

[3] https://github.com/interpretml/DiCE

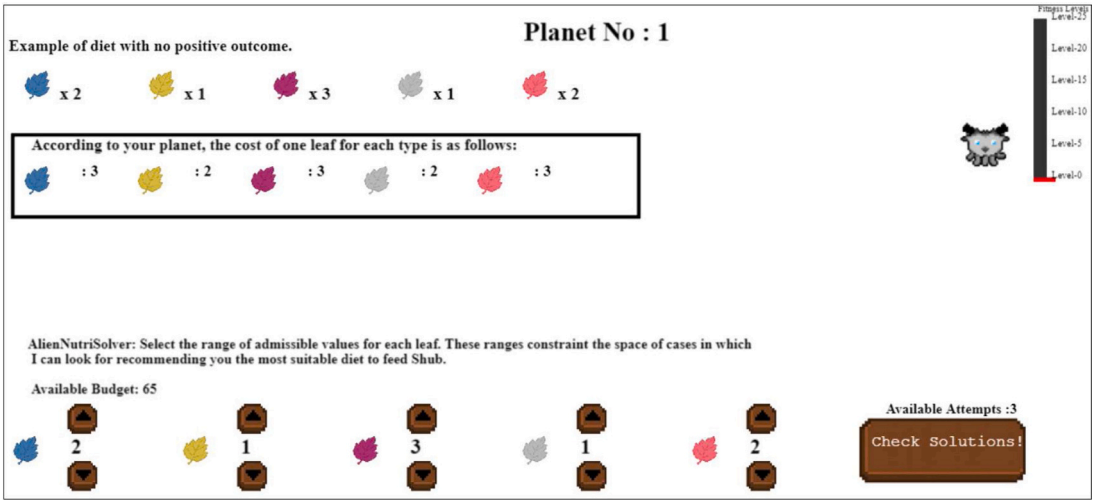

Fig. 1. The learner's task is to go for those combinations of plants by selecting from drop-down menus that can be searched in the available budget. After each search/selection, the learner feeds the plants to Shub and waits for the outcome.

3.2. System implementation

The development of the Alien Nutri-Solver framework draws inspiration from the Alien Zoo framework (Kuhl et al., 2023b), serving as an extension to the original concept of the CL-XAI framework (Suffian et al., 2024b). This adaptation involves a division between the front end, which designs the game interface for user interaction, and the back end, responsible for delivering AI predictions and their related explanations. The web interface utilizes Phaser3,[4] an HTML5 game framework driven by JavaScript. The system's backend, on the other hand, is founded on Python3, leveraging the sklearn[5] package for supporting ML algorithms. The underlying ML model is trained using the synthetic plant data (Kuhl et al., 2023b) described in Section 3.1.2, and it predicts the fitness of the Shub, thereby influencing the next steps in the game. The user input reaches this model via the front end, allowing an analysis of the potential for yielding positive outcomes and consequently the Shub's fitness. This Python-based system utilizes UFCE and DiCE to generate CEs for the underlying ML algorithm. We implemented the Alien Nutri-Solver framework as open-source software to promote Open Science. The interested reader is kindly referred to the following GitHub repository for further details (https://github.com/msnizami/HCEvaluations4UFCE).

3.3. Evaluation measures

In this user study, we evaluate the goodness of explanations generated by the user input-driven CE methods, in comparison with a control group. User evaluations are crucial for validating new explanation methods, as they ensure that explanations are understandable and useful from a human perspective. Such evaluations provide insights into how effectively the models communicate their reasoning, aligning algorithmic outputs with user expectations and enhancing trust in AI systems. There are many (human-centered) metrics for qualitative evaluation of XAI methods (Hoffman et al., 2018; Kuhl et al., 2023b; Longo et al., 2024; Vilone and Longo, 2021). It is worth noting that user understanding requires the development of a user's "mental model" regarding a system's internal operations (Hsiao et al., 2021). The concept of a "mental model" originates from psychological theories, referring to an individual's internal representation of the people, objects, and environments they interact with (Richardson et al., 1994; Staggers and Norcio, 1993). In this study, we expect the user's mental model to accurately reflect the system explained through XAI explanations. Following the framework proposed by Hoffman et al. (2018), assessing the user's practical comprehension of the AI system occurs during the explanation generation phase. This assessment involves evaluating the user's cognitive processes, which measure the quality of the explanation (referred to as "Explanation Goodness") and the level of satisfaction with it (referred to as "User Satisfaction"). Assessing a user's comprehension and satisfaction presents a significant challenge. To tackle this, we have expanded the Alien Nutri-Solver platform to assign problem-solving tasks to users. Task performance acts as a gauge of the understanding of the explanations provided. In this way, we can validate the effectiveness of XAI approaches in producing high-quality explanations. Research on user trust extends across various decision-making applications, including image classification (Buçinca et al., 2020) and loan approval (Schoeffer et al., 2022). Additionally, investigations into user trust within recommendation systems have been conducted (Kunkel et al., 2019; Ooge et al., 2022).

We evaluate explainers in the following dimensions:

1. **Understanding/Learning**. Explanations play a pivotal role in facilitating the construction of precise mental models, which can be categorized as follows: global understanding, signifying a general comprehension of a system's functioning; local understanding, signifying insight into a specific decision made by the system; and functional understanding, representing a grasp of the system's capabilities and intended usage (Chen et al., 2023; Hoffman et al., 2018; Hsiao et al., 2021; Kuhl et al., 2022). We will focus on the local understanding of the user's mental model. In addition, we will evaluate the perceived actionability of explanations (Vilone-scale (Vilone and Longo, 2021)). This could be evaluated from the objective scores (task performance) and self-reported answers (Kuhl et al., 2023a).
2. **Satisfaction/Usability**. User satisfaction can be assessed using subjective measures through the administration of a questionnaire, for example, the *Explanation Satisfaction Scale* introduced by Hoffman et al. (2018) and later refined by van der Waa et al. (2021) and adapted by Kuhl et al. (2023b). This scale is a reliable and psychometrically robust means of gauging user satisfaction with a system's explanation.
3. **Trust**. Much of the existing research on trust evaluation utilizes two main categories of measures: self-reported and observed trust (Papenmeier et al., 2019). Self-reported trust is typically assessed through user-filled questionnaires, while observed trust is determined by human agreement with decisions made by the model. In this study, we assess self-reported trust.

[4] https://phaser.io/
[5] https://scikit-learn.org/stable/

Table 1
Table of survey items related to Understanding/Learning, Satisfaction, and Trust in the user study.

RQs	Evaluation	Item No. and Item statements
RQ1	Understanding	**1** : (Explanation group) What do you think: Which plants were relevant to increase the fitness of Shub? (Control group): What do you think: Which plants were relevant to increase the fitness of Shub?
	(Quantitative Evaluation)	**2**: (Explanation group) What do you think: Which plants were irrelevant to increase the fitness of Shub? (Control group) What do you think: Which plants were irrelevant to increase the fitness of Shub? **3**: (Explanation group) I often used the "Help" button to get suggestions on what choice would have led to a better result. (Control group) I would have liked to have a "Help" button to get suggestions on what choice would have led to a better result. 4: (Explanation group) I did not understand the suggestions on what choice would have led to a better result provided by the "Help" button. (Control group) I needed no support to understand which selection choices would have led to a better result. **5**: (Explanation group) I have learned from the suggestions on what choice would have led to a better result provided by the "Help" button how to select a good diet to increase the fitness of the Shub.
RQ2	Satisfaction	**6**: (Explanation group) I found that suggestions on what choice would have led to a better result provided by the "Help" button are useful to increase the fitness of the Shub.
	(Qualitative Evaluation)	**7**: (Explanation group) I did not use the suggestions on what choice would have led to a better result provided by the "Help" button to increase the fitness of Shub. **8 (catch item)**: (Both Explanation and Control groups) To show you are paying attention to this question, please select "I prefer not to answer". **9**: (Explanation group) I found inconsistencies in the suggestions on what choice would have led to a better result provided by the "Help" button. (Control group) I found inconsistencies in the behaviour of the AlienNutriSolver. **10**: (Explanation group) From the suggestions on what choice would have led to a better result provided by the "Help" button, I do not understand how the AlienNutriSolver works. (Control group) From interacting with the system, I do not understand how the AlienNutriSolver works. **11**: (Explanation group) The suggestions on what choice would have led to a better result provided by the "Help" button are satisfying. **12**: (Explanation group) The suggestions on what choice would have led to a better result provided by the "Help" button have sufficient detail. **13**: (Explanation group) The suggestions on what choice would have led to a better result provided by the "Help" button seem incomplete. **14**: (Explanation group) The suggestions on what choice would have led to a better result provided by the "Help" button tell me how to use the AlienNutriSolver.
RQ3	Trust	**15**: (Both Explanation and Control groups) I trust the predictions of the AlienNutriSolver. **16**: (Explanation group) I do not trust the suggestions on what choice would have led to a better result provided by the "Help" button.
	(Qualitative Evaluation)	**17**: (Explanation group) I am confident in the suggestions on what choice would have led to a better result provided by the "Help" button. I feel that they work well. **18**: (Explanation group) I do not feel safe when I rely on the suggestions on what choice would have led to a better result provided by the "Help" button, I will get the right answers. **19**: (Explanation group) I am wary of the suggestions on what choice would have led to a better result provided by the "Help" button. **20**: (Explanation group) I like using the suggestions on what choice would have led to a better result provided by the "Help" button for decision-making.

3.4. Study design

In this section, we outline the design of the user study aimed at conducting an empirical investigation to gain a thorough understanding of the employed explainers within the Alien Nutri-Solver framework. The study seeks to determine whether and to what extent these explainers influence participants' desirable behavior when they are tasked with and without explanations.

We look for answers to the following research questions:

1. **RQ1: How UFCE and DiCE methods affect individuals' understanding for better task performance?** We evaluate whether users can enhance their understanding of complex data relationships through explanations. By providing CEs, we aim to help users better grasp the system's intricacies. Users will also gain the ability to explicitly recognize relevant and irrelevant input features, indicating proficient capability to discern critical data factors. We evaluate whether users can perform better to maximize the objective scores. We also evaluate how well they performed to obtain objective scores during the game as well as self-reported responses in the post-game survey.
2. **RQ2: How UFCE and DiCE methods affect individual's satisfaction?** We evaluate whether user satisfaction is enhanced thanks to the given explanations. We anticipate that better task performance can lead to enhanced satisfaction in the explanations. A post-game survey is employed to gather feedback from users regarding their satisfaction with the explanations.
3. **RQ3: How UFCE and DiCE methods influence user's trust?** We evaluate whether users' trust is enhanced thanks to the given explanations. We anticipate that better task performance and satisfaction with the explanations can lead to enhanced trust in the system. A post-game survey collects feedback from users regarding their trust in the explanations.

Further, Table 1 presents a summarized mapping about the RQs, evaluation metrics, and corresponding survey items. A comprehensive detail about this mapping table is provided in Supplementary Materials. By examining these research questions, we aim to uncover how explanation customization and their quality influence the user's learning process (objectively and subjectively).

3.4.1. Experimental treatment

We have adopted a between-subject design (Rong et al., 2023).

For the user-input task concerning fitness prediction, we design three treatments: one control and two experimental (DiCE and UFCE), as follows:

- **Control (no explanation)**: Participants are not provided with any explanation regarding the model's prediction for each task.

We have avoided providing any help button even with generic tips for the control group to maintain the integrity of evaluating the specific intervention being tested. The purpose of the control group is to serve as a baseline without external assistance, which allows for a clearer assessment of how the feedback-driven explanations impact user performance and understanding.

- **UFCE (counterfactual explanation)**: We elucidate the model's prediction for each task by examining alterations in feature values that could lead to an opposing model prediction. Leveraging UFCE's capability for user input to define feature value ranges, it pinpoints the smallest change required in a feature to reverse the model's prediction (while keeping other feature values constant). Additionally, if users encounter difficulties in finding a solution, they are guided with supportive assistance (a Help button), offering hints or guidance to facilitate the discovery of optimal solutions.
- **DiCE (counterfactual explanation)**: This treatment is similar to UFCE, however, we utilized DiCE to provide explanations suggesting the smallest change required in a feature(s) to reverse the model's prediction. For the sake of a fair comparison, similar supportive assistance and guidance was offered in this treatment as it was in UFCE.

3.4.2. Participants

At the beginning of March 2024, the study was carried out by distributing a public link to the web server hosting the game study among the students of Multiomics,[6] an online educational startup offering multiple learning and training courses to students with different educational backgrounds. In Multiomics, they believe in the facilitation of scientific research activities, hence, we targeted 120 participants with different educational backgrounds in science disciplines. We carefully selected the individuals who lacked fundamental knowledge of AI, a prerequisite for our experiment. These participants were divided into distinct groups within a between-subjects design. For each treatment, individuals were randomly assigned to one of three groups: Control, UFCE, or DiCE. Before participation, all participants had to accept the informed consent electronically, indicating their agreement by clicking a designated box. We excluded contributions from participants whose data did not meet our quality standards (see Section 3.4.4). Participants did not receive any monetary benefit as it was a voluntary participation. However, after completing the study they were debriefed through a link to the online repository containing the information regarding the study. Additionally, they were provided with a brief informative session on AI and XAI and their applications. Participants were presented with an information sheet and provided informed consent by ticking a box, affirming their understanding that participation was anonymous and that they could withdraw at any stage of the game and post-game survey. This study took place in an online laboratory setting, with a facilitator ready to assist participants and handle any technical queries during the game. The facilitator ensured participants understood the basics of the experiment and provided guidance throughout. The presence of the facilitator offers multiple benefits, including enhanced participant understanding, resolution of technical issues, real-time monitoring, and ensuring the integrity of the user study. However, the facilitator's role was strictly limited to assisting in the administration of the tasks and ensuring consistency across sessions. The facilitator did not interfere with or assist in the decision-making on the provided tasks. This approach was implemented to ensure uniformity in task execution across all participants.

[6] https://wemultiomics.com/courses/

3.4.3. Experimental procedure

Upon arrival, participants were randomly allocated to one of three groups: Control, UFCE, or DiCE. Participants initially familiarized themselves with general information and specifics regarding their voluntary participation and consent. The experiment was composed of two phases: a game and a survey.

At the start, users were provided with comprehensive information regarding the study's objectives, procedures, expected duration, their right to withdraw, confidentiality measures, and contact details for the primary investigator. Should users opt not to participate, they have the option to close the window. Alternatively, users indicate their agreement by pressing a designated button. Subsequently, a detailed page presents information about the game, featuring images of the Shub creature to be fed and the assortment of available plants.

Written instructions clarified that the fitness of the Shub can be influenced by the selection of healthy or unhealthy combinations of leaves per plant, with a maximum limit of six leaves per plant. Users are granted the flexibility to choose any preferred combination of plants. Additional guidance is provided to assist users in maximizing the Shub's fitness level to facilitate learning for the task. Moreover, users receive feedback on which choices could have yielded better results. Clicking the "Start" button indicates that the user is ready to commence the game phase.

Game Phase. The user interface is designed to provide learners with an intuitive and captivating experience as they embark on their journey of nurturing a Shub and exploring the intricate connections between plant combinations and growth outcomes. Comprising various screens, each screen is crafted to deliver essential information and interactivity, ensuring an engaging learning environment. In the configuration screen, depicted in Fig. 1, an avatar representing a Shub dynamically reflects its fitness level on a vertical bar, offering real-time feedback as the chosen plant combination affects the Shub's health. The bar's limits indicate the optimal and unsatisfactory fitness levels. In addition, the five different plants are displayed at the top, forming the Shub's diet, with a customizable test input presented at the bottom, prompting learners to select combinations from drop-down menus for each plant.

Additionally, users are provided with information on available attempts and budgets for improving health, along with the time cost associated with investing in each leaf of every plant. The costs of leaves for each plant, which vary across different planets, are also indicated. As users increase the number of leaves, the cost is deducted from the budget, with the updated budget displayed accordingly. Before starting the game, users are advised to exercise caution regarding costs and budget when selecting leaves. Furthermore, a Help button becomes available when the user's provided inputs do not result in an improved diet. In Fig. 2, it can be observed, a Help button appears when there was no improvement in the diet. Upon pressing the Help button, the system offers suggestions to guide the user towards better diet plans. In general cases, Help buttons are included to help-out from the start. However, in this user study, the help button was designed to appear only after users encountered difficulty, such as failed attempts, to encourage initial exploration and independent learning relying on the solutions provided by the system. Providing help from the start could lead to some users relying solely on the help button, which would not reflect genuine learning or problem-solving. Participants are initially assigned a Shub with a fitness level of 0. To submit their choices, participants click a "Check Solution!" button in the bottom right corner of the screen, which, upon pressing, changes to "Loading". The illustrations of different game screens are provided as Supplementary Material in Appendix C. The underlying machine learning model predicts the new fitness based on the user's input. Subsequently, the explainer method computes a CE, presents it to the user for review, and feeds it to Shub. Upon pressing the "Feeding Time!" button, a brief progress scene is displayed, during which the underlying ML model updates the Shub's fitness score. The user is then directed to the next planet (configuration screen), visualizing the impact of the current choice through written

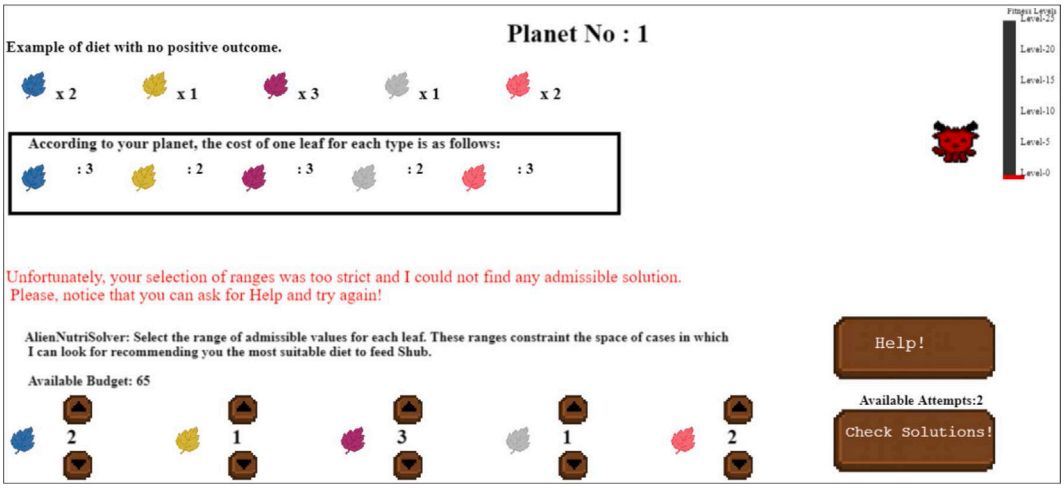

Fig. 2. Help button: When the user selected ranges to constitute the diet resulted in no valid solution for a better diet. A help button appears and by clicking it the user receives feedback to look for suggestions and to try again.

information and an animated Shub. Participants play the game for three planets.

Survey Phase. After playing the game phase, participants engage with a series of questions aimed at assessing various aspects (question items are presented in Table 1). As per different treatments, control and experimental groups respond to questions designed according to their game environment. The control participants encounter a total of 8 questions regarding understanding, satisfaction, and trust. The experimental participants (UFCE and DiCE) engage in a series of 20 questions regarding understanding, satisfaction, and trust. Initially, the survey items seek to evaluate users' explicit knowledge regarding the significance of plants for task success (for all treatments, covered in items 1 and 2 of the postgame survey, Table 1). Participants conveyed the subjective nature of their explanation requests using the Help button and their understanding of explanations in the experimental group, while the control group expressed their need for explanations and support (covered in items 3 and 4 of the postgame survey, Table 1). Subsequently, participants provide subjective assessments of the usefulness and satisfaction derived from the experience (explanation group, covered in items 5 and 6 of the post-game survey, Table 1). Similarly, participants provide subjective assessments on satisfaction and trust items, presented in detail in Table 1 (Supplementary Material contain comprehensive details for each item of survey items). The last 3 items of the survey phase concentrate on collecting demographic information, specifically regarding the participant's gender, age, and English language proficiency.

Upon concluding the study, participants receive gratitude for their participation. Additionally, participants are given the option to access a link that provides comprehensive debriefing information.

3.4.4. Data quality

Due to the inherent nature of web-based studies, there is a possibility of incomplete participation by some users. To ensure data integrity, predefined criteria are established beforehand. Only participants who fully engage with the game and complete the post-game questionnaire are included in further analysis. Participants are identified as inattentive users if they fail to answer the correct response to catch an item during the specific phase of the study. Similarly, users exhibiting consistent responses of solely positive or negative valence during the survey phase are categorized as straight-liners. Individuals meeting any of these criteria are excluded from subsequent analysis to uphold data quality standards.

3.4.5. Statistical analysis and measures

All the statistical tests carried out in this user study, were conducted with R-4.3.1 (R Core Team et al., 2013). We developed the code for performing the statistical analyses in RStudio, adapting the script from Kuhl et al. (2023b) and expanding it to suit our specific requirements. We focused on comparing experimental conditions and controls, as well as comparing experimental groups.

We conducted a power analysis for a one-way ANOVA under different assumptions of large and medium effect size to determine the statistical power of the conducted user study. We conducted significance tests to compare groups of experimental and control conditions. We applied the Kruskal–Wallis test (McKight and Najab, 2010) to run a non-parametric analysis of differences between all three groups. A significant value of $p < 0.05$ was taken to suggest that at least one group differs from the others. If the Kruskal–Wallis test yielded significance, we proceeded to compare multiple pairs of groups with post-hoc Dunn's test (Dinno and Dinno, 2017), Bonferroni corrected to control for multiple comparisons. Conversely, if the Kruskal–Wallis test did not yield significance, no further post-hoc analyses would be conducted.

This procedure was used to check for potentially confounding differences between all three groups in terms of covariates age and gender, judged relevance of plants (i.e., understanding, RQ1), and mean fitness scores (i.e., objective task performance, RQ1). Assessments of potential group differences between the two experimental groups UFCE and DiCE for Satisfaction (RQ2) and Trust (RQ3) were based on the data from the post-game survey via the Wilcoxon–Mann–Whitney U-test (Nachar et al., 2008) for ordinal data. We report the corresponding U statistic based on the sum of ranks, the effect size r, and p.

4. Results

This section describes the results to answer the designed research questions to assess the objective and subjective measures of different groups of users when tasked in an unknown domain. We collected data from 101 participants, among whom 86 played the game and responded to the post-game survey, completely. We only analyzed data from participants who fully engaged in the game and completed the post-game survey (see Table 2). Among the participants, there were males (n = 39), females (n = 44), and those who chose not to disclose their gender (n = 3). Ages ranged from 18 to 34 years.

From those 86 participants, 27 users were in the control group (15 female, 12 male, median age group is 18–24y), and 23 users were in the UFCE group (15 female, 7 male, 1 other, median age group is 18–24y), and 36 users were in the DiCE explanation group (14 female, 20 male,

Table 2
Demographic information of participants (m denotes male, f denotes female, and o denotes other).

Before quality assurance measures (N = 86)					
–	Control	UFCE	DiCE	H(2)	**p-value**
N	27	23	36	–	–
Gender	15 f /12 m	15 f/7 m/1 o	14 f/20 m/2 o	4.014	0.134
Age	18–24 y	18–24 y	18–24 y	0.176	0.915
After quality assurance measures (N = 70)					
–	Control	UFCE	DiCE	H(2)	**p-value**
N	24	20	26	–	–
Gender	13 f/11 m	13 f/7 m	10 f/16 m	5.398	0.067
Age	18–24 y	18–24 y	18–24 y	0.279	0.869

2 other, median age group is 18–24y). Before applying data quality criteria, there were no significant differences between groups in terms of age (H(2)=0.176,p = 0.916) or gender (H(2)=4.015,p = 0.134).

A power analysis for a one-way ANOVA with three groups (n1 = 27, n2 = 23, n3 = 36) was conducted under different assumptions about the effect size. For a large effect size (f = 0.40), the estimated power was 0.91 at $\alpha = 0.05$, indicating that the study is adequately powered to detect substantial differences between groups. In contrast, for a medium effect size (f = 0.25), the estimated power was 0.65, which may not be sufficient to reliably detect medium-sized effects. Therefore, while the study is well-powered for detecting large effects, caution is exercised while interpreting the results for medium effects, as they are not detected with high confidence for few items.

4.1. Quality criteria

Before going deeper into the results, we should apply some quality criteria to our data. Namely, sub-quality data should be removed as follows.

Identify "straight liners" in survey part. We identified users consistently providing uniform responses during the survey phase. The objective was to detect user IDs engaged in "straight-lining", wherein responses exclusively lean towards either positive or negative valence.

In-attentive participants. We identified participants who demonstrated inattentiveness and struggled to provide accurate responses to catch items.

Removing users with irregular data. We removed irregular data as follows:

- we removed 0 users that straight-lined in the survey;
- we removed 3 users from the control group and 10 users from the experimental groups (with 2 participants from the UFCE group and 8 participants from the DiCE group) who were not attentive to the catch item;
- we removed 3 users due to technical issues of not correctly storing their data (with 1 participant from the UFCE group and 2 participants from the DiCE group).

Final clean data. To sum up, in our final data we have 70 users, with 24 users in the control group (13 female, 11 male, median age group is 18–24y), and 20 users in the UFCE group (13 female, 7 male, median age group is 18–24y), and 26 users in the DiCE group (10 female, 16 male, median age group is 18–24y).

In this dataset, there were again no significant differences between groups in terms of age (H(2)=0.279,p = 0.870) or gender (H(2)=5.399,p = 0.067).

4.2. Response to research questions

The central inquiry in this large-scale user study is whether users experience advantages when provided with CEs (UFCE and DiCE) as compared to users who do not receive CEs during the process of learning and identifying relationships within an unfamiliar dataset while engaging with the Alien Nutri-Solver framework. Based on this inquiry, the aim is to evaluate UFCE with human participants and check if UFCE performs better than control and DiCE groups in terms of user understanding and task performance, satisfaction with CEs, and trust in CEs. This aim is to be achieved by answering the designed research questions, and each research question is answered with one or more analyses of data collected in the user study. In the following subsequent sections, the research questions are answered with data analyses.

4.3. RQ1: Effects on individuals' understanding and task performance

The research question RQ1 inquires about the effects on the understanding and task performance of individual users when they are provided with explanations versus the case without explanations. This evaluation involves metrics derived from participants' judgments to identify relevant plants, task performance, objective scores obtained during the game, as well as self-reported responses in the post-game survey. We analyzed the objective scores from the user in the game phase and self-reported responses in the post-game survey to see whether users were able to explicitly recognize relevant and irrelevant input features. We performed three analyses on the following constructs: *Judged Relevance of Plants*, *Task Performance*, and *Subjective Understanding in Survey Phase*.

4.3.1. Judged relevance of plants

We expect that users receiving explanations can more clearly state which plants were relevant and not for the fitness of Shub.

The plots in Fig. 3 illustrate users' perceived relevance of various plants (features). The blue rectangular boxes delineate the ground truth of plants 2, 4, and 5 deemed relevant for the Shub's fitness. Conversely, plants 1 and 3 were deemed irrelevant. Analysis reveals that over 65% of UFCE users correctly identified plant 2, and similar behavior of DiCE users has been shown for plant 4.

The mean number of matches of user judgments and ground truth is shown in Fig. 4 with statistical significance indicated by asterisks. There are significant differences between groups in terms of matches between judged plants and ground truth plants for both relevant and irrelevant plants (H(2)=9.920,p = 0.007). Post-hoc test revealed a significant difference between Control and DiCE (Z = −2.618, adjusted p = 0.026), between Control and UFCE (Z = −2.823, adjusted p = 0.014), but not between UFCE and DiCE (Z = −0.455, adjusted p = 1). It can be stated that those who got explanations were better at picking out important features than those who did not.

4.3.2. Task performance

We observe how users who received CEs and utilized the Help button can perform better than the control users in the formulation of a nutrient diet for the better fitness of Shub, leading to higher fitness scores.

In the game, there were three tasks each to be performed on a specific *planet* (e.g., task-1 on planet-1). The users were assigned a consolidated budget for all tasks to select leaves to constitute the diet for a Shub. For a successful finish of the task, the system credits the user a score of 5, thus, after the successful completion of all three tasks it will assign the user a score of 15.

The plots in Fig. 5 show the mean fitness scores per group and indicate the statistical differences and significance among the groups. There are significant differences between groups in terms of the final fitness score obtained per group (H(2) = 24.729,p = 0.000004). Post-hoc test revealed a significant difference between Control and DiCE (Z = −2.501, adjusted p = 0.037), and between Control and UFCE (Z = −4.971, adjusted p = 0.000002), and also between DiCE and UFCE (Z = −2.680, adjusted p = 0.022). It can be stated that users receiving an explanation did achieve higher fitness scores as compared to those

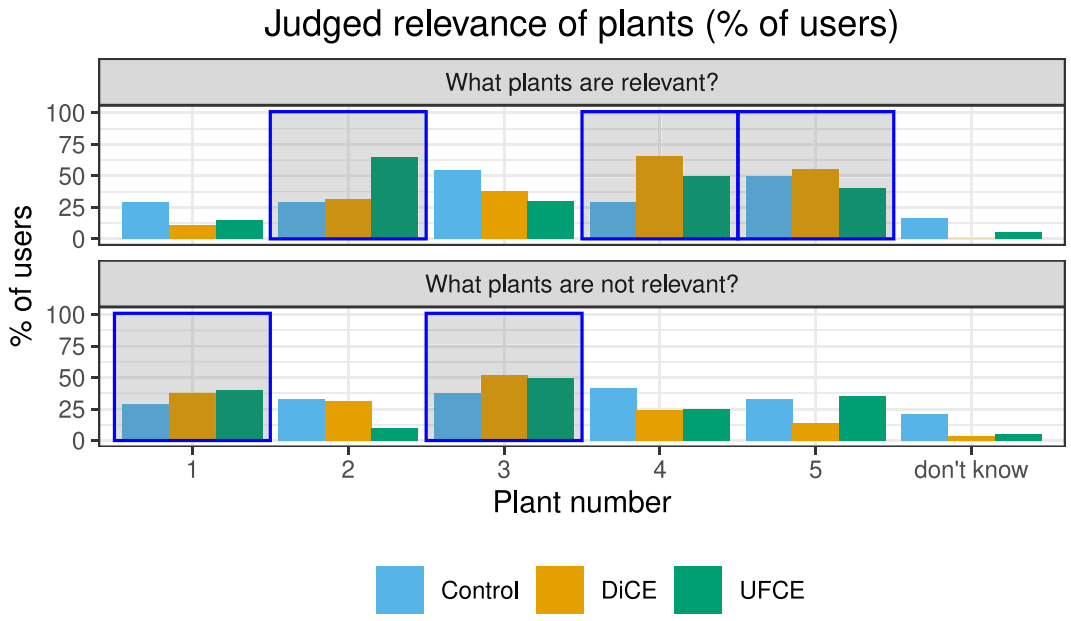

Fig. 3. RQ1: Judged relevance of plants (understanding). The users assessing relevant plants and irrelevant plants, blue rectangular boxes are the ground truths.

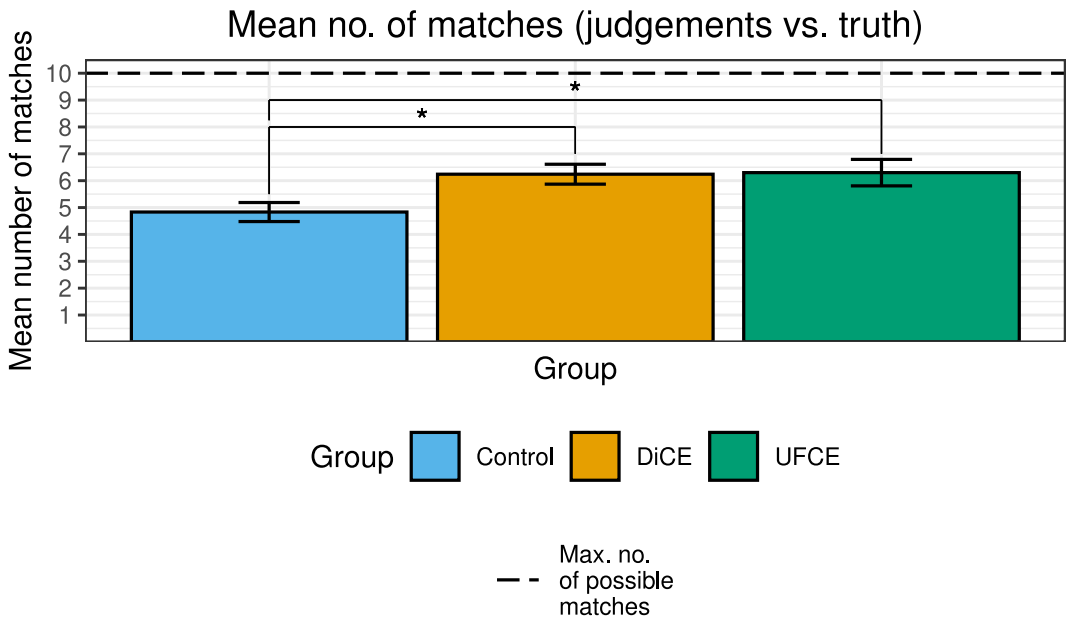

Fig. 4. RQ1: Judged relevance of plants (understanding) and statistics. Mean number of matches between user judgments and ground truths, asterisk * denote statistical significance ($p < 0.05$).

who did not. Further, users in the UFCE group outperformed users in the DiCE group.

For users who were not able to achieve the maximum scores, it could be due to their style of playing, in which they used higher costs for less important plants. To validate the aspect of not performing better, we validate objectively the mean changes in plants for the users in different groups to analyze whether their changes were for the right plants.

Fig. 6 illustrates the mean changes in plants. It can be observed that participants of DiCE have changed plant-1 and plant-3 more often which is a wrong strategy according to the underlying data distribution to achieve better fitness scores.

The trend of task performance endorses to the hypothesis that the provided explanations resulted in enhanced learning and task performance with systematic differences in experimental groups driven by disparities in individual user behavior.

4.3.3. *Subjective understanding in survey phase*

In the survey phase, items 3 and 4 are differently presented in the explanation groups and control group. In the explanation groups, item 3 asks about the participant's behavior for "usage of help button to understand explanations" and item 4 asks about "whether the user understands the choices for better fitness or not". In the control group, as there was no help button and no explanation suggestions, item 3 asks "whether the user needed the help button to understand choices" and

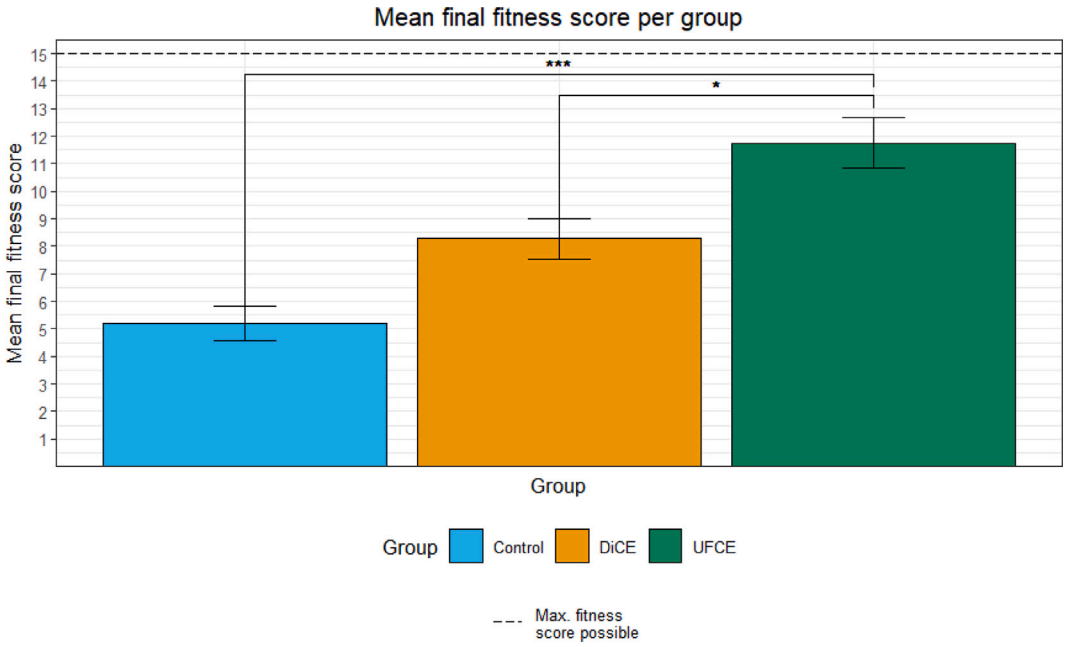

Fig. 5. RQ1 (Task performance): Mean fitness scores, where * indicates $p < 0.05$ and *** indicates $p < 0.001$.

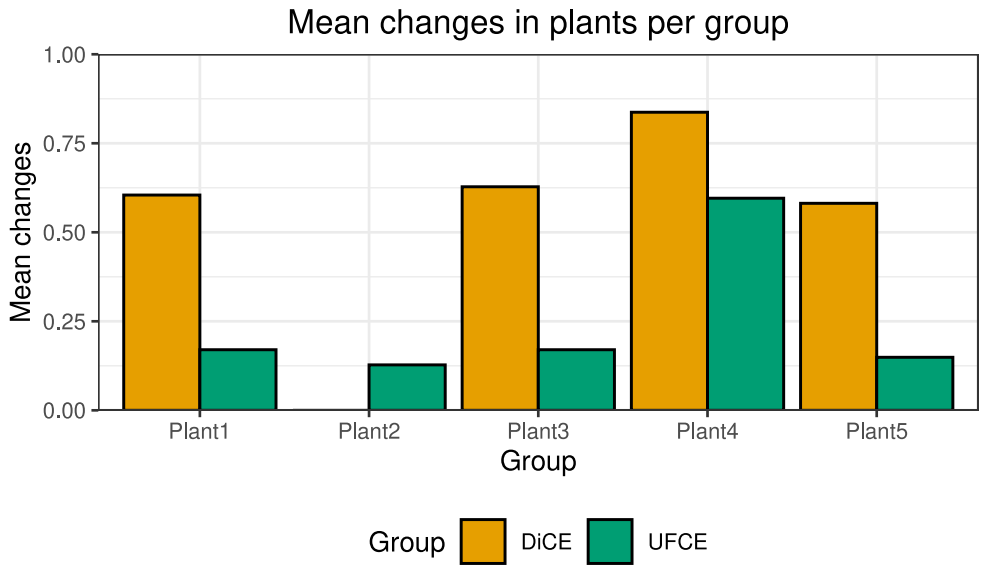

Fig. 6. RQ1 (Task performance): Mean changes in plants per group.

item 4 asks "whether the user did not need to get the help button". We first present the explanation group results with a statistical test to see if there is any difference between the UFCE and DiCE, and then present the control group results.

In Fig. 7, the stacked bars denoting user responses for UFCE and DiCE are plotted and annotated with an indication of no statistical significance (n.s.). On to the comparison for the Likert scale, the Wilcoxon–Mann–Whitney U test revealed that there is no significant difference between the users in the UFCE group (M = 4, SEM = 0.229) and users in the DiCE group (M = 4.034, SEM = 0.116) in terms of using the help button (U = 310,p = 0.637, r = 0.067).

Similarly, there is no significant difference between the users in the UFCE group (M = 2.263, SEM = 0.168) and users in the DiCE group (M = 2.703, SEM = 0.198) in terms of understanding the suggestions (U = 206,p = 0.167, r = −0.203).

Fig. 8 shows the stacked bar denoting user responses on items 3 and 4 for the control group. It can be observed that almost all users of the control group needed support in the form of a help button. In addition, they strongly disagreed with the following statement "They do not need support to understand choices".

Survey item 5 is related to the actionability of explanations and therefore it corresponds to explanation groups. Fig. 9 shows the mean user responses for both DiCE and UFCE with statistical significance.

The Wilcoxon–Mann–Whitney U test revealed that there is a significant difference in user responses of UFCE (M = 3.85, SEM = 0.208) versus explanations generated by DiCE (M = 3.392, SEM = 0.165) in terms of actionability of explanations (U = 367.5,p = 0.038, r = 0.298). Accordingly, UFCE explanations were found more actionable than DiCE explanations.

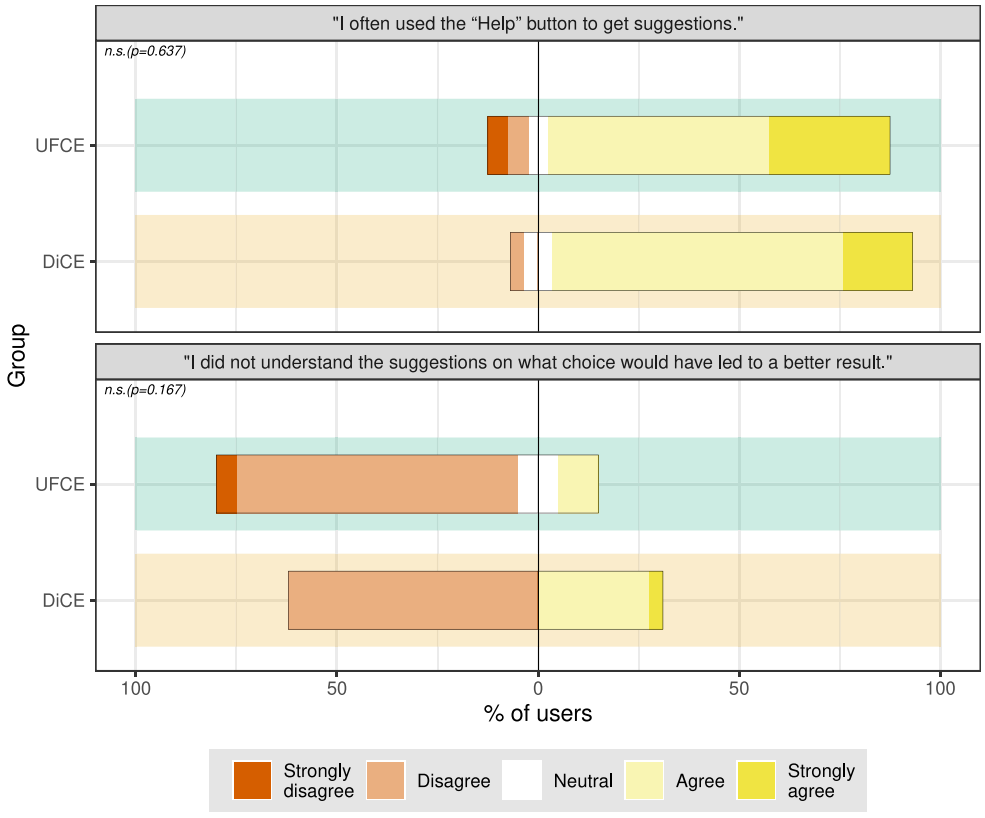

Fig. 7. RQ1 (Subjective understanding): Responses to survey items 3 and 4 in Explanation groups.

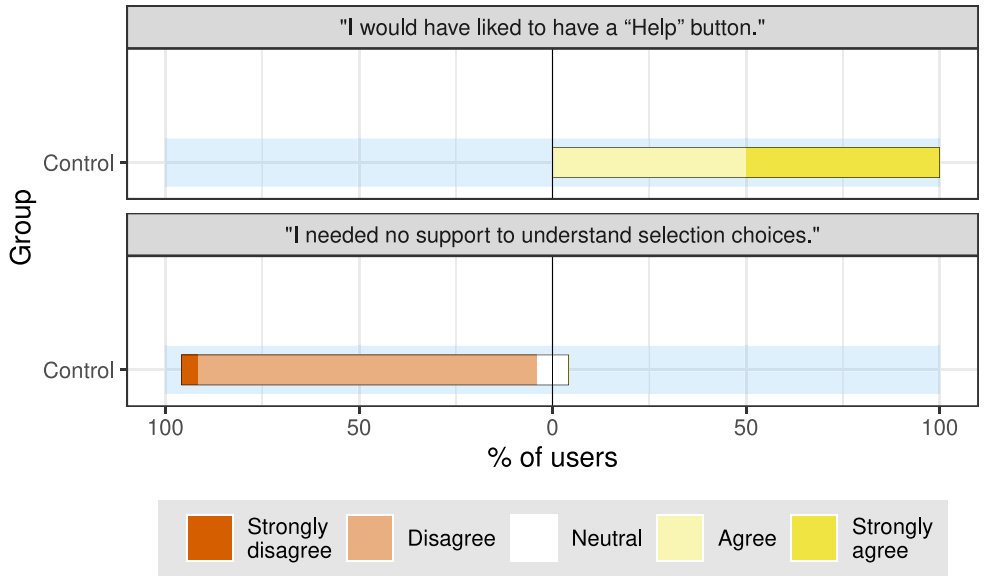

Fig. 8. RQ1 (Subjective understanding): Responses to survey items 3 and 4 in the Control group.

4.4. RQ2: How do explanations affect individual's satisfaction?

The research question RQ2 inquires about the individual's level of satisfaction with the Alien Nutri-Solver framework depending on the fact of receiving (or not receiving) explanations during the game. The survey items from 6 to 14 are designed to assess the user's satisfaction level. Notice that, item 8 is a catch item to check the attentiveness of users, which was used for filtering participants and ensuring data quality, but it is excluded from the analyses in this section. These survey items are customized for explanation and control groups. Items 6–7 and 9–14 assess the satisfaction for explanation groups, while items 9–10 are customized for the control group (the sequence of items in the control group was different in actual implementation). As the items assess different aspects for explanation and control groups, we analyze their results separately. For the explanation groups, survey items assess the helpfulness of explanations, explanation usage, robustness of explanations, perceived interpretability of explanations, explanation satisfaction, detail of explanations, completeness of explanations, and instructional level of explanations. On the other hand, for the control

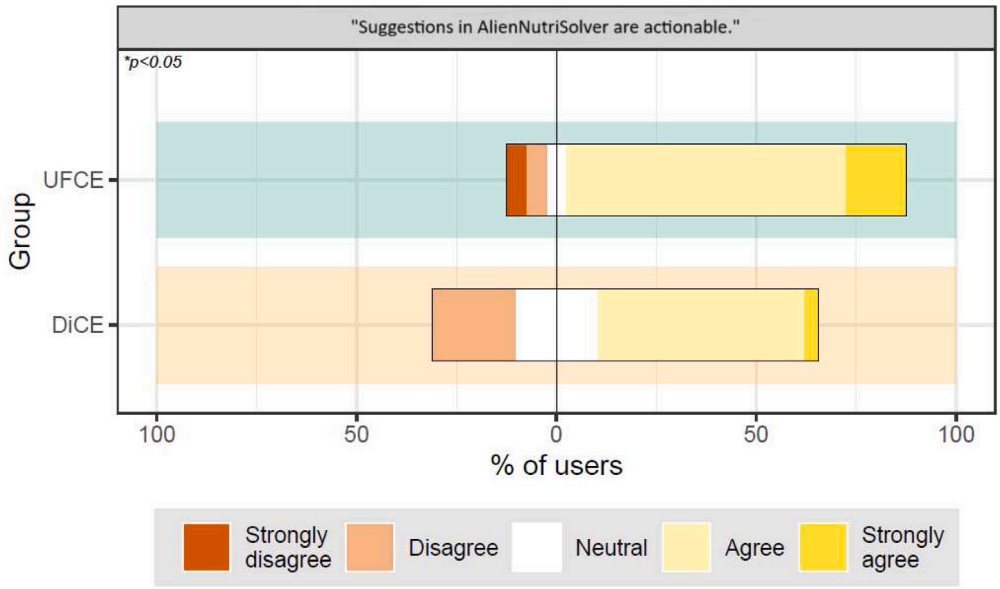

Fig. 9. RQ1 (Subjective understanding): Responses to survey item 5, actionability of explanations in explanation groups with significant difference of $p < 0.05$.

group, items 9 and 10 assess the perceived robustness and perceived interpretability of the Alien Nutri-Solver framework.

We present the analyses of results for all survey items with their statistical significance. We pay attention to: *Explanation helpfulness and usage*, *Explanation perceived robustness and interpretability*, and *Explanation satisfaction*.

4.4.1. Explanation helpfulness and usage

The plots in Fig. 10 illustrate mean responses about users' self-reported satisfaction on survey items 6 and 7 in terms of explanation usefulness and usage, respectively. For item 6 "Suggestions are useful to increase the fitness of the Shub", we compared the responses of users in the UFCE group (M = 4.055, SEM = 0.098) and users in the DiCE group (M = 3.655, SEM = 0.142). The Wilcoxon–Mann–Whitney U-test revealed that there is not a significant difference in responses regarding the usefulness of suggestions (U = 325,p = 0.060, r = 0.273). Similarly, we compared responses of item 7 "I did not use the suggestions" for users in the UFCE group (M = 2.25, SEM = 0.175) and users in the DiCE group (M = 2.827, SEM = 0.186). The analysis revealed that there is a significant difference in the usage of suggestions (U = 194.5,p = 0.026, r = −0.316).

4.4.2. Explanation perceived robustness and interpretability

The plots in Fig. 11 illustrate mean responses about users' self-reported satisfaction on survey items 9 and 10 in terms of explanation robustness and perceived interpretability, respectively. For item 9, the Wilcoxon–Mann–Whitney U-test revealed that there is not a significant difference in responses in the UFCE group (M = 2.533, SEM = 0.273) and users in the DiCE group (M = 2.88, SEM = 0.185) regarding robustness in terms of inconsistencies in the explanations (U = 144,p = 0.200, r = −0.202). Similarly, for item 10 about the working of the Alien Nutri-Solver framework, the analysis revealed that there is not a significant difference in users in the UFCE group (M = 2.588, SEM = 0.192) and users in the DiCE group (M = 3.137, SEM = 0.190) regarding the perceived interpretability (U = 173.5,p = 0.076, r = −0.260).

For the control group, items 9 and 10 are analyzed as well. The plots in Fig. 12 illustrate mean responses about users' self-reported satisfaction in terms of robustness and perceived interpretability. It can be observed that users in the control group responded neutrally to both aspects.

4.4.3. Explanation satisfaction

The plots in Fig. 13 illustrate mean responses about users' self-reported satisfaction on items 11 and 12 in terms of explanation satisfaction and sufficient details, respectively.

For item 11 about satisfaction, we compared the responses of users in the UFCE group (M = 3.777, SEM = 0.172) versus the responses of users in the DiCE group (M = 3.413, $SEM = 0.168$).

The Wilcoxon–Mann–Whitney U test revealed no significant difference (U = 317,p = 0.132, r = 0.219). Similarly, for item 12, the Wilcoxon–Mann–Whitney U test revealed that there is not a significant difference between the responses of users in the UFCE group (M = 3.444, SEM = 0.231) versus responses of users in the DiCE group (M = 3.642, SEM = 0.156) regarding the sufficient details in the explanations (U = 216.5,p = 0.320, r = −0.146).

The plots in Fig. 14 illustrate mean responses about users' self-reported satisfaction on survey items 13 and 14 concerning helpfulness and completeness of suggestions made by the Alien Nutri-Solver framework.

For item 13 about whether suggestions are helpful, we compared the responses of users in the UFCE group (M = 3.416, SEM = 0.229) versus responses of users in the DiCE group (M=2.964, $SEM = 0.158$) with a Wilcoxon–Mann–Whitney U test. It showed that there is not a significant difference (U = 218,p = 0.118, r = 0.246). Similarly, for item 14 about the completeness of suggestions, we compared the responses of users in the UFCE group (M = 3.789, SEM = 0.210) versus responses in the DiCE group (M = 3.551, SEM = 0.145). The analysis showed that there is not a significant difference (U = 326,p = 0.177, r = 0.194).

The above analysis indicates that no significant difference was found between the users in both groups. Since, no significant difference was observed, it does not necessarily imply equal outcome and satisfaction.

4.5. RQ3: How do explanations influence individual's trust?

The research question RQ3 inquires about the individual's trust in the Alien Nutri-Solver framework depending on the use (or not use) of automated explanations. The survey items from 15 to 20 are designed to assess this issue. Item 15 is the same for both groups: explanation and control. Thus, we analyze the results of item 15 by running the Kruskal–Wallis test. The rest of the items from 16–20 are only asked to participants in the explanation groups. These items assess trust in the explanation, confidence in the explanation, perceived safety when

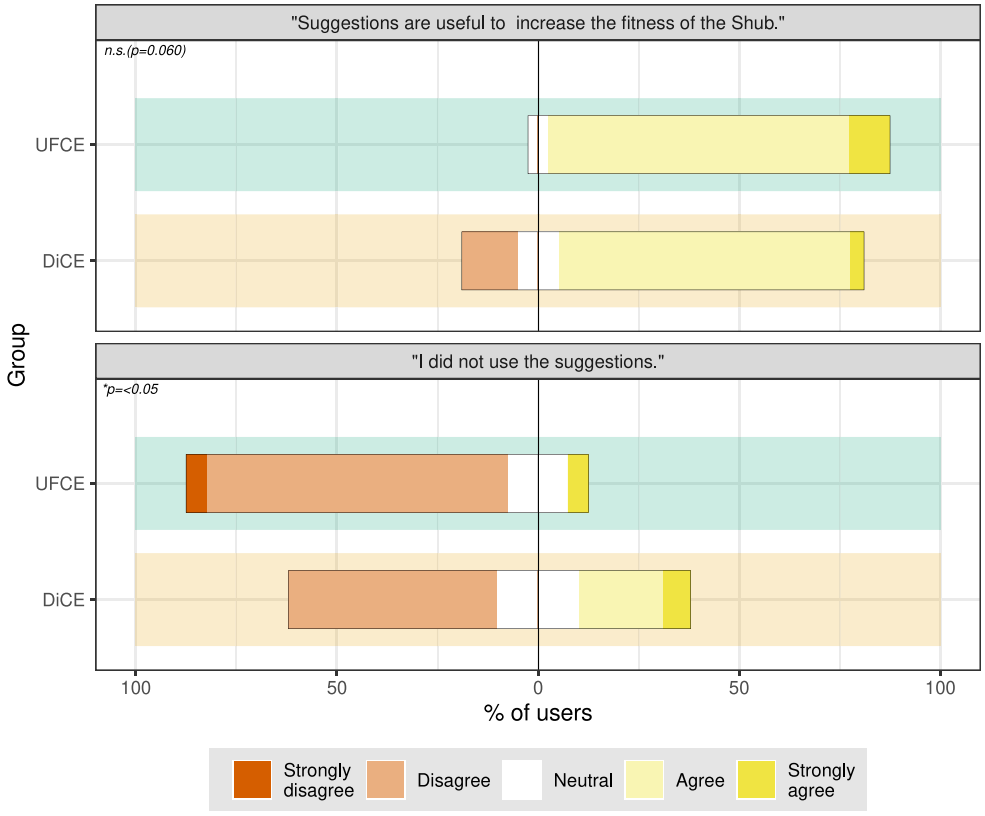

Fig. 10. RQ2 (Satisfaction): Mean responses to survey items 6 and 7 for explanation usefulness and usage with significant differences.

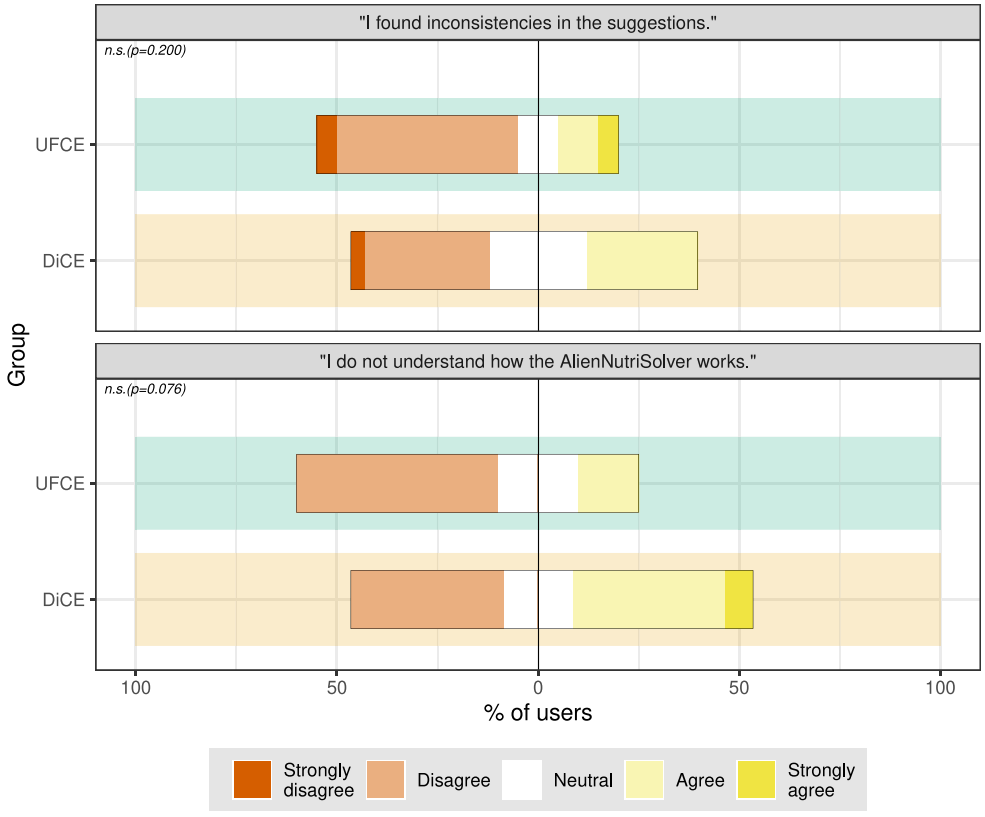

Fig. 11. RQ2 (Satisfaction): Mean responses to survey items 9 and 10 for explanation robustness and perceived interpretability for explanation groups.

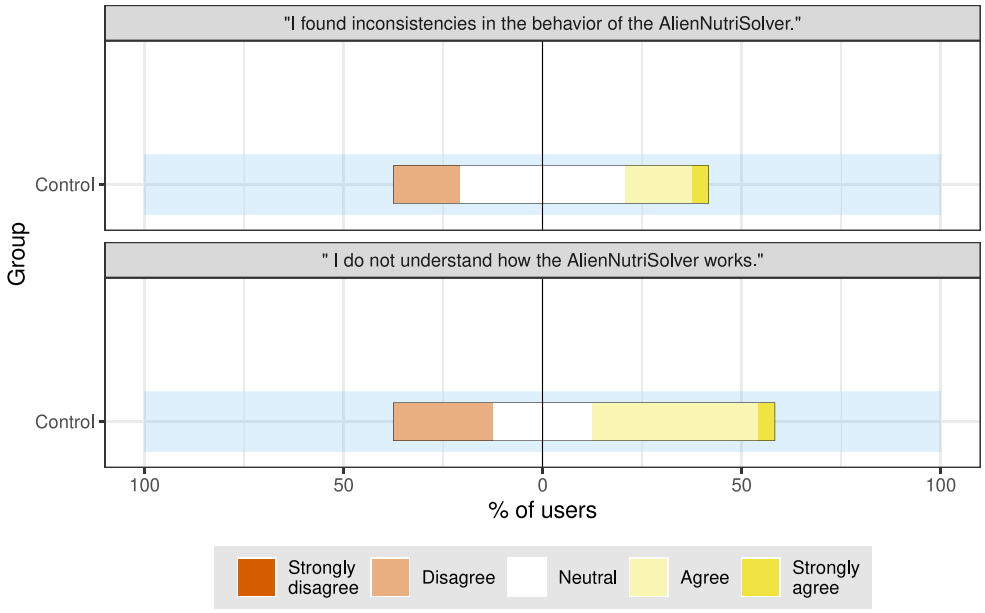

Fig. 12. RQ2 (Satisfaction): Mean responses to survey items 9 and 10 about system robustness and perceived interpretability for the control group.

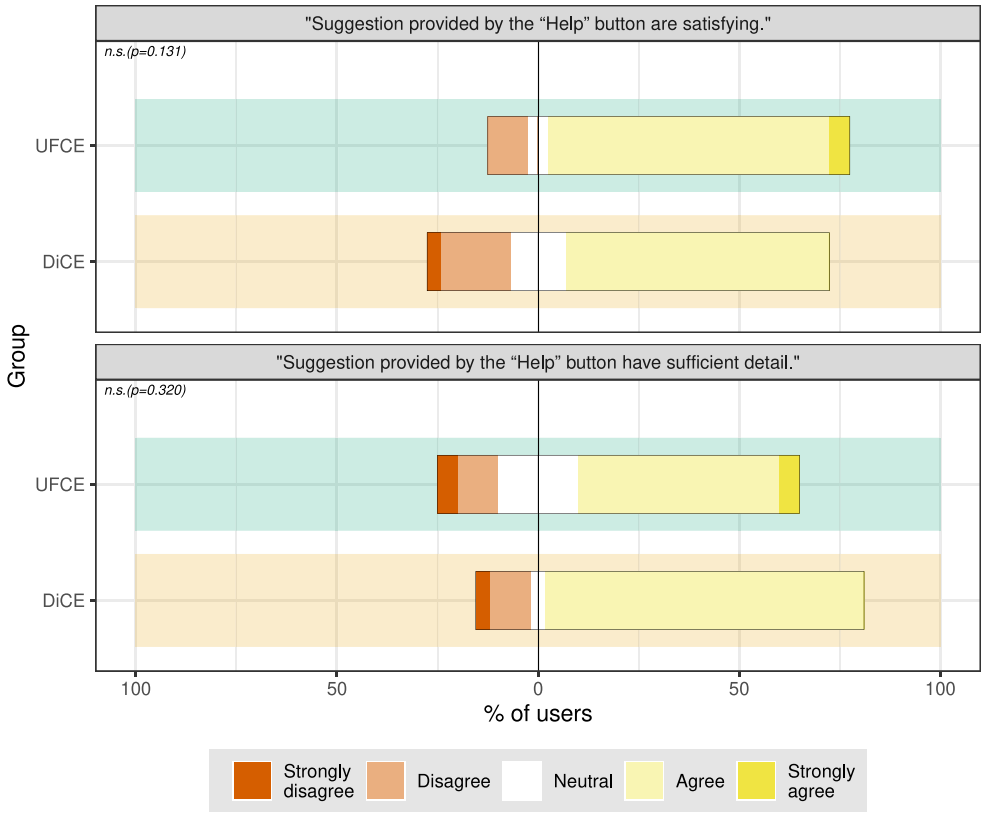

Fig. 13. RQ2 (Satisfaction): Mean responses to survey items 11 and 12 for explanation satisfaction and sufficient details in explanation groups.

trusting the explanation, worry about the reliability of the explanation, and user preference for the explanation.

First, we present the analyses of the results of item 15. Then, we discuss the analyses on the rest of the items.

4.5.1. Trust in the alien nutri-solver framework

The plots in Fig. 15 show the mean trust in the Alien Nutri-Solver framework by all groups.

The analysis revealed that there is no significant difference between the groups regarding trust in the system ($H(2) = 0.003, p = 0.951$).

4.5.2. Trust and confidence in the explanations

The plots in Fig. 16 illustrate mean responses about users' self-reported trust on items 16 and 17 in terms of trust in the explanations and confidence in the explanations, respectively.

We compared the responses of users for item 16 in the UFCE group ($M = 2.25$, $SEM = 0.144$) and in the DiCE group ($M = 2.428$, $SEM =$

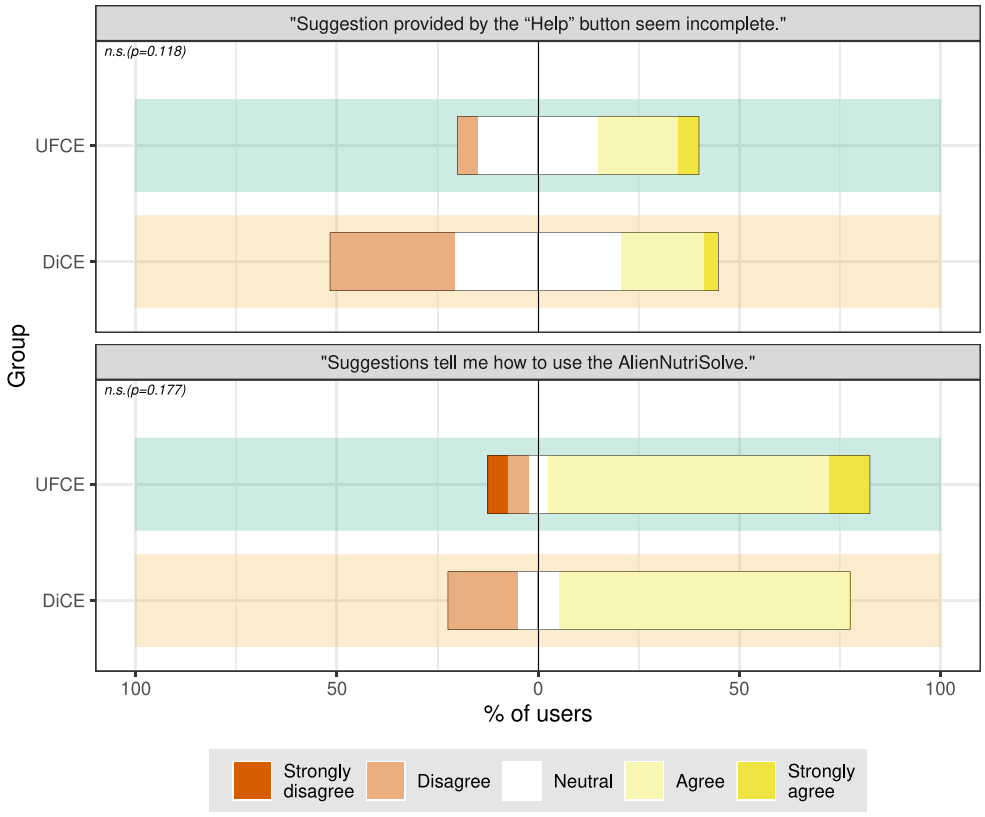

Fig. 14. RQ2 (Satisfaction): Mean responses to survey items 13 and 14 for explanation helpful usage and incomplete suggestion in explanation groups.

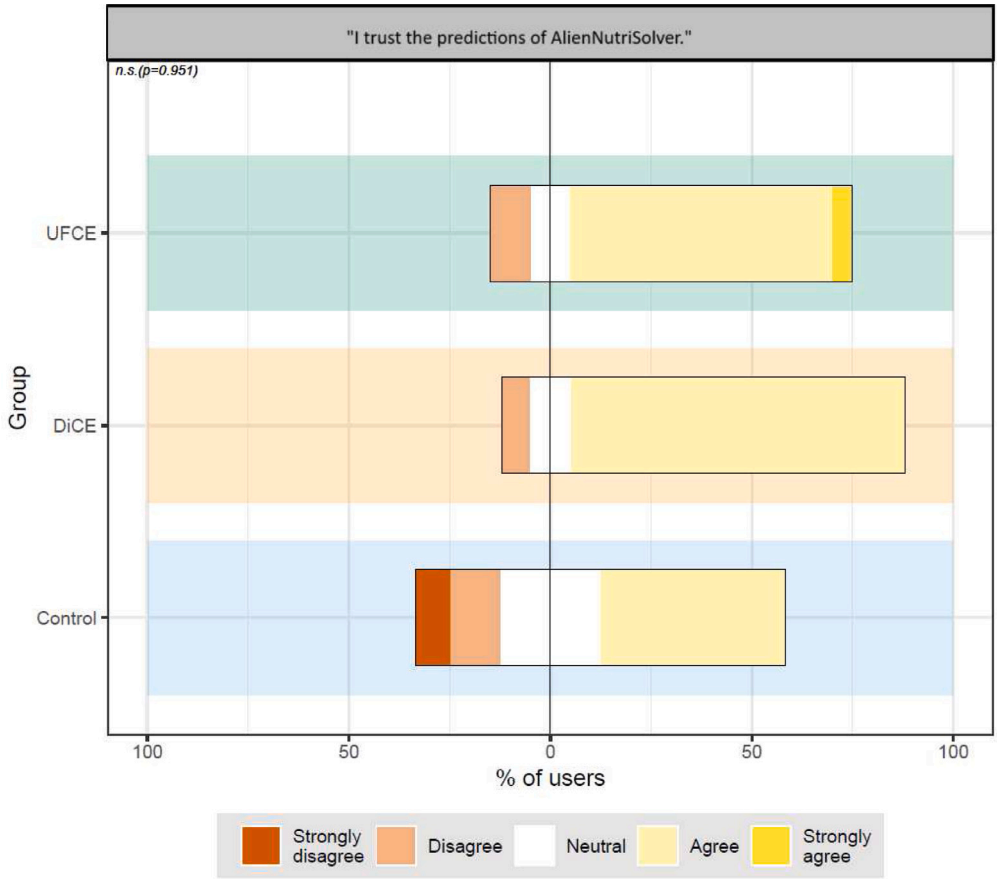

Fig. 15. RQ3 (Trust): Mean trust of users, where n.s. indicates no significant statistical difference.

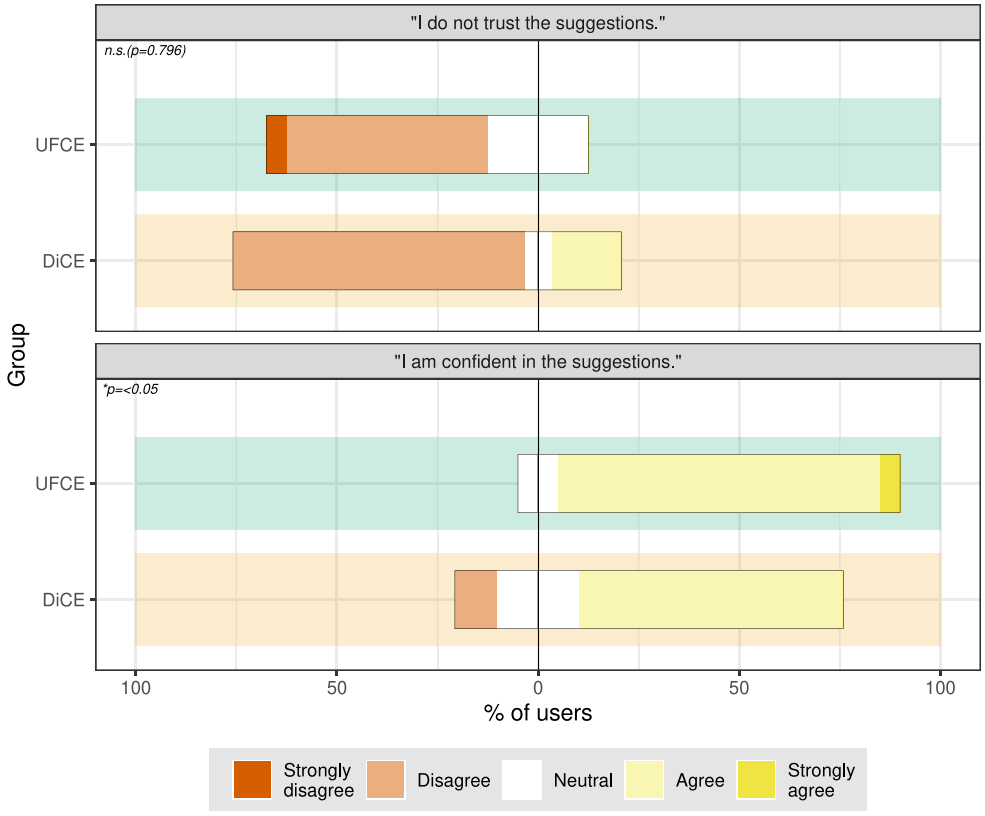

Fig. 16. RQ3 (Trust): Mean responses to survey items 16 and 17 for explanation trust and confidence in explanation groups.

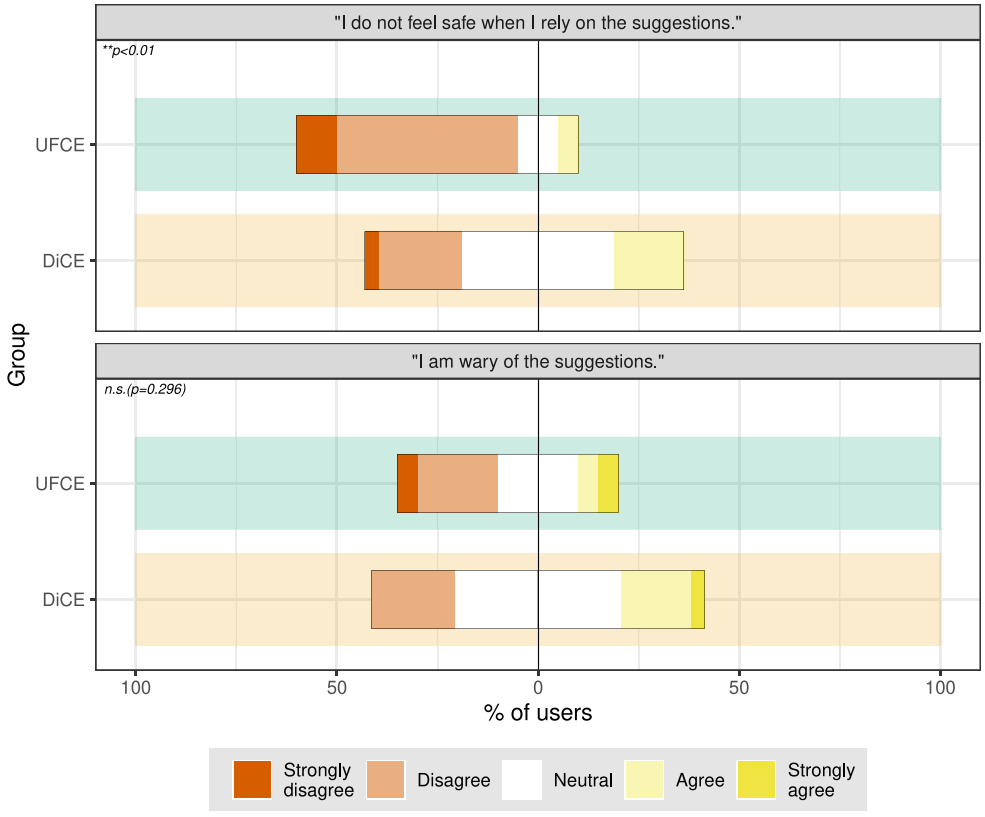

Fig. 17. RQ3 (Trust): Mean responses to survey items 18 and 19 for perceived safety and wary nature in explanation groups.

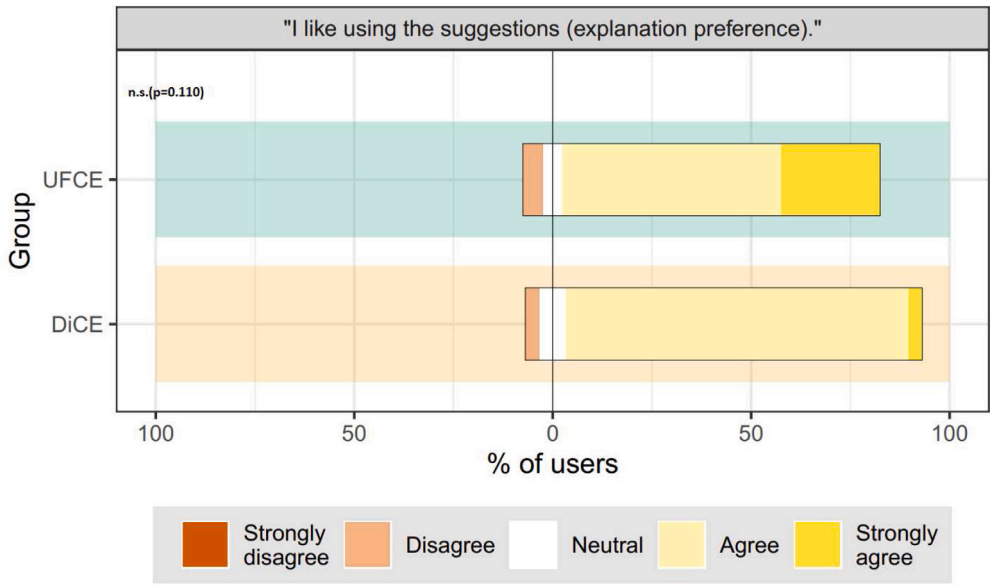

Fig. 18. RQ3 (Trust): Responses to survey item 20, preference of explanations in explanation groups.

0.149). The analysis showed that there is not a significant difference (U = 215,p = 0.796, r = −0.038). Similarly, we compared the responses of users for item 17 in the UFCE group (M = 3.947, SEM = 0.092) and the DiCE group (M = 3.571, SEM = 0.130). The analysis showed that there is a significant difference (U = 336,p = 0.04, r = 0.288). Users in the UFCE group are more confident than users in the DiCE group.

4.5.3. Perceived safety and wary behavior concerning the explanations

The plots in Fig. 17 illustrate mean responses to items 18 and 19 in terms of "feel safe when relying on suggestion" and "wary of suggestions", respectively.

We compared the responses of users for item 18 in the UFCE group (M = 2.142, SEM = 0.205) versus the responses of users in the DiCE group (M = 2.869, SEM = 0.169). The analysis revealed that there is a significant difference (U = 82.5,p = 0.009, r = −0.426). Users using UFCE reported feeling more safe (confident) in finding a suitable solution when relying on explanations, compared to those using DiCE, particularly in the context of identifying a better diet for the Shub. Similarly, we compared the responses of users for item 19 in the UFCE group (M = 2.727, SEM = 0.332) versus the responses of users in the DiCE group (M = 3.041, SEM = 0.164). The analysis showed that there is not a significant difference (U = 104,p = 0.296, r = −0.176).

The survey item 20 is related to the preference of explanations in the explanation groups. Fig. 18 illustrates the mean user responses for preference of explanation for both DiCE and UFCE. The Wilcoxon–Mann–Whitney U test revealed that there is no significant difference (U = 315.5,p = 0.110, r = 0.232) between the user responses in the UFCE group (M = 4.111, SEM = 0.178) versus user responses in the DiCE group (M = 3.896, SEM = 0.090).

The above self-reported answers of users to items 15–20 indicate that users in the UFCE group were more confident in the given explanations than users in the DiCE group. In addition, users of UFCE showed more trust that they feel safe when relying on the suggestions as compared to DiCE users. Thus, UFCE significantly got better responses on two items of trust than DiCE, thereby endorsing the trustworthiness of its explanations.

5. Discussion

In the user study, the primary objective was to assess the ability of UFCE and DiCE to enhance cognitive learning. To do so, we considered human-centered evaluation metrics such as task performance, understanding, satisfaction and trust. Namely, we designed tasks within the game and crafted survey items for post-game evaluation, aiming for both objective and subjective assessment of user experience.

One of the limitations encountered during the data acquisition process was related to technical issues with data export. Specifically, there were instances where a few records became mixed, resulting in data integrity concerns. Consequently, these mixed records had to be excluded from the dataset under study, resulting in the loss of data from three participants. To address this issue, we undertook manual and comprehensive verification processes to ensure the integrity of the data across all groups.

Despite the inherent interdependence among three key features (plants) within the underlying data for the assigned tasks, participants belonging to explanation groups notably improved the health of the Shub throughout the experiment more than participants in the control group.

During the experiment, users within the explanation groups consistently demonstrated their ability to identify relevant plants for the task across multiple attempts. However, it is essential to note that this proficiency should not be immediately interpreted as evidence of users developing comprehensive mental models of the underlying system, given the substantial enhancement in task performance.

In addition to objective measures that quantify system understanding, this study also incorporates various subjective measures through survey items to delve into perceived understanding, satisfaction, and trust. Users' evaluations of the explanations reveal that they found them to be more helpful and usable for improving the fitness of the Shub. Further, a significant proportion of users responded positively regarding their understanding of the explanations.

In summary, findings derived from subjective data are in agreement with findings derived from objective data. The potential reasons for UFCE receiving better evaluation compared to DiCE could be attributed to:

- Survey item 5: Actionability judgment reflects better responses for UFCE, leading to better objective performance compared to DiCE.
- Survey item 7: UFCE users indicated significantly more reliance on suggestions compared to DiCE users.
- Survey item 17: UFCE users exhibited significantly more confidence in the suggestions than DiCE users.
- Survey item 18: UFCE users expressed significantly more feelings of safety when relying on the suggestions compared to DiCE users.

Furthermore, this study found that UFCE users relied significantly more on suggestions compared to DiCE users. Statistical validation confirmed this finding, showing that the mean change in selecting the correct plants when generating explanations was higher for UFCE users than for DiCE users. As a consequence of this engagement, UFCE users perceived the suggestions provided to them as more actionable, consequently yielding superior outcomes in the study.

The controlled nature of tasks designed in this study may come across as artificial. However, such a setup is necessary to isolate and evaluate the specific effects of explanation methods, especially when working with participants who are novices to the task domain. This control minimizes external biases and ensures the validity of comparisons across groups. We argue that having an experimental setting where measuring goodness of explanations do not depend on a priori knowledge of the user, it is essential to draw general conclusions from the study, ensuring the comparison result is fair and independent of the cohort of assessors.

Our conducted user study faced slightly underpowered for medium size effect. Therefore, while the study is well-powered for detecting large effects, caution should be exercised while interpreting the results for medium effects, as they are not detected with high confidence for few items. This analysis highlights the study's strong ability to detect large effects but also reflects limitations when considering more subtle differences.

6. Conclusion

In this paper, the human-centered evaluations of CEs were conducted through a web-based game framework, Alien Nutri-Solver, focusing on various aspects such as task performance, usability, user satisfaction, and trust. A user study is carried out employing between-subjects treatment (mainly comparing UFCE with DiCE). The findings revealed that the participants from the explanation groups notably enhanced the Shub's health throughout the experiment compared to the control group. Further, UFCE users exhibited significantly higher reliance on suggestions compared to DiCE users, supported by statistical validation. Moreover, UFCE users demonstrated superior actionability judgment, higher confidence, and greater feelings of safety compared to DiCE users. Additionally, the heightened engagement with the Alien Nutri-Solver framework led to superior study outcomes, with participants in the UFCE group statistically outperforming participants in DiCE group, and both groups surpassing the control group. These findings are particularly significant given the underlying assumptions, underscore the importance of human-centered XAI, and advocate for meaningful cognitive involvement of users.

CRediT authorship contribution statement

Muhammad Suffian: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ulrike Kuhl:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Alessandro Bogliolo:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Jose M. Alonso-Moral:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Ethics approval and consent of participation

The experiments conducted for human-centered evaluation involved human participants voluntarily according to the protocols approved by the Ethics Committee in the University of Urbino (Approved in session no. 82 of 23 May 2024 of Ethics Committee for Human Experimentation (CESU) at University of Urbino). The study conducted adhered to ethical guidelines, posing minimal risks to participants. Participants electronically consented by signing the agreement form, indicating their understanding and willingness to participate. This information form helped participants familiarize themselves with the study and understand their rights. All data collected was anonymized, and recorded files were transcribed. The information form and informed consent are shown in Appendix, and the anonymized data are available at the following GitHub link: https://github.com/msnizami/HCEvaluations4UFCE/tree/main/User_Study_Data/data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially funded by MIMIT, Italy, under FSC project "Pesaro CTE SQUARE", CUP D74J22000930008. Jose Maria Alonso-Moral was supported by MCIN/AEI/ 10.13039/501100011033 (grant PID2021-123152OB-C21), but also by the Galician Ministry of Culture, Education, Professional Training, and University, Spain (grants ED431C2022/19 and ED431G2019/04), all grants were co-funded by the European Regional Development Fund (ERDF/FEDER program). Ulrike Kuhl was supported by the research training group "Dataninja" (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia, Germany.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijhcs.2025.103484.

Data availability

I have shared the link to data and code related to this article.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M., 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–18.

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160. http://dx.doi.org/10.1007/s11055-020-00914-1.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Ser, J.D., Díaz-Rodríguez, N., Herrera, F., 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Inf. Fusion 101805. http://dx.doi.org/10.1016/j.inffus.2023.101805.

Archibald, M.M., Munce, S.E., 2015. Challenges and strategies in the recruitment of participants for qualitative research. Heal. Sci. J. 11, 34–37.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

Artelt, A., Hammer, B., 2021. Efficient computation of contrastive explanations. In: 2021 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–9.

Buçinca, Z., Lin, P., Gajos, K.Z., Glassman, E.L., 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In: Proceedings of the 25th International Conference on Intelligent User Interfaces. pp. 454–464.

Burkart, N., Huber, M.F., 2021. A survey on the explainability of supervised machine learning. J. Artificial Intelligence Res. 70, 245–317.

Byrne, R.M., 2016. Counterfactual thought. Annu. Rev. Psychol. 67, 135–157.

Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: A survey on methods and metrics. Electronics 8 (8), 832.

Chander, B., John, C., Warrier, L., Gopalakrishnan, K., 2024. Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness. ACM Comput. Surv..

Chen, V., Liao, Q.V., Wortman Vaughan, J., Bansal, G., 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. Proc. ACM Human-Comput. Interact. 7 (CSCW2), 1–32.

Chromik, M., Schuessler, M., 2020. A taxonomy for human subject evaluation of black-box explanations in XAI. In: ExSSATEC@IUI.

Dai, X., Keane, M.T., Shalloo, L., Ruelle, E., Byrne, R.M., 2022. Counterfactual explanations for prediction and diagnosis in XAI. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. pp. 215–226.

Dandl, S., Molnar, C., Binder, M., Bischl, B., 2020. Multi-objective counterfactual explanations. In: Parallel Problem Solving from Nature–PPSN XVI: 16th International Conference, PPSN 2020, Leiden, the Netherlands, September 5-9, 2020, Proceedings, Part I. Springer, pp. 448–469.

Del Ser, J., Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., Saranti, A., Holzinger, A., 2024. On generating trustworthy counterfactual explanations. Inform. Sci. 655, 119898.

Dinno, A., Dinno, M.A., 2017. Package 'dunn. test'. CRAN Repos 10, 1–7.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., et al., 2017. Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics. DSAA, IEEE, pp. 80–89.

Goodman, B., Flaxman, S., 2017. European Union regulations on algorithmic decision-making and a "right to explanation". AI Mag. 38 (3), 50–57.

Guidotti, R., 2022. Counterfactual explanations and how to find them: literature review and benchmarking. Data Min. Knowl. Discov. 1–55. http://dx.doi.org/10.1007/s10618-022-00831-6.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51 (5), 1–42.

Hoffman, G., 2019. Evaluating fluency in human–robot collaboration. IEEE Trans. Human-Mach. Syst. 49 (3), 209–218.

Hoffman, R.R., Miller, T., Klein, G., Mueller, S.T., Clancey, W.J., 2023. Increasing the value of XAI for users: A psychological perspective. KI-Künstliche Intell. 1–11.

Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J., 2018. Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.

Hsiao, J.H., Ngai, H.H.T., Qiu, L., Yang, Y., Cao, C.C., 2021. Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI). arXiv preprint arXiv:2108.01737.

Keane, M.T., Kenny, E.M., Delaney, E., Smyth, B., 2021. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. arXiv preprint arXiv:2103.01035.

Kuhl, U., Artelt, A., Hammer, B., 2022. Keep your friends close and your counterfactuals closer: Improved learning from closest rather than plausible counterfactual explanations in an abstract setting. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. ACM, New York, NY, USA, pp. 2125–2137. http://dx.doi.org/10.1145/3531146.3534630.

Kuhl, U., Artelt, A., Hammer, B., 2023a. For better or worse: The impact of counterfactual explanations' directionality on user behavior in XAI. In: Longo, L. (Ed.), Explainable Artificial Intelligence. Springer Nature Switzerland, Cham, pp. 280–300.

Kuhl, U., Artelt, A., Hammer, B., 2023b. Let's go to the alien zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning. Front. Comput. Sci. 5, 20.

Kunkel, J., Donkers, T., Michael, L., Barbu, C.-M., Ziegler, J., 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–12.

Lai, V., Chen, C., Smith-Renner, A., Liao, Q.V., Tan, C., 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 1369–1385.

Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16 (3), 31–57.

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., et al., 2024. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Inf. Fusion 106, 102301.

McKight, P.E., Najab, J., 2010. Kruskal-wallis test. Corsini Encycl. Psychol. 1.

Mohseni, S., Zarei, N., Ragan, E.D., 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans. Interact. Intell. Syst. (TiiS) 11 (3–4), 1–45.

Mothilal, R.K., Sharma, A., Tan, C., 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, New York, NY, USA, pp. 607–617. http://dx.doi.org/10.1145/3351095.3372850.

Mueller, S.T., Veinott, E.S., Hoffman, R.R., Klein, G., Alam, L., Mamun, T., Clancey, W.J., 2021. Principles of explanation in human-AI systems. In: Proceedings of the AAAI Workshop on Explainable Agency in Artificial Intelligence. AAAI-2020.

Nachar, N., et al., 2008. The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. Tutor. Quant. Methods Psychol. 4 (1), 13–20.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C., 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Comput. Surv. 55 (13s), 1–42.

Ooge, J., Kato, S., Verbert, K., 2022. Explaining recommendations in E-learning: Effects on adolescents' trust. In: 27th International Conference on Intelligent User Interfaces. pp. 93–105.

Papenmeier, A., Englebienne, G., Seifert, C., 2019. How model accuracy and explanation fidelity influence user trust. arXiv preprint arXiv:1907.12652.

Patel, M.X., Doku, V., Tennakoon, L., 2003. Challenges in recruitment of research participants. Adv. Psychiatr. Treat. 9, 229–238. http://dx.doi.org/10.1192/apt.9.3.229.

Pehlivan, C.N., 2024. The EU artificial intelligence (AI) act: An introduction. Glob. Priv. Law Rev. 5 (1).

R Core Team, R., et al., 2013. R: A language and environment for statistical computing. URL: https://www.R-project.org/.

Richardson, G.P., Andersen, D.F., Maxwell, T.A., Stewart, T.R., 1994. Foundations of mental model research. In: Proceedings of the International System Dynamics Conference. EF Wolstenholme, pp. 181–192.

Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., Kasneci, E., 2023. Towards human-centered explainable ai: A survey of user studies for model explanations. IEEE Trans. Pattern Anal. Mach. Intell..

Samek, W., Müller, K.R., 2019. Towards explainable artificial intelligence. Explain. AI: Interpret. Explain. Vis. Deep. Learn. 5–22.

Schoeffer, J., De-Arteaga, M., Kuehl, N., 2024. Explanations, fairness, and appropriate reliance in human-AI decision-making. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–18.

Schoeffer, J., Kuehl, N., Machowski, Y., 2022. "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 1616–1628.

Sokol, K., Flach, P., 2020. One explanation does not fit all. KI-Künstliche Intell. 34 (2), 235–250.

Staggers, N., Norcio, A.F., 1993. Mental models: concepts for human-computer interaction research. Int. J. Man-Mach. Stud. 38 (4), 587–605.

Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M., 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access 9, 11974–12001. http://dx.doi.org/10.1007/s11055-020-00914-1.

Stepin, I., Suffian, M., Catala, A., Alonso-Moral, J.M., 2024. How to build self-explaining fuzzy systems: From interpretability to explainability [AI-eXplained]. IEEE Comput. Intell. Mag. 19 (1), 81–82. http://dx.doi.org/10.1109/MCI.2023.3328098.

Suffian, M., Alonso-Moral, J.M., Bogliolo, A., 2024a. Introducing user feedback-based counterfactual explanations (UFCE). Int. J. Comput. Intell. Syst. 17 (1), 123. http://dx.doi.org/10.1007/s44196-024-00508-6.

Suffian, M., Graziani, P., Alonso, J.M., Bogliolo, A., 2022. FCE: Feedback based counterfactual explanations for explainable AI. IEEE Access 10, 72363–72372. http://dx.doi.org/10.1109/ACCESS.2022.3189432.

Suffian, M., Kuhl, U., Alonso-Moral, J.M., Bogliolo, A., 2024b. CL-XAI: Toward enriched cognitive learning with explainable artificial intelligence. In: Aldini, A. (Ed.), Software Engineering and Formal Methods. SEFM 2023 Collocated Workshops. Springer Nature Switzerland, Cham, pp. 5–27.

Suffian, M., Stepin, I., Alonso-Moral, J.M., Bogliolo, A., 2023. Investigating human-centered perspectives in explainable artificial intelligence. In: CEUR Workshop Proceedings, Vol. 3518. pp. 47–66.

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

Voigt, P., Von dem Bussche, A., 2017. The EU General Data Protection Regulation (GDPR). In: A Practical Guide, Vol. 10, No. 3152676, first ed. Springer International Publishing, Cham, pp. 1–383.

van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M., 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial Intelligence 291, 103404.