frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 21 March 2023
DOI 10.3389/fcomp.2023.1087929

OPEN ACCESS

EDITED BY
Kasun Amarasinghe,
Carnegie Mellon University, United States

REVIEWED BY
Nils Jansen,
Radboud University, Netherlands
Mark T. Keane,
National University of Ireland, Ireland

*CORRESPONDENCE
Ulrike Kuhl
✉ ukuhl@techfak.uni-bielefeld.de

SPECIALTY SECTION
This article was submitted to Theoretical Computer Science, a section of the journal Frontiers in Computer Science

RECEIVED 02 November 2022
ACCEPTED 01 March 2023
PUBLISHED 21 March 2023

CITATION
Kuhl U, Artelt A and Hammer B (2023) Let's go to the Alien Zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning. *Front. Comput. Sci.* 5:1087929. doi: 10.3389/fcomp.2023.1087929

Let's go to the Alien Zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning

Ulrike Kuhl[1,2]*, André Artelt[2] and Barbara Hammer[2]

[1]Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University, Bielefeld, Germany, [2]Machine Learning Group, Faculty of Technology, Bielefeld University, Bielefeld, Germany

Introduction: To foster usefulness and accountability of machine learning (ML), it is essential to explain a model's decisions in addition to evaluating its performance. Accordingly, the field of explainable artificial intelligence (XAI) has resurfaced as a topic of active research, offering approaches to address the "how" and "why" of automated decision-making. Within this domain, counterfactual explanations (CFEs) have gained considerable traction as a psychologically grounded approach to generate *post-hoc* explanations. To do so, CFEs highlight what changes to a model's input would have changed its prediction in a particular way. However, despite the introduction of numerous CFE approaches, their usability has yet to be thoroughly validated at the human level.

Methods: To advance the field of XAI, we introduce the Alien Zoo, an engaging, web-based and game-inspired experimental framework. The Alien Zoo provides the means to evaluate usability of CFEs for gaining new knowledge from an automated system, targeting novice users in a domain-general context. As a proof of concept, we demonstrate the practical efficacy and feasibility of this approach in a user study.

Results: Our results suggest the efficacy of the Alien Zoo framework for empirically investigating aspects of counterfactual explanations in a game-type scenario and a low-knowledge domain. The proof of concept study reveals that users benefit from receiving CFEs compared to no explanation, both in terms of objective performance in the proposed iterative learning task, and subjective usability.

Discussion: With this work, we aim to equip research groups and practitioners with the means to easily run controlled and well-powered user studies to complement their otherwise often more technology-oriented work. Thus, in the interest of reproducible research, we provide the entire code, together with the underlying models and user data: https://github.com/ukuhl/IntroAlienZoo.

KEYWORDS

explainable AI, human-grounded evaluation, user study, experimental framework, counterfactual explanations, usability, human-computer interaction

1. Introduction

In a step toward accountable and transparent machine learning (ML), the European Union mandates safeguards against automated decision-making with the General Data Protection Regulation (GDPR) 2016/679 (European Union, 2016). Specifically, the GDPR states that a person subjected to automated decision-making may obtain an explanation of the given decision. As a result, there has been an upswing of technical explainable artificial intelligence (XAI) approaches on how to make ML explainable (Arrieta et al., 2020; Chou et al., 2022).

Alongside novel explainability approaches, authors have proposed evaluation criteria and guidelines to systematically assess XAI approaches in terms of their usability (Doshi-Velez and Kim, 2017; Arrieta et al., 2020; Davis et al., 2020; Sokol and Flach, 2020a). This theoretical groundwork sparked several practical validation frameworks, commonly evaluating explanations in terms of accuracy and fidelity (White and d'Avila Garcez, 2020; Pawelczyk et al., 2021; Sattarzadeh et al., 2021; Arras et al., 2022), or robustness (Artelt et al., 2021). However, while XAI taxonomies repeatedly emphasize the need for human-level validation of explanation approaches (Doshi-Velez and Kim, 2017; Sokol and Flach, 2020a), user evaluations of XAI approaches often face limitations concerning statistical power and reproducibility (Keane et al., 2021). What is more, counter-intuitive findings underscore the importance of acknowledging the human factor when evaluating XAI approaches (Byrne, 2019). For instance, participants easily simulate predictions of clear models with few features, however, this does not lead users to adjust their behavior more closely in line with the model's predictions, or enable them to detect faulty predictions (Poursabzi-Sangdeh et al., 2021). Similarly, introducing a theoretically motivated plausibility constraint on generated explanations may actually decrease usability for users in certain settings (Kuhl et al., 2022).

Driven by this shift toward a user-centered focus on explainability (Miller, 2019), counterfactual explanations (CFEs) receive special attention as a supposedly useful, human-friendly, and psychologically comprehensible solution (Byrne, 2019; Miller, 2019; Keane et al., 2021). CFEs for ML correspond to *what-if* scenarios, highlighting necessary changes in a model's input that trigger a desired change in the model's output (i.e., "if you earned US$ 200 more per month, your loan would be approved").

The contrastive nature of CFEs, emphasizing why a specific outcome occurred instead of another, strongly resembles human cognitive reasoning (Hilton and Slugoski, 1986; Lipton, 1990; Lombrozo, 2012; Byrne, 2019; Miller, 2019). Humans routinely and automatically engage in counterfactual thinking (Roese, 1997; Goldinger et al., 2003), mentally changing the internal representation of relevant facts to simulate a counterfactual scenario, while maintaining the factual representation in parallel (Byrne, 2016). When doing so, they tend to mentally undo events that are exceptional, controllable, recent, actionable, and plausible (see Byrne, 2019, for a recent review on psychological aspects of CFEs for XAI).

Building on extensive psychological research on counterfactual thinking in humans, Epstude and Roese (2008) emphasize its beneficial role for regulating one's behavior to improve future performance. According to their *Functional Theory of Counterfactural Thinking*, disparities between the current state and an ideal goal state triggers spontaneous counterfactual thought (Roese and Epstude, 2017). In the same vein, Markman and McMullen (2003) argue that this type of comparative thought mode may help to prepare for the future by guiding the formation of intentions, thus changing prospective behavior.

Based on these an similar accounts, XAI research often treat explanations formulated as counterfactuals as intuitively useful, and readily human-usable (Guidotti et al., 2018; Stepin et al., 2019; Artelt and Hammer, 2020; Dandl et al., 2020). However, it is not clear whether these encouraging insights from psychology may be exactly transferred for the use of CFEs in XAI, making it imperative to validate technical CFEs approaches providing explainability for ML models at the user level (Doshi-Velez and Kim, 2017; Offert, 2017). Yet, according to a recent review, only one in three counterfactual XAI papers include user-based evaluations, often with limited statistical validity and little opportunity to reproduce the presented results (Keane et al., 2021).

Studies that do examine CFE approaches from a user-perspective give cause for cautious confidence in their usability. van der Waa et al. (2021) demonstrate that CFEs, compared to example based and no-explanation control variants, enable users interacting with a hypothetical decision support system for diabetes patients to correctly identify features relevant for a system's prediction. Additionally, their data suggest that CFEs have a positive effect on perceived system comprehensibility compared to no explanation. In a scenario similarly inspired by real-life, Warren et al. (2022) present participants with a simulated AI system predicting whether an individual meets the legal blood alcohol content limit to be still allowed to drive. In this setting, users initially presented with either counterfactual or causal explanations of the system's decisions, show greater prediction accuracy themselves, compared to no-explanation controls. While receiving CFEs in contrast to causal explanations provided only a small advantage in terms of objective performance, participants judged counterfactual explanations to be more satisfying and trustworthy than causal ones (Warren et al., 2022). This is in line with earlier work revealing that participants judge counterfactual style explanations to be subjectively more intuitive and useful than, e.g., visualizing feature importance scores (Le et al., 2020).

However, this positive evidence is not unanimous. Users tasked to learn how an automatic system behaves indeed show some understanding of the types of rules governing said system after receiving counterfactual-style explanations, as compared to receiving no-explanation (Lim et al., 2009). Yet, only participants that are presented with explicit feedback stating why the system behaved in a certain way perform consistently better across several metrics, including perceived understanding. In a recent experiment, participants tasked both with prediction and diagnosis show similar performance patterns in both agricultural and abstract domains, independent of receiving CFEs, pre-factual explanations, or mere control descriptions (Dai et al., 2022). Strikingly, user ratings of helpfulness were also comparable across all explanation conditions, at odds with the presumed advantage of CFEs in terms of their usability. Lage et al. (2019) demonstrate that users show consistently higher response times when asked to answer counterfactual-style questions, indicating increased cognitive load for this type of task, a factor that may actually hinder usability.

Even though user evaluations are essential to evaluate the efficacy of explanation modes, designing an effective user study is no easy feat. A well-designed study closely considers the respective explainees, and the reason for explaining (Adadi and Berrada, 2018; Sokol and Flach, 2020a), simultaneously taking into account confounding factors and available resources (Doshi-Velez and Kim, 2017). Further, researchers need to ensure comparable conditions across participants, while systematically varying XAI approaches, underlying ML models, or data distributions.

On top of these methodological challenges, XAI user evaluations often suffer from a series of limitations. While there is general agreement that one explanation mode fitting all scenarios is unlikely to exist (Sokol and Flach, 2020b), not all

studies clearly formulate the given explanation purpose and target group (Le et al., 2020). Without an explicit classification of the experimental context, however, research runs the risk to reach all-too-general conclusions like declaring one mode of explaining universally superior to another. Further, a common methodological limitation concerns limited statistical power due to low participant numbers (Lim et al., 2009; Akula et al., 2020), often driven by the effortful nature of user studies. For reasons of simplicity, some approaches provide participants with explanations that follow a certain XAI approach, but were actually designed by the researchers themselves (Narayanan et al., 2018; Lage et al., 2019; van der Waa et al., 2021). Such a *Wizard of Oz* approach, with a human behind the scenes playing the role of an automatic system (Dahlbäck et al., 1993), allows perfect control over materials encountered by participants. However, it fails to account for potential variability in the results of ML algorithms, imperfectly mimicking the user experience "in the wild."

Moreover, some evaluations exclusively focus on assessing perceived usability. Using questionnaires and surveys is a prominent approach to ask participants how well they like or understand a certain explainability method, but may be affected by response bias and the difficulty of capturing the complexity and nuances of user experiences. Additionally, it is unclear whether subjective evaluations translate into tangible behavioral effects (Hoffman et al., 2018). In fact, a recent study fails to show a correlation between perceived system understandability and any objective measure of performance (van der Waa et al., 2021). Therefore, while surveys are valuable tool to assess users' satisfaction with and trust in an XAI approach in a systematic and quantitative manner, complementing them with behavioral measures is key to draw comprehensive conclusions.

Further, experimental designs in XAI typically let participants passively study pre-selected examples of a system's input and output values, together with a corresponding explanation (Lim et al., 2009; Le et al., 2020; van der Waa et al., 2021). However, evidence from educational science shows that interactive activities grant deeper understanding (Chi and Wylie, 2014), suggesting greater efficacy of designs prompting user action. Last, many designs reported are difficult to exactly reproduce as experimental code, ML models, and underlying data are not openly available. This lack of shared resources severely hampers replication studies and adaptation of frameworks according to novel research purposes near to impossible.

Overall, the lack of openly accessible and engaging user study designs that enable direct comparisons between different CFE implementations, models, and data sets motivates the current work. To advance the field of XAI, we introduce the Alien Zoo, an engaging, web-based and game-inspired experimental framework. The Alien Zoo provides means to evaluate the usability of a specific and very prominent variant of post-hoc, model agnostic explanations for ML, namely CFEs (Artelt and Hammer, 2019), targeting novice users in a domain-general context. Presenting this game-type scenario in a low-knowledge domain, we aim to equip research groups and practitioners with an easily adaptable design. Within the constraints concerning the explanation target (i.e., novice users), and the abstract explanation setting, it is suitable for empirical evaluations of various research questions concerning usability of CFE for XAI. Thus, we aspire to narrow the gulf between the increasing interest in generating human-friendly explanations for automated decision-making, and the limitations given current user-based evaluations.

As a proof of concept, we demonstrate the efficacy and feasibility of the Alien Zoo approach in a user study, showing a beneficial impact of providing CFEs on user performance in the proposed iterative learning task. Providing the entire code, together with the underlying data and scripts used for statistical evaluation,[1] our hope is that this framework will be utilized by other research groups and practitioners.

2. Materials and methods

2.1. The Alien Zoo framework

The increasing number of user studies in the domain of XAI is met by an increasing number of recommendations and guidelines concerning design principles to be taken into account (Davis et al., 2020; Mohseni et al., 2021; van der Waa et al., 2021). When constructing the Alien Zoo framework, we closely follow the recommendations put forward by van der Waa et al. (2021).

2.1.1. Use case and experimental context

The effectiveness of an explanation decisively depends decisively on the reason for explaining, and the intended target audience (Adadi and Berrada, 2018; Arrieta et al., 2020; Mohseni et al., 2021). Both aspects determine the choice of an appropriate use case, and thus the experimental context.

On the one hand, the user's explanation needs may vary dramatically across individuals. Users who already possess a lot of applicable domain knowledge may find more sophisticated explanations more useful than novice users (Doshi-Velez and Kim, 2017). Prior domain knowledge and user beliefs may impact how and even if users meaningfully engage with provided explanations (Lim et al., 2009). Moreover, explainees equipped with AI expertise perceive and evaluate provided explanations differently than users that lack this kind of knowledge (Ehsan et al., 2021). On the other hand, the explanation's purpose profoundly affects requirements a given XAI approach ought to meet. For instance, users tasked to compare different models may benefit from different explanation modes than those who want to gain new knowledge from a predictive model or the data used to build it (Adadi and Berrada, 2018). Consequently, generalizability of conclusions beyond a given use case, context, and target group is limited and needs to be treated with upmost caution (Doshi-Velez and Kim, 2017; Sokol and Flach, 2020a).

In the Alien Zoo framework, we focus on an abstract experimental context: Participants act as zookeepers for a fictional alien species called shubs (Figure 1A). The participants' main task is to determine the best combination of plants to feed them (Figure 1B). Importantly, for a user starting the game, it is not clear what plants (or which plant combination) makes up a nutritious

1 Available at: https://github.com/ukuhl/IntroAlienZoo.

diet, causing their pack to thrive. In regular intervals, participants receive CFEs together with their past choices, highlighting an alternative selection that would have led to a better result. Thus, the current use case is that of assisting novice users without any prior experience to gain new knowledge from a predictive model about the data used to build it. Consequently, Alien Zoo user studies correspond to human-grounded evaluations, with participants engaging in "counterfactual simulation" (Doshi-Velez and Kim, 2017). Thus, our setting falls into the "explaining to discover" category for explainability, evaluating whether providing CFEs to novice users enhances their ability to extract yet unknown relationships within an unfamiliar dataset (Adadi and Berrada, 2018).

2.1.2. Constructs and their relations

Clear definitions of utilized constructs and their interrelations are crucial to enrich XAI user evaluations with a solid basis for scientific theory (van der Waa et al., 2021). Alien Zoo user evaluations focus on three constructs: subjective usability, system understanding, and task performance. Figure 2 depicts a causal diagram showing the expected relations between these constructs. Specifically, we posit that providing CFEs positively impacts a user's system understanding, as well as their subjective usability. Consequently, increased system understanding will enable users to better perform the task at hand.

The given proof of concept study described in Section 2.2 compares user performance when receiving CFEs with a no explanation a control. When provided with CFEs, we expect participants to gain a better understanding of decisive features, and the best combination thereof, in the underlying data. Consequently, we anticipate increased system understanding to improve task performance. Given how humans engage in counterfactual thinking automatically on a day-to-day basis (Sanna and Turley, 1996; Roese, 1997; Goldinger et al., 2003), we expect that explanations formulated as counterfactuals also have a positive impact on subjective understanding.

Finally, it is crucial to consider subjective usability as a construct separate from system understanding. A participant's action does not necessarily correspond to their perceived system understanding, strongly suggesting that user behavior and self-report do not measure the same construct (van der Waa et al., 2021).

2.1.3. Measurements

The Alien Zoo framework assesses the constructs subjective usability, system understanding, and task performance through objective behavioral and subjective, self-report measures (Figure 2). Given evidence of a disparity between perceived system understanding and objective performance (van der Waa et al., 2021), addressing both aspects may provide a holistic usability assessment of CFE for ML.

First, we expect participants to recall and apply the information provided by CFEs to improve their feeding choice. Ultimately, this translates to an increase in the participant's capacity to correctly identify the decisive factors in the data used to train the shub growth model, both in the study game and survey phase. While this capacity is not directly measured during the game, we acquire corresponding self-reports *via* the post-game survey, also determining to what extent users develop an explicit understanding of the data structure.

Second, we expect that system understanding has a positive effect on task performance. Measures assessing task performance include the number of aliens in the pack over the duration of the game (henceforth referred to as pack size). This value indirectly quantifies the extent of user's understanding of relevant and irrelevant features in the underlying data set, as a solid understanding leads to better feeding choices. Similarly, we expect time needed to reach a feeding decision over trials to be indicative of how well-participants can work with the Alien Zoo (henceforth referred to as decision time). As we assume participants to become more automatic in making their plant choice, we expect this practice effect to be reflected as decreased decision time (Logan, 1992).

Third, self-reports acquired *via* the post-game survey assess different aspects of how participants judge the subjective usability of explanations provided (for a full list of all survey items, see Supplementary Table 1).

2.1.4. System implementation

The implementation of the Alien Zoo realizes a strict separation of the front end creating the game interface participants interact with, and the back end providing the required ML functionality, a webserver hosting the study, and databases for data acquisition. The web interface employs the JavaScript-based Phaser 3, an HTML5 game framework.[2] The back end of the system is Python3-based (Python Programming Language; RRID:SCR_008394), with the sklearn package (RRID:SCR_019053; Pedregosa et al., 2011) supporting ML processes. An underlying ML model trained on synthetic plant data to predict the alien pack's growth rate determines the behavior of the game. This model receives input from the user end to update the current number of shubs. To ensure flexibility in terms of potential models, we employ the CEML toolbox (Artelt, 2019) to compute CFEs.[3] CEML is a Python toolbox for generating CFEs, supporting many common machine learning frameworks to ensure availability of a wide range of potential ML algorithms. Thus, the Alien Zoo provides a highly flexible infrastructure to efficiently investigate different intelligibility factors of automatically generated CFEs.

2.1.5. Advantages and limitations of the Alien Zoo framework

The proposed framework offers several advantages for investigating the usability of CFEs in XAI.

First, the abstract nature of the task eliminates potential confounding effects of prior user knowledge: it is safe to say that any user is a novice when it comes to feeding aliens, eliminating the possibility of misconceptions or prior beliefs.

2 https://phaser.io/

3 https://github.com/andreArtelt/ceml

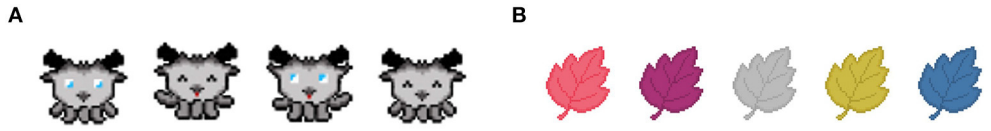

FIGURE 1
Integral components of the Alien Zoo framework: **(A)** An exemplary group of shubs, the small alien species inhabiting the zoo. **(B)** Plants available to the participants for feeding.

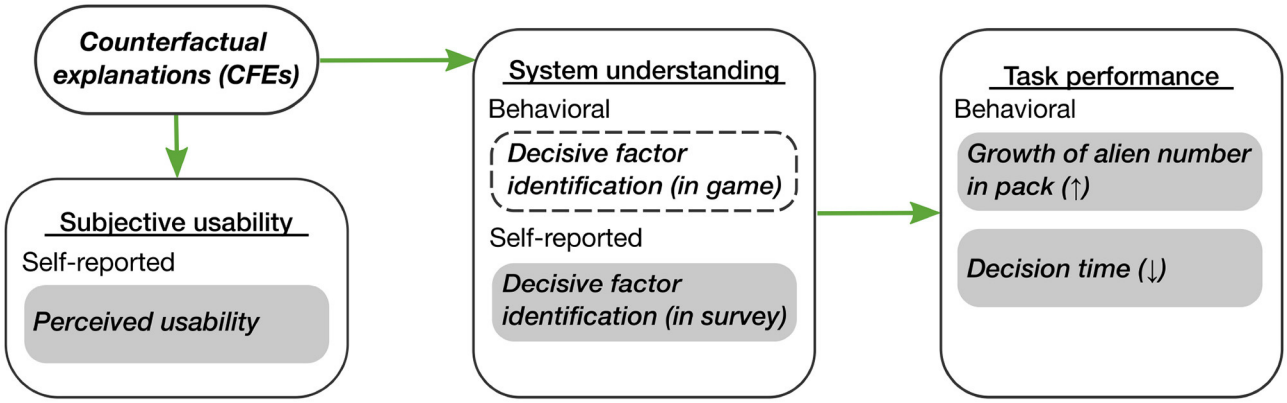

FIGURE 2
Causal diagram describing expected effects of counterfactual explanations on the constructs subjective usability, system understanding, and task performance investigated in the Alien Zoo framework. Green arrows depict expected positive effects. Opaque gray boxes show the measures for the respective construct, divided into behavioral and self-reported measurements. Arrows behind measures depict the expected direction of positive effects. The dashed box shows a behavioral factor reflecting improved system understanding indirectly assessed *via* its mediating role on task performance.

Second, the interactive and game-like design promotes user engagement over iterative rounds of user action and feedback. Evidence from educational science motivates this choice, demonstrating that learner's level of commitment affects the learning outcome, and that interactive activities foster deeper understanding (Chi and Wylie, 2014).

Third, the web-based infrastructure allows for prompt data collection from a large number of participants, which is essential for gathering results with sufficient statistical power, a common shortcoming of previous work (Keane et al., 2021). For instance, data acquisition from 90 participants in the current proof of concept study (Section 2.2) took five days *via* Amazon Mechanical Turk (AMT), including the initial quality assessment. Acknowledging concerns regarding validity of data acquired this way, we demonstrate the feasibility of using the given online approach to obtain meaningful data if appropriate quality measures are in place (see Section 2.2.4).

Fourth, the Alien Zoo confronts participants with feedback from real XAI methods based on reproducible ML models. Acknowledging that *Wizard of Oz* designs are the preferable option in terms of control over what participants experience and consistency of presented explanations (Browne, 2019; Jentzsch et al., 2019), we posit that employing explanations genuinely produced from ML models provides more truthful insights into the user-AI-experience.

Moreover, the Alien Zoo combines both subjective and objective measures to gain a comprehensive understanding of CFE usability in XAI. The post-game survey quantifies how users experience the task, and in how far they deem information provided by CFEs helpful and usable. Importantly, these subjective measures are complemented by objective measures (i.e., task performance, and decision time). Given that the complex association between subjective evaluation and user behavior is yet poorly understood (Hoffman et al., 2018), the proposed framework opens up relevant research opportunities.

Last, in the interest of reproducible research, we fully share data and code of the Alien Zoo framework on GitHub (RRID:SCR_002630; https://github.com/ukuhl/IntroAlienZoo), encouraging research groups and practitioners alike to adapt and utilize the implementation.

Naturally, while addressing prominent limitations of previous work, the Alien Zoo framework is not without constraints itself.

One of the largest limitations of our paradigm concerns its generality. In its current form, the Alien Zoo provides usability insights when providing CFEs as feedback in an iterative learning design, targeting an abstract domain for novice users. Whether and how far observations based on the proposed design generalize to other tasks, domains and target groups, remains to be shown in dedicated validation studies adapting the specific aspects in question. Likewise, while the proposed framework promises to be a valuable tool for investigating specific aspects of CFEs in XAI,

it does not provide a patent solution for addressing all potentially relevant dimensions. For instance, all Alien Zoo features share the same, continuous feature type (i.e., number of leaves per plant), impeding investigations of more fundamental feature properties that have recently been shown to impact understanding (Warren et al., 2022), and thus the effectiveness of an explanation in a current setting.

2.2. Empirical proof of concept study

In the following, we empirically investigate the efficacy and feasibility of the Alien Zoo framework. To this effect, to employ it to run a user study examining the impact of providing CFEs on user performance as compared to no explanations in the proposed Alien Zoo iterative learning task. The study consists of two experiments that primarily vary in terms of the complexity of the underlying data used for model building. Specifically, growth rate in Experiment 1 depends on the best combination of three plants, while this is reduced to the best combination of two plants in Experiment 2 (see Section 2.2.6). The critically low learning rate of control participants in Experiment 1 motivated our decision to run the second experiment. Thus, we investigated whether users that do not receive explanations generally fail to learn in the Alien Zoo setting, even given a simpler configuration.

2.2.1. Hypotheses

The guiding question of the empirical proof of concept study is whether users benefit from receiving CFEs when tasked to identify relationships within an unknown data set when interacting with the Alien Zoo framework.

We evaluate this question using an interactive iterative learning task, in which users repeatedly select input values for an ML model. Throughout the experiment, users receive feedback at regular intervals. Either we show them an overview of their choices alone (control condition), or we show them this overview alongside CFEs, highlighting how changes in their past choices may have led to better results (CFEs condition). *Via* this approach, the interaction between repeated actions by users and corrective feedback allows us to assess system understanding objectively through task performance over a series of decisions.

We hypothesize that providing CFEs compared to no explanations indeed helps users in the task at hand. Specifically, we assume that exposure to alternative feeding choices that would lead to better results enables users to build a more accurate mental model of the underlying data distribution.

We recruited novice users and designed the task around an abstract scenario in order to gain insight into the usability of CFEs. By using this approach, we can protect against possible differences in domain knowledge or misconceptions about the task setting that might impact task performance (van der Waa et al., 2021).

Consequently, we address the following three hypotheses.

2.2.1.1. Hypothesis 1

We expect users that receive CFEs on top of a summary of their past choices to outperform users without explanations in discovering unknown relationships in data, both in terms of objective and subjective measures. Specifically, we anticipate that participants in the CFEs condition (a) produce larger pack sizes, thus showing greater learning success, (b) become faster in making their choice as a sign of more automatic processing, and (c) are able to explicitly identify relevant and irrelevant input features.

2.2.1.2. Hypothesis 2

In terms of subjective understanding, we predict a marked group difference. We expect that users that receive CFEs will subjectively find their feedback more helpful and usable. Furthermore, we posit that those users will also judge CFE feedback to be more helpful for other users.

2.2.1.3. Hypothesis 3

Both feedback variants (overview of past choices and overview of past choices + CFEs) are relatively straight-forward. Thus, when evaluating users' understanding of the feedback themselves, and their evaluation of timing and efficacy of how feedback is presented, we do not expect to see group differences. Further, it will be interesting to see if users differ in terms of needing support to understand the provided feedback. In the CFE condition, it is conceivable users may wish for additional help for interpreting this added information.

2.2.2. Participants

We conducted the study in early March 2022 on AMT. We restricted access to the study to users that (a) belong to the high performing workers on the platform, and have been granted the Mechanical Turk Masters Qualification, (b) have a work approval rate of at least 99%, and (c) did not participate before in any Alien Zoo tasks we ever ran on AMT.

For each experiment, we recruited 45 participants, randomly assigned to either *CFE* or *control* (i.e., no explanation) group. We based the choice for this relatively small sample size on the results of an a-priori power analysis run based on pilot data from a previous experiment employing the same framework (Kuhl et al., 2022). The corresponding a-priori power analysis showed that power levels surpassed 80% for sample sizes >40 by trial 11, assuming a medium effect size. In the current study, we expected a quite large effect given the assumed boost in performance by receiving explanations compared to receiving no additional information. Thus, we decided to close data acquisition after 45 participants submitted their data in each experiment. Note, however, that a particular advantage of the proposed design is its online nature, allowing comfortable acquisition of larger data sets for well-powered studies. All participants gave informed electronic consent by providing click wrap agreement prior to participation. Participants received payment after first data quality assessment.

Contributions from participants whose data showed insufficient quality (see Section 2.2.4) were rejected. Affected users received US$ 1 base compensation for participation, paid *via* the bonus system. This concerned 6/45 (Experiment 1) and 2/45 (Experiment 2) participants, respectively. All remaining participants received a base pay of US$ 3 for participation. The five best performing users in each experiment received an additional bonus of US$ 1. We included information about the prospect of a bonus in the experimental instructions, to motivate users to

comply with the task (Bansal et al., 2019). The Ethics Committee of Bielefeld University, Germany, approved this study.

2.2.3. Experimental procedure

The experiment consists of a game and a survey phase. Accepting the task on AMT redirects participants to a web server hosting the study.

Users are first notified of the purpose, procedure, and expected duration of the study, their right to withdraw, confidentiality and contact details of the primary investigator. If a user does not wish to participate, they may close the window. Otherwise, users confirm their agreement *via* button press. As soon as they indicate agreement, participants get secretly allotted to one of the experimental conditions (either *control* or *CFE*) *via* random assignment.

A subsequent page provides detailed information about the Alien Zoo game. Specifically, it illustrates images of the aliens to be fed, and the variety of plants they may use for feeding. Written instructions state that a pack size can be increased or decreased by choosing healthy or unhealthy combinations of leaves per plant. The maximal number of leaves per plant is limited to six, and users may freely select any combination of plants they find preferable. Subsequent instructions direct the user to maximize the number of aliens, so-called shubs, in order to qualify as a top player to receive an additional monetary bonus. Further, written information establishes that participants will receive a summary of their past choices after two rounds of feeding. Users in the CFEs condition also learn that they will be provided with feedback on what choice would have led to a better result on these occasions.

Clicking a "Start" button at the end of the page indicates that the user is ready to start the game phase. This button appears with a delay of 20 s in an effort to prevent participants from skipping the instructions.

2.2.3.1. Game phase

Figure 3 visualizes the general flow of scenes displayed during the game phase. This phase begins with a padlock scene, where participants make their first feeding selection (left side in Figure 3 and Supplementary Figure 1). All available plant types alongside upward and downward arrow buttons appear on the right side of this scene. The same leaf icon in different colors represents the different plants (Figure 1B). While each participant encounters the same 5 plant colors, their order is randomized for each participant in order to avoid confounding effects. During the first trial, the top of the page notes that clicking on the upward and downward arrows increases and decreases the number of leaves of a specific plant, respectively. In each subsequent trial, the top of the page holds a summary of the previous trial's choice, together with the previous and current pack size. Furthermore, the page shows a padlock displaying the current pack of animated shubs. Participants receive a pack of 20 shubs to begin. Participants submit their choice by clicking a "Feeding time!" button in the bottom right corner of the screen.

While users watch a short progress scene, the underlying ML model predicts the new growth rate based on the user's input. Our implementation subsequently updates the pack size based on the model's decision, and computes a CFE. Within three seconds, a new padlock appears, visualizing the impact of the current choice in terms of written information and animated shubs. The choice procedure repeats after odd trials.

Users receive feedback after even trials, accessible *via* a single "Get feedback!" button replacing the choice panel on the right-hand side of the screen. The feedback button directs users to an overview of past two feeding choices, and the impact on pack size. Users in the *CFE* condition are additionally presented with the intermittently computed CFEs, illustrating an alternative choice that would have led to a better result for each of the past two trials. If users select a combination of plants that lead to maximal increase in pack size, no counterfactual will be computed. In these cases, users learn that they were close to an optimal solution in that round.

Hitting a "Continue!" button appearing after 10 s on the right-hand side of the screen, users proceed with the next trial, encountering a new padlock scene. We included this delay to ensure that users spend sufficient time with the presented information to be able to draw conclusions for their upcoming feeding decisions. Each experiment in this paper consists of 12 trials (i.e., 12 feeding decisions). Users receive feedback after even trials.

Two additional attention checks assess attentiveness of users during the game phase, implemented after trials 3 and 7. Said attention checks request participants to type in the current number of shubs in the last feeding round. Participants receive immediate feedback on the correctness of their answer, alongside a reminder to stay attentive to every aspect of the game at all times. The game then continues with the subsequent progress scene.

After the user made 12 feeding decisions, the game phase of the study ends.

2.2.3.2. Survey phase

In the survey phase, users answer a series of questions. Survey items first assess user's explicit knowledge of plant relevance for task success (items 1 and 2), and second subjective judgements of usability and quality of feedback provided *via* an adapted version of the System Causability Scale (Holzinger et al., 2020).

A final set of three self-report measures assesses potential confounding factors. They address whether users understand the feedback provided, whether they feel they need support for understanding it, and how they evaluate the timing and efficacy of feedback. The last two items of the survey phase collect demographic information on participant's gender and age.

On the final page of the study, users are thanked for their participation and receive a unique code to provide on the AMT platform to prove that they completed the study and qualify for payment. To ensure anonymity, we encrypt payment codes and delete them as soon as users received payment.

Finally, participants may choose to follow a link providing full debriefing information.

2.2.4. Data quality criteria

Due to the nature of web-based studies, some users may attempt to game the system, claiming payment without providing adequate answers. Thus, a priori defined criteria ensure sufficient data quality.

Users qualify as speeders based on their decision time in the padlock scene, if they spent less than two seconds to make their

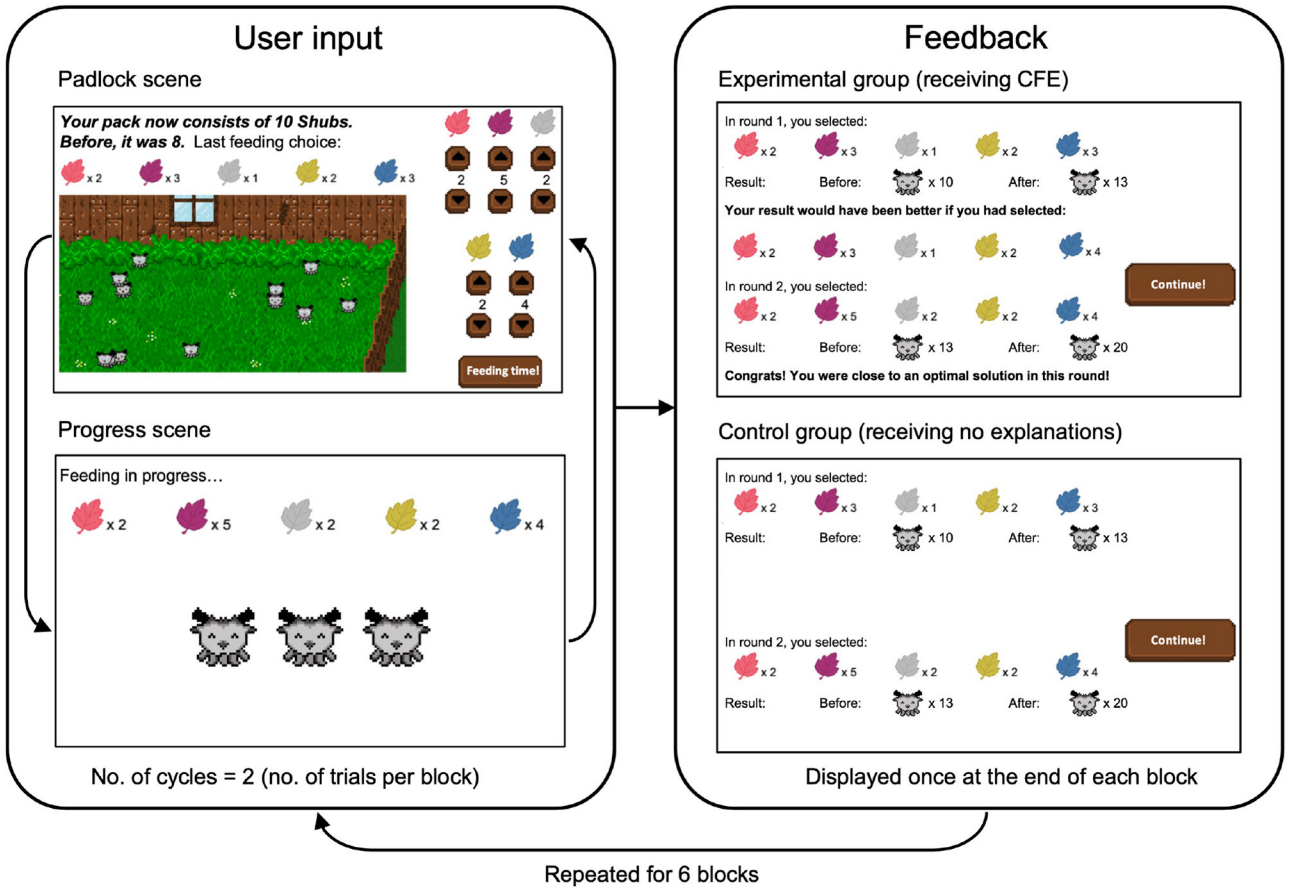

FIGURE 3
General flow of scenes displayed during the game phase. Note that this pattern was disrupted after trials 3 and 7 for an additional attention scene, asking participants to indicate the current number of shubs in their pack. A more detailed overview of scenes within a block can be found in Supplementary Figure 1.

plant selection in at least four trials. Users qualify as inattentive participants if they fail to give the correct number of shubs in both attention trials (game phase). Likewise, we categorize participants as inattentive users if they fail to select the requested answer when responding to the catch item in the survey phase (see Supplementary Table 1). Finally, users qualify as straight-liners if they keep choosing the same plant combination despite not improving in three blocks or more (game phase), or if they answer with only positive or negative valence in the survey phase.

By excluding data of individuals that were flagged for at least one of these reasons from further analysis, we maintain a high level of data quality.

2.2.5. Statistical analysis

We perform all statistical analyses using R-4.1.1 (R Core Team, 2021, R Project for Statistical Computing; RRID:SCR_001905), using experimental condition (*control* and *CFE*) as independent variable. Staying true to our longitudinal design, linear mixed models examine effects of experimental condition over the 12 experimental trials (R package: lme4 v.4_1.1-27.1; RRID:SCR_015654) (Bates et al., 2015). In the model evaluating differences in terms of user performance, number of shubs generated serves as dependent variable. In the model evaluating differences in terms of user's reaction time, decision time in each trial serves as dependent variable. Each model includes the fixed effects of group, trial number, and their interaction. The random-effect structure includes a by-subjects random intercept. We decided to follow this approach, as linear mixed models account for correlations of data drawn from the same participant (Detry and Ma, 2016; Muth et al., 2016). To compare model fits, we rely on the analysis of variance function of the stats package in base R. η_p^2 values denote effect sizes (effectsize v.0.5) (Ben-Shachar et al., 2020). We follow up significant main effects or interactions by computing pairwise estimated marginal means, with respective effect sizes reported in terms of Cohen's *d*. To account for multiple comparisons, all *post-hoc* analyses reported are Bonferroni corrected.

We evaluate data acquired during the survey phase depending on item type. The first two items assess user's explicit knowledge of plant relevance, or irrelevance, for task success.

We aim to obtain a unified measure of user knowledge, appreciating correct answers but also penalizing incorrect ones. Therefore, we use the number of matches between user input and ground truth (i.e., number of plants correctly identified as relevant or irrelevant) per participant per item. Distributions of match data

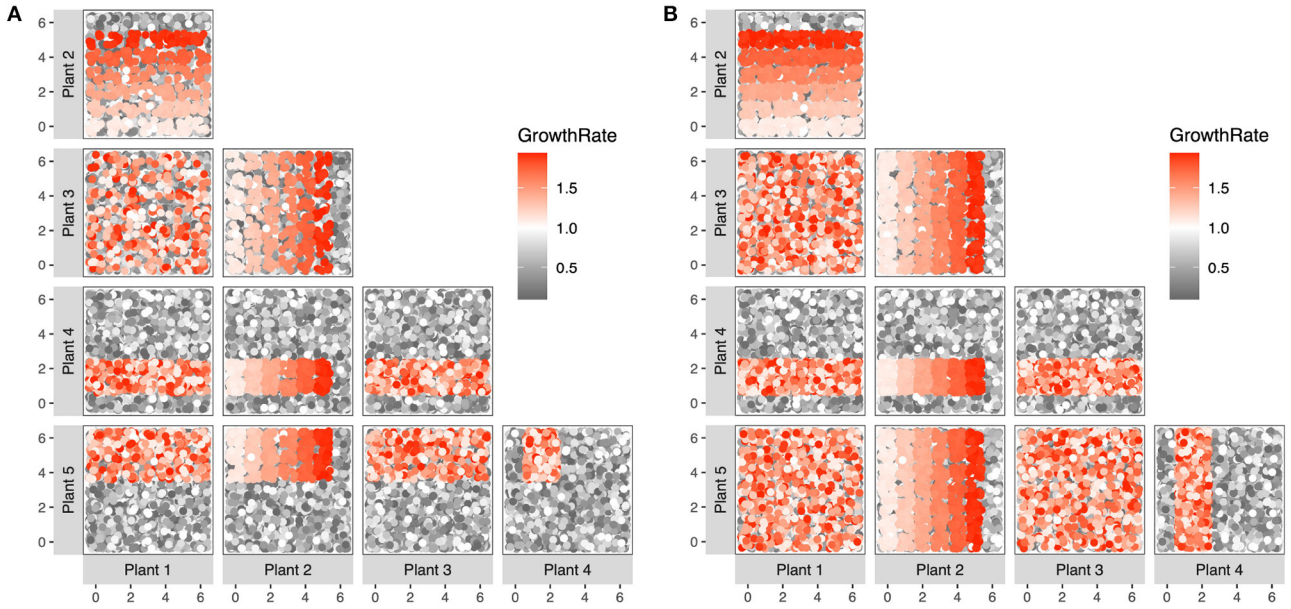

FIGURE 4
Distribution of synthetic data used for model training. Each point in each scatter plot represents the combination of two plant values, colored according to the corresponding growth rate of that point. Gray values indicate growth rate values below 1 (leading to pack size decreases), and red values code values above 1 (leading to pack size increases). **(A)** Experiment 1: The growth rate scales linearly with plant 2, depending on values of plant 4 and plant 5. **(B)** Experiment 2: The growth rate scales linearly with plant 2, depending on values of plant 4. For clear rendering, only 0.2% of all training data are shown, with data points jittered around their true integer values.

was tested for normality using the Shapiro-Wilk test, followed up by the non-parametric Wilcoxon-Mann-Whitney U-test in case of non-normality, and the Welch two-sample t-test otherwise for group comparisons. We follow the same approach to compare age and gender distributions. Finally, we gauge group differences of ordinal data from the Likert-style items, using the non-parametric Wilcoxon-Mann-Whitney U-test. Effect sizes for all survey data comparisons are given as r.

2.2.6. Models

To predict the growth rate and thus ultimately the new pack size given the user input in each trial, we train a decision tree regression model for each experiment. Decision trees consecutively split the data along a series of if-then-else rules, thus approximating the underlying data distribution (Shalev-Shwartz and Ben-David, 2014). Decision trees are powerful enough to model our synthetic data set with sufficient accuracy, while allowing for efficient computation of CFE (Artelt and Hammer, 2019).[4] The current implementation uses the Gini splitting rule of CART (Breiman et al., 1984). To maintain comparable model outputs for all users within throughout one experiment, we use the same decision tree model once build in the beginning.

4 Note, however, that the Alien Zoo framework itself does not depend on a specific model, and could potentially used with other regression models as well.

2.2.6.1. Hyperparameter tuning

To ensure the models reliably present the respective underlying data structure without overfitting, we choose tree depth that yield a high R^2 value and minimizes the mean squared error (MSE) when evaluated on test data. As a further sanity check, we ensured that inputting the perfect solution into the model reliably yields no CFE (i.e, eliciting the feedback that one is close to an optimal solution). Overly complex models are prone to overfit, picking up dependencies in the structure of the randomly chosen features. CFEs generated on the basis of such a model may suggest changes in irrelevant features, thus leading participants on a garden-path. Thus, for the more complex data set in Experiment 1, we use a maximal tree depth of 7 (model performance on test data: $R^2 = 0.893$, MSE $=$ 0.037), while the tree model in Experiment 2 was trained with a maximal tree depth of 5 (model performance on test data: $R^2 = 0.888$, MSE $= 0.039$).

2.2.6.2. Training data

The underlying data in Experiment 1 were generated according to the following scheme: The growth rate scales linearly with values 1–5 for plant 2, iff plant 4 has a value of 1 or 2 AND plant 5 is not smaller than 4 (Figure 4A). For Experiment 2, we reduced the dependency to two relevant features, such that growth rate scales linearly with values 1–5 for plant 2, iff plant 4 has a value of 1 or 2 (Figure 4B). In both experiments, the linear relationship does not hold for value 6 of plant 2, to prevent a simple maximization strategy with respect to this feature.

Growth rate may take a value between 0.1 and 1.9.[5] In each trial, the respective model predicts the new growth rate based on the current user input. Subsequently, the new growth rate (range 0.1–1.9) is converted into a corresponding value between −10 and 10 in our implementation, that gets then added to the current number of shubs to update the pack size. Note that our implementation prevents pack size from shrinking below two.

Each synthetic data set contains all possible plant—growth rate combinations 100 times, yielding 1,680,700 data points. For final model training, we balance the data set by first binning the samples based on their label (growth rate), and then applying Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al., 2002) using the bins as class labels. The final data set is obtained by removing the binning.

3. Results

The empirical part of the current paper investigates whether the proposed Alien Zoo framework is suitable to study the effect of providing automatically generated CFEs for users tasked to learn about yet unknown relationships in a data set. We used an abstract setting to circumvent any confounding effects from previous knowledge of the users.

3.1. Experiment 1

In Experiment 1, we acquired data from 45 participants (Table 1), tasked to identify relationships within an unknown data set. To ensure sufficient task complexity, we opted for a comparatively complex interdependence of three features.

3.1.1. Participant flow

From 45 participants recruited *via* AMT, we exclude data from participants who failed both attention trials during the game ($n = 2$), and straight-lined during the game despite not improving ($n = 4$). No participant in this cohort qualified as speeder, gave an incorrect response for the catch item in the survey, or straight-lined in the survey. Thus, the final analysis includes data from 39 participants (Table 1). Note that for one user in the CFE condition, logging of responses for the first two survey items ("Which plants were [not] relevant to increase the number of Shubs in your pack?") failed. Thus, we excluded this user in the evaluation of these two items, but included them in all remaining analysis.

On average, the final 39 participants in Experiment 1 needed 17 m:42 s (±01 m:16 s SEM) from accepting the task on AMTs to inserting their unique payment code.

5 Originally, the prediction was conceived to be used as a factor, enabling exponential growth in perfect cases. This was changed because it meant that individual people might achieve very high pack sizes, in turn disproportionally driving potential effects.

3.1.2. Objective measures of usability

Hypothesis 1 posits that users benefit from receiving CFEs compared to no explanations in the Alien Zoo framework. To address this hypothesis, we compare data from participants in both groups in terms of pack size produced over time, decision time, and matches between ground truth and indicated plants. Figure 5A depicts the development of average pack size as well as average decision time per group. While users receiving CFEs clearly show a positive trajectory, users receiving no explanation did not show any trace of improvement over the course of this experiment. In fact, no user in the control condition managed to increase their pack size from the minimal attainable number of two by trial 12. A significant interaction of factors trial number and group [$F_{(11,407)} = 6.649$, $p < 0.001$, $\eta_p^2 = 0.153$] in the corresponding linear mixed effects model confirms this stark discrepancy. Follow-up analysis reveal significant differences between groups from trial 9 onward [$t_{(56.7)} \geq 2.461$, $p \leq 0.0169$, $d \geq 1.711$]. Additionally, there is a significant main effect of trial number [$F_{(1,407)} = 15.758$, $p < 0.001$, $\eta_p^2 = 0.299$], but no significant main effect of group [$F_{(1,37)} = 3.755$, $p = 0.060$, $\eta_p^2 = 0.092$].

Participants in either group showed a marked decrease in decision time over the curse of the study, especially after the very first trial (Figure 5B). A significant main effect of factor *trial number* [$F_{(11,407)} = 13.025$, $p < 0.001$, $\eta_p^2 = 0.260$] confirms this observation. Corresponding *post-hoc* analyses show significant differences between trial 1 and all other trials [all $t_{(407)} \geq 5.189$, $p < 0.001$, $d > 1.175$]. Moreover, decision time for trial 4 as the initial trial after the first in-game attention question, stands out. Users require significantly more time to reach a feeding decision in trial 4 compared to trial 5 [$t_{(407)} = 3.755$, $p = 0.013$, $d = 0.850$], trial 7 [$t_{(407)} = 4.020$, $p = 0.005$, $d = 0.911$], trial 10 [$t_{(407)} = 3.397$, $p < 0.049$, $d = 0.769$], and trial 11 [$t_{(407)} = 3.537$, $p < 0.030$, $d = 0.801$]. Neither the main effect of factor *group* [$F_{(1,37)} = 3.976$, $p = 0.054$, $\eta_p^2 = 0.097$], nor the interaction between factors *trial number* and *group* [$F_{(11,407)} = 0.965$, $p = 0.477$, $\eta_p^2 = 0.025$] reach significance.

Thus, these results verify our hypothesis that providing CFEs in the AlienZoo not just facilitates, but enables learning in the first place, given the poor performance of participants in the control group.

3.1.3. Assessing user's explicit knowledge

In terms of mean number of matches between user judgments of plant relevance for task success and the ground truth, participants receiving CFEs could explicitly identify relevant plants (*control*: mean number of matches between user input and ground truth = 1.895 ± 0.072 *SE*; *CFE*: mean number of matches = 3.000 ± 0.286 *SE*; $U = 281.5$, $p = 0.001$, $r = 0.517$) as well as irrelevant plants (*control*: mean number of matches between user input and ground truth = 2.421 ± 0.176 *SE*; *CFE*: mean number of matches = 3.210 ± 0.224 *SE*; $U = 264.5$, $p = 0.009$, $r = 0.422$) more easily than users receiving no explanation (see Figure 5C).

3.1.4. Measures of subjective usability

Hypothesis 2 posits that providing CFEs compared to no explanation increases user's subjective understanding. To assess this

TABLE 1 Demographic information of participants in Experiment 1.

	Before quality assurance measures ($N = 45$)				After quality assurance measures ($N = 39$)			
	Control	CFE	U-value[a]	p-value	Control	CFE	U-value[a]	p-value
N	22	23	..	..	19	20	..	..
Gender[b]	5 f/17 m	5 f/18 m	255.5	0.950	4 f/15 m	5 f/15 m	182.5	0.788
Age (*Mdn*)[c]	35–44 y	35–44 y	225.5	0.516	35–44 y	35–44y	143	0.168

[a]Non-parametric Wilcoxon-Mann-Whitney U-test.
[b]f, female; m, male.
[c]*Mdn*, median age band (options: 18–24 y, 25–34 y, 35–44 y, 45–54 y, 55–64 y, 65 y, and over).

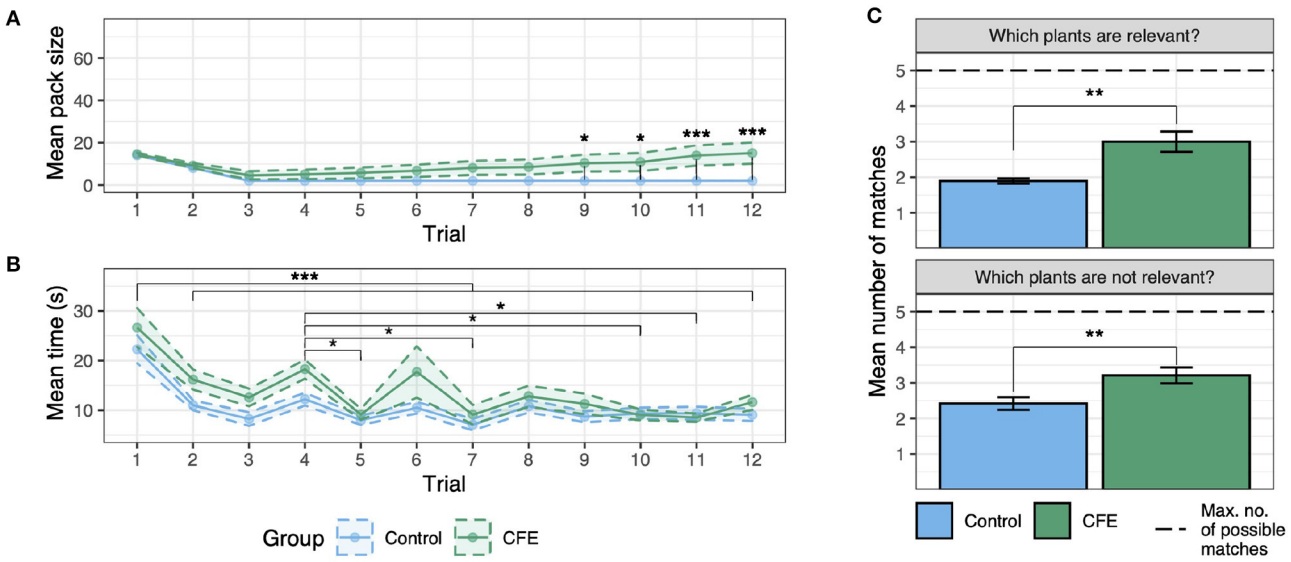

FIGURE 5
Experiment 1: Development of **(A)** mean pack size per group by trial, **(B)** mean decision time per group by trial, and **(C)** mean number of matches between user judgments and ground truth for survey items assessing relevant plants and irrelevant plants, respectively. Shaded areas in **(A, B)**, and error bars in **(C)** denote the standard error of the mean. Asterisks denote statistical significance with $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$, respectively.

notion, we analyze participant judgments on relevant items in the post-game survey.

Visual assessment of user responses suggest large discrepancies between groups in items assessing feedback's helpfulness and usability (Figure 6A). This notion is confirmed by the corresponding statistical assessment. Groups differ when judging whether presented feedback (i.e., summary of past choices only vs. summary + CFEs) was helpful to increase pack size (*control* condition: $M = 1.789 \pm 0.282$ *SE*; *CFE* condition: $M = 3.700 \pm 1.285$ *SE*; $U = 306.5$, $p < 0.001$, $r = 0.540$). Similarly, participants receiving CFEs on top of a summary of their past choices significantly differed in terms of reported subjective usability (*control* condition: $M = 1.210 \pm 0.096$ *SE*; *CFE* condition: $M = 3.450 \pm 0.294$ *SE*; $U = 351$, $p < 0.001$, $r = 0.759$). Strikingly, however, there is no significant difference between groups for estimated usefulness of feedback for others (*control* condition: $M = 3.632 \pm 0.244$ *SE*; *CFE* condition: $M = 3.350 \pm 0.335$ *SE*; $U = 175$, $p = 0.674$, $r = 0.067$).

3.1.5. Mode of presenting feedback and CFEs

In conflict with Hypothesis 3, survey responses reflecting user's subjective understanding of feedback show that groups differ in terms of understanding the feedback as such (Figure 6B). While a considerable proportion of both groups responds positively about understanding the feedback, the *control* group leans significantly more to giving positive judgements (*control* condition: $M = 4.105 \pm 0.252$ *SE*; *CFE* condition: $M = 3.7 \pm 0.309$ *SE*; $U = 312.5$, $p < 0.001$, $r = 0.567$). When indicating their need for support for understanding, both groups reply with a comparable, more balanced response pattern (*control* condition: $M = 2.684 \pm 0.351$ *SE*; *CFE* condition: $M = 2.900 \pm 0.383$ *SE*; $U = 205$, $p = 0.674$ $r = 0.067$). User judgements on timing and efficacy of presented feedback is consistently high across groups (*control* condition: $M = 4.316 \pm 0.188$ *SE*; *CFE* condition: $M = 3.950 \pm 0.246$ *SE*; $U = 154.5$, $p = 0.289$ $r = 0.170$).

3.1.6. Identification of inconsistencies

Our explanatory analysis revealed that users in groups did not differ in finding inconsistencies in the feedback provided (*control* condition: $M = 2.316 \pm 0.265$ *SE*; *CFE* condition: $M = 2.95 \pm 0.352$ *SE*; $U = 232$, $p = 0.232$, $r = 0.192$, see Figure 6C).

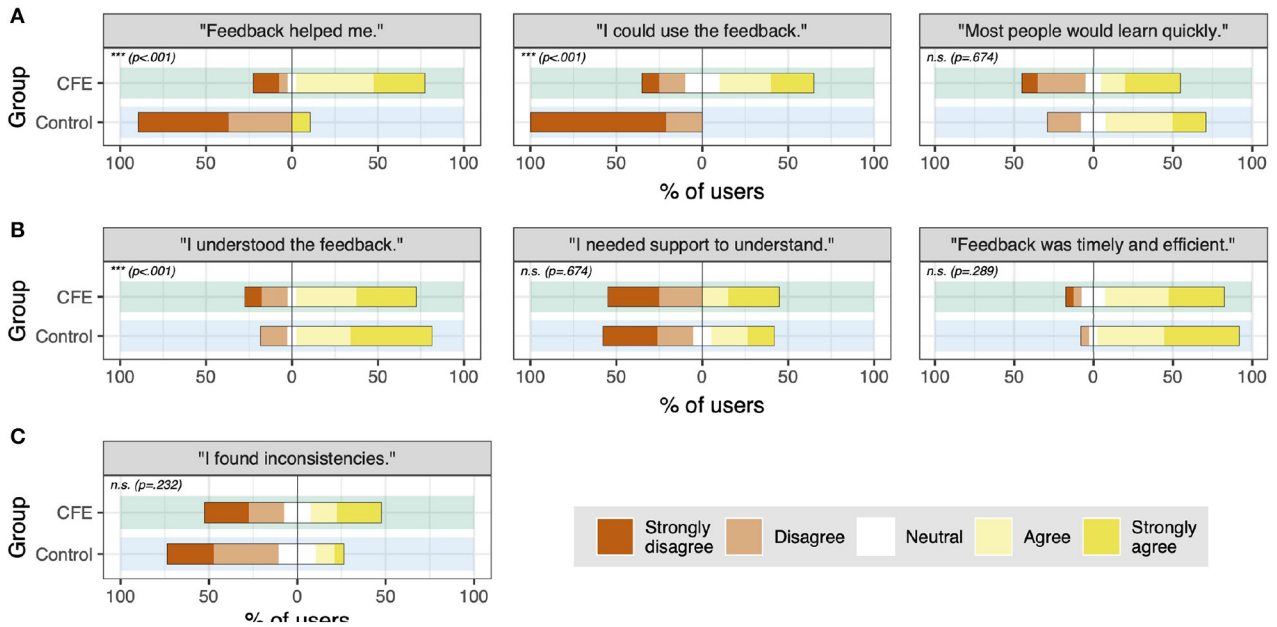

FIGURE 6
Experiment 1: Overview of user responses in post-game survey (adapted from Holzinger et al., 2020) per group. **(A)** Depicts user replies in survey items relevant for hypothesis 2, **(B)** depicts user replies in survey items relevant for hypothesis 3, and **(C)** depicts replies relevant for our last exploratory analysis. Statistical information including the respective *p*-value is given within each item's box (n.s., not significant).

3.2. Experiment 2

In Experiment 2, we acquired data from 45 additional participants facing the same task as in Experiment 1 (Table 2). The underlying data used for model training was simpler, including the interdependence of two and not three features.

3.2.1. Participant flow

From 45 participants recruited *via* AMT, we exclude data from participants who failed both attention trials during the game ($n = 1$), and straight-lined during the survey ($n = 1$). No participant in this cohort qualified as a speeder, gave an incorrect response for the catch item in the survey, or straight-lined in the game part of the study. Thus, the final analysis includes data from 43 participants (Table 2).

On average, the these 43 participants in Experiment 2 needed 14 m:25 s (± 01 m:07 s SEM) from accepting the task on AMTs to inserting their unique payment code.

3.2.2. Objective measures of usability

In Experiment 2, we successfully replicate the beneficial effect of providing CFEs compared to no explanations in the Alien Zoo approach already seen in Experiment 1. This is noteworthy, given the less complex interdependencies within the underlying data set.

As in Experiment 1, average pack size per group increases significantly faster when CFEs are given [Figure 7A; significant interaction of factors trial number and group; $F_{(11,\ 451)} = 32.748$, $p < 0.001$, $\eta_p^2 = 0.444$]. In contrast to Experiment 1, some users in the *control* condition increases their pack size over the course of the experiment, in line with our expectation given the simpler data set. Still, follow-up analyses reveal significant differences between groups from trial 5 onward [all $t_{(67.8)} \geq 2.384$, $p \leq 0.020$, $d \geq 1.467$]. Additionally, there is a significant main effect of trial number [$F_{(1,\ 451)} = 62.556$, $p < 0.001$, $\eta_p^2 = 0.604$], as well as a significant main effect of group [$F_{(1,\ 41)} = 16.909$, $p < 0.001$, $\eta_p^2 = 0.292$].

Similar to Experiment 1, participants in both groups showed a decrease in decision time over the curse of the study, evident after the very first trial (Figure 7B). A significant main effect of factor *trial number* [$F(11,\ 451) = 4.991$, $p < 0.001$, $\eta_p^2 = 0.109$] confirms this observation. Corresponding post-hoc analyses show significant differences between trial 1 and all other trials [all $t_{(451)} \geq 3.432$, $p \leq 0.043$, $d \geq 1.740$], except for trials 2 and 10. Neither the main effect of factor *group* [$F_{(1,\ 41)} = 2.758$, $p = 0.104$, $\eta_p^2 = 0.063$], nor the interaction between factors *trial number* and *group* [$F_{(11,\ 451)} = 1.439$, $p = 0.152$, $\eta_p^2 = 0.034$] reach significance.

Overall, these results support the initial findings from Experiment 1, emphasizing the beneficial role of providing CFEs in the Alien Zoo for successful task completion.

3.2.3. Assessing user's explicit knowledge

Unlike Experiment 1, there is no statistically meaningful difference between groups in terms of number of matches between user judgments of plant relevance for task success and the ground truth (*control*: mean number of matches between user input and ground truth = 3.000 ± 0.207 *SE*; *CFE*: mean number of matches = 3.182 ± 0.260 *SE*; $U = 255$, $p = 0.554$, $r = 0.090$) as well as irrelevant plants (*control*: mean number of matches between user input and ground truth = 2.857 ± 0.221 *SE*; *CFE*: mean number of matches = 2.819 ± 0.284 *SE*; $U = 223.5$, $p = 0.860$, $r = 0.221$),

TABLE 2 Demographic information of participants in Experiment 2.

	Before quality assurance measures ($N = 45$)				After quality assurance measures ($N = 43$)			
	Control	CFE	U-value[a]	p-value	Control	CFE	U-value[a]	p-value
N	21	24	..	..	21	22	..	..
Gender[b]	9 f/11 m/1 nb	11 f/13 m	238	0.725	9 f/11 m/1 nb	9 f/13 m	229	0.967
Age (*Mdn*)[c]	35–44 y	35–44 y	280	0.497	35–44 y	35–44 y	253	0.571

[a]Non-parametric Wilcoxon-Mann-Whitney U-test.
[b]f, female; m, male; nb, non-binary.
[c]*Mdn*, median age band (options: 18–24 y, 25–34 y, 35–44 y, 45–54 y, 55–64 y, 65 y, and over).

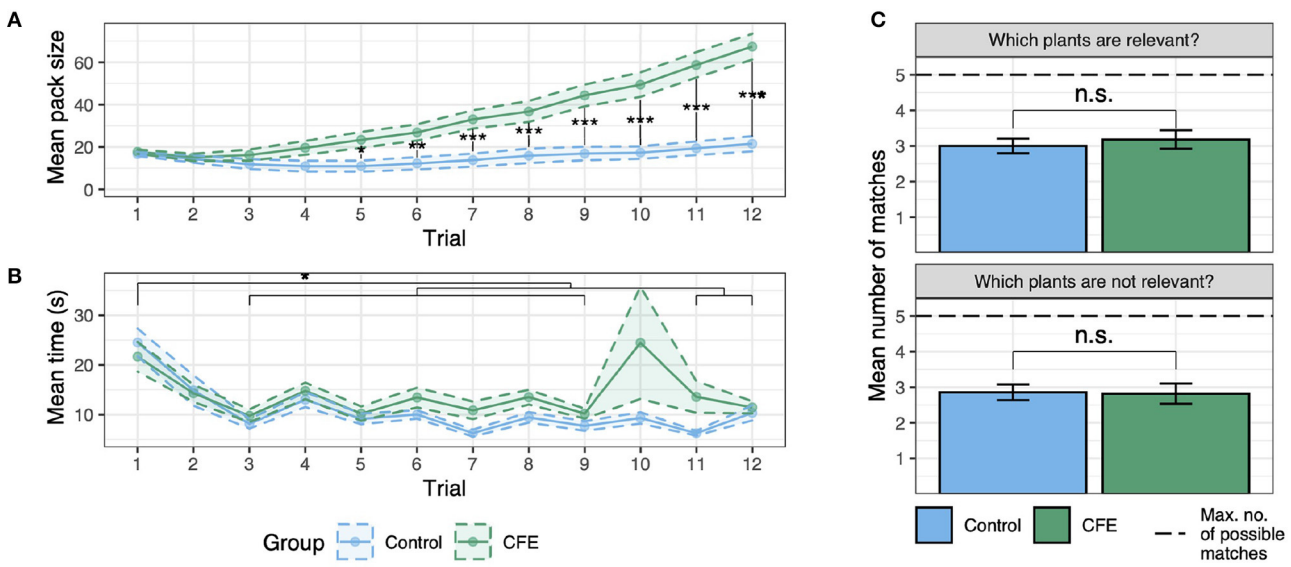

FIGURE 7
Experiment 2: Development of **(A)** mean pack size per group by trial, **(B)** mean decision time per group by trial, and **(C)** mean number of matches between user judgments and ground truth for survey items assessing relevant plants and irrelevant plants, respectively. Shaded areas in **(A)** and **(B)**, and error bars in **(C)** denote the standard error of the mean. Asterisks denote statistical significance with $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$, respectively, and n.s., not statistically significant (i.e., $p > 0.05$).

indicating greater success in building up explicit knowledge even without explanations, given the simpler data set (see Figure 7C).

Thus, given the current data, the advantage of building better explicit knowledge when CFEs are available seems to disappear.

3.2.4. Measures of subjective usability

In stark contrast to Experiment 1, the majority of users from both groups in Experiment 2 shared a positive feeling that provided feedback was helpful and usable (Figure 8A). The difference in response patterns still differs significantly between groups, in terms of subjective helpfulness (*control* condition: $M = 3.714 \pm 0.277$ *SE*; *CFE* condition: $M = 4.682 \pm 0.153$ *SE*; $U = 351$, $p = 0.001$, $r = 0.489$) and subjective usability (*control* condition: $M = 4.048 \pm 0.189$ *SE*; *CFE* condition: $M = 4.591 \pm 0.157$ *SE*; $U = 325$, $p = 0.012$, $r = 0.385$). Extremely favorable user judgements from the *CFE* group likely drive this effect, due to strong agreement by a large proportion of users from this cohort.

As in Experiment 1, there is no significant difference between groups for estimated usefulness of feedback for others (*control* condition: $M = 3.952 \pm 0.212$ *SE*; *CFE* condition: $M = 4.409 \pm 0.157$ *SE*; $U = 294.5$, $p = 0.091$, $r = 0.258$).

3.2.5. Mode of presenting feedback and CFEs

In accordance with Hypothesis 3, survey responses reflecting user's subjective understanding of feedback show that groups did not differ in terms of understanding the feedback as such (Figure 8B; *control* condition: $M = 4.571 \pm 0.111$ *SE*; *CFE* condition: $M = 4.409 \pm 0.157$ *SE*; $U = 306$, $p = 0.05$, $r = 0.610$). Likewise, users in both groups indicate strongly that they do not wish for support to understand feedback provided (*control* condition: $M = 1.905 \pm 0.248$ *SE*; *CFE* condition: $M = 2.318 \pm 0.250$ *SE*; $U = 284$, $p \leq 0.177$ $r = 0.206$). User judgments on timing and efficacy of presented feedback is consistently high across groups (*control* condition: $M = 4.334 \pm 0.174$ *SE*; *CFE* condition: $M = 4.591 \pm 0.107$ *SE*; $U = 266.5$, $p = 0.334$ $r = 0.147$).

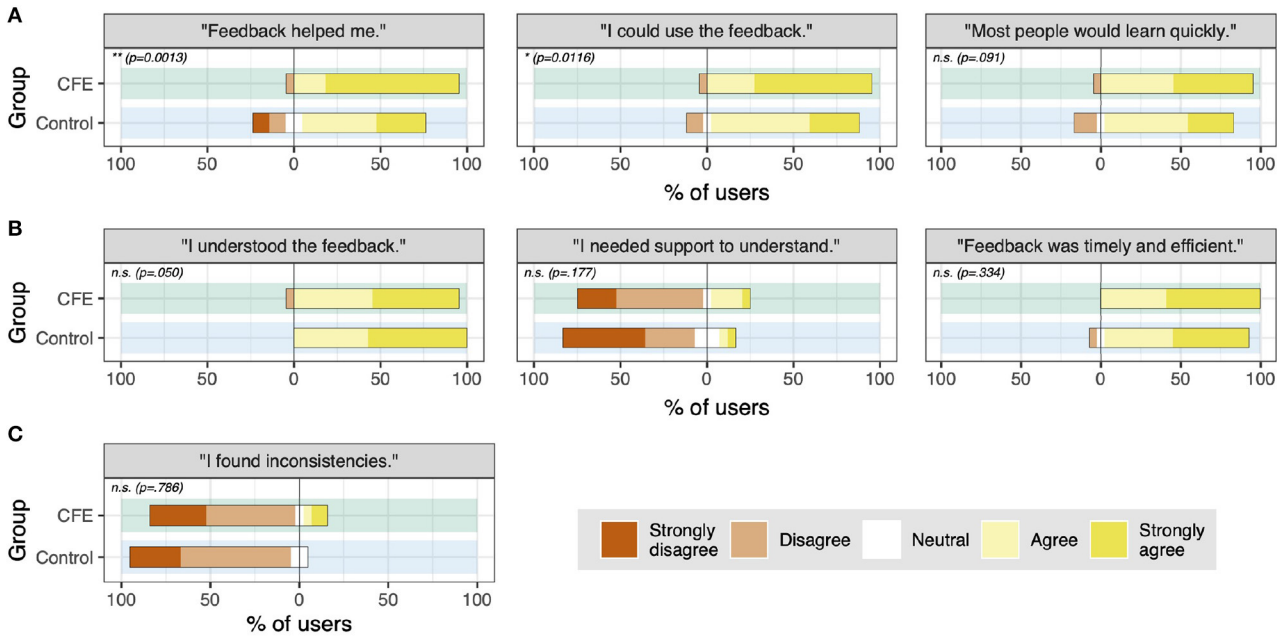

FIGURE 8
Experiment 2: Overview of user responses in post-game survey (adapted from Holzinger et al., 2020) per group. **(A)** Depicts user replies in survey items relevant for hypothesis 2, **(B)** depicts user replies in survey items relevant for hypothesis 3, and **(C)** depicts replies relevant for our last exploratory analysis. Statistical information including the respective p-value is given within each item's box (n.s., not significant).

3.2.6. Identification of inconsistencies

Analysis of the final survey item reveals that users in both groups did not differ in finding inconsistencies in the feedback provided (*control* condition: $M = 1.810 \pm 0.131$ *SE*; *CFE* condition: $M = 2.091 \pm 0.254$ *SE*; $U = 241.5$, $p = 0.786$, $r = 0.041$, see Figure 8C).

4. Discussion

In the empirical proof of concept study, we investigate the efficacy and feasibility of the Alien Zoo framework. To this end, we examine the impact of providing CFEs on user performance as compared to no explanations. Based on objective behavioral variables and subjective self-reports, we assess understanding and usability of CFE-style feedback. Our results reveal the potential of the Alien Zoo framework to study the usability of CFEs approaches.

Most notably, merely providing a summary of past choices does not necessarily enable users to gain insight into the system. This becomes especially clear considering the poor task performance of *control* participants in the more complex Experiment 1. Given the comparatively complex interdependence of three features in the underlying data, none of the *control* participants manage to increase their pack size in the course of the experiment.

Participants receiving CFEs for their choices, however, are able to manipulate the system more efficiently. While both experiments vary in terms of the complexity of the underlying data used for model building, the observation of CFEs participants outperforming their peers in the *control* group, is consistent. Interestingly, the *control* group in Experiment 2 did indeed manage to improve their pack size to some extent, but providing explanations puts users at a definite advantage. In fact, 100% of all users in the experimental condition in Experiment 2 correctly determine that plant 2 is a relevant feature. This observation not only supports the claim that CFEs are a very intuitive and meaningful way of explaining in XAI (Wachter et al., 2017), but clearly demonstrates their effectiveness in the current setting.

Intriguingly, our results diverge from those of empirical XAI studies that find no beneficial effect of providing CFEs on user's task performance (Lim et al., 2009; van der Waa et al., 2021). For instance, Lim et al. (2009) review various explanation approaches in the domain of context-aware systems. Their evidence suggests that users receiving counterfactual style *what-if*-explanations have no advantage over control users when manipulating abstract features (labeled *A*, *B*, and *C*) to explore their influence on abstract predictions (labeled *a* or *b*).

Decisive differences between both experimental tasks may explain this stark contrast. First, the Alien Zoo revolves around an engaging setting (i.e., feeding aliens to make the pack grow), as opposed to the non-specific nature of the system in Lim et al. (2009). Second, we offer users different rounds of action and feedback in alternating learning and testing steps, making the Alien Zoo truly interactive. In contrast, users in Lim et al. (2009) undergo an initial evaluation section displaying explanation after explanation, followed by a separate test phase. Learners obtaining deeper understanding through hands-on activities rather than passive studying is well established in educational science (Chi and Wylie, 2014), potentially explaining discrepancies in terms of

observed user behavior. Thus, we suggest that future XAI usability studies should put a strong focus on goal-directed and interactive tasks to be maximally effective.

All users across conditions and experiments significantly decrease their decision time, apparent already after the very first trial. This effect most likely reflects how participants initially familiarize themselves with the interface. Another slight increase is observable for trial 4, right after the first in-game attention question appeared. We assume that users took this trial to re-focus after this unsuspected disruption. From there on, decision times consistently level out for both groups. Thus, despite the performance benefit, we have no evidence that providing CFEs leads to faster, and this more automatic, decision-making. Reaction time measures are a staple of experimental psychology, typically linked to simple, perceptual experiments. The current study, in contrast, employs a complex, a game-type scenario invoking high-level decision-making that may not be sensitive enough or too short to reveal subtle group differences. Further, the experimental groups may have been slowed down by the increased cognitive load imposed by CFE, an effect previously demonstrated to lead to higher—not lower—response times (Lage et al., 2019). At the same time, having no additional help to go by, some no-explanation controls may have approached the task using fast and simple heuristics, relying on intuition more that time-consuming elaboration (Kahneman, 2011). Consequently, decision time decreases for both groups may be driven by different underlying factors (increasing automaticity vs. intuitive decision-making).

In the more complex Experiment 1, users in the experimental group identify plants relevant for the task more reliably compared to users in the control group. Interestingly, in the simpler Experiment 2, this significant difference vanishes. This might reflect the greater success of control users to see through the system in this simpler setting, even without explanations. However, this should not be taken as evidence that users across groups indeed build up comparable mental models of the underlying system, given the considerable difference in task performance. In fact, one caveat of the current analysis may be insufficient sensitivity of the measure of matches between user input and ground truth, possibly diluting noteworthy effects. For instance, 100% of all users in the experimental condition, but only 57% of all control participants, could determine that plant 2 is a relevant feature in Experiment 2 (see Supplementary Figures 2, 3). The current measure does not capture this detail, calling for careful interpretation of the corresponding null-effect.

On top of the objective measures quantifying system understanding, we assess various subjective measures to tap into perceived usability. Across both experiments, the experimental groups judged their CFE-style feedback as being more helpful and usable compared to the control group (Figure 6A, respectively). Thus, providing CFEs does not just improve user's performance, but also their subjective usability of the system.

Surprisingly, despite variable responses in terms of helpfulness and usability of presented feedback for oneself, the estimated usefulness for others is not different across groups. In fact, a larger proportion of control users in Experiment 1 reported favorably on this item, even though they found feedback of little help and limited usefulness. This astonishing result is difficult to interpret without access to more detailed qualitative data from those participants. Maybe these users are demotivated by their poor turnout, feeling that they perform exceptionally bad compared to the average person.

Participant responses to items in place to assess potential confounding factors reveal an interesting pattern that merits closer inspection (Figure 8B). In Experiment 1, a considerable proportion of both groups responds positively about understanding the feedback. However, the *control* group leans significantly more toward agreement. This might reflect higher cognitive load the CFE group, as they receive a more crowded, information-heavy screen. While in line with findings suggesting that counterfactual style questions impose a larger cognitive load on participants (Lage et al., 2019), this interpretation is unlikely as this effect vanishes in the simpler Experiment 2. This fact rather suggests that the increased task difficulty drives this effect. Response patterns on the other two control items are more consistent. Users across groups and across experiments state that they need little support to understand the feedback provided. Similarly, an overwhelming majority of all users across all groups indicate that feedback was timely and efficient, backing the efficacy of the Alien Zoo framework despite its relatively complex game-like setup.

Survey items depicted in Figure 8B is set in place to assess potential confounding factors, possibly impacting the efficiency of the Alien Zoo framework. We assumed that across experiments, consistent group difference with respect to these items would inform us about potential design flaws. While one group difference emerges, however, it is not consistent across experiments. This clearly indicates that the respective item (*"I understood the feedback."*) not just evaluates general understanding, but also reflects the underlying task difficulty. A possible explanation for this may be that there is still room for improvement for a clean identification of confounds. In lack of a standard inventory for assessing subjective usability in XAI user studies, we rely on an adapted version of the System Causability Scale (Holzinger et al., 2020). Prominent alternatives for test instruments include are the DARPA project's Explanation Satisfaction and Trust scales (Hoffman et al., 2018), shown to effectively assess differences in terms of satisfaction and trust (Warren et al., 2022). Further scales exist, often highly specific to particular applications (Cahour and Forzy, 2009; Heerink et al., 2010) Consequently, there is essential similarity between some items and individual specificity across scales. Perfecting subjective measures of XAI satisfaction may the a worthwhile future endeavor for future research.

Finally, our exploratory analysis reveals that groups in both experiments do not differ in finding inconsistencies in the feedback provided. This acts as a further quality measure for the CFE approach, trusted to generate feasible and sound explanations. While this verdict is virtually unanimous across users in the simpler Experiment 2, some users in both groups in Experiment 1 indicate that they did indeed determine inconsistencies. While a minority, this observation merits a comment. We cannot exclude that some users in the CFE group indeed receive feedback in different runs that, when taken together, does not perfectly align. It is important to keep in mind that CFEs are local explanations, highlighting what would lead to better results in a particular instance. Variability, especially in terms of the irrelevant features, may indeed exist. To uncover whether such effects cause fundamental problems, we intentionally moved away from the

perfect, hand-crafted explanations assessed classical *Wizard of Oz* designs more prominently used in the community (Narayanan et al., 2018; Lage et al., 2019; Sokol and Flach, 2020b; van der Waa et al., 2021), and used predictions from real ML-models. However, the observation that a small proportion of users in the control group indicate that they found inconsistencies, is much more puzzling. These users merely see a correct summary of their past choices as feedback, and thus inconsistencies are impossible. Given that this survey item was the very last, it may be a sign of participants' loss of attention or fatigue. Identifying the actual underlying reasons requires collecting quantitative data, e.g., *via* in depth user interviews. These measurements require moving away from the accessible web-based format, and perform complementary evaluations in an in-person, lab-based setting.

4.1. Limitations of the empirical study

Several limitations of the empirical proof of concept study deserve detailed discussion.

A cautionary note regards the general efficacy of CFEs for human users. CFEs are local explanations, focusing on how to undo one past prediction. Thus, it is very unlikely that users are able to form an accurate mental model of the entire underlying system solely based on a sparse set of these specific explanations. This is a short-coming, given that completeness is an important prerequisite for this process (Kulesza et al., 2013). It remains an avenue for future research to show situations that severely impact usability of CFEs, as they are unable to provide a complete picture.

Another point to keep in mind is the potential problem of users falling victim to confirmation bias after receiving the first round of CFE feedback (Wang et al., 2019). In essence, we cannot rule out that some users generate a faulty initial hypothesis, and subsequently look for confirming evidence for that faulty initial hypothesis only. This may have greater impact on the control group, given that they have very little evidence to go by choosing the best plant combination. Still, it also needs to be acknowledged as a possible issue for the CFE participants. Consequently, such a strategy would hamper learning profoundly, and we cannot rule out that some lower performing users indeed follow it. While exploring the impact of confirmation bias for CFEs in XAI is outside the scope of this work, the issue deserves more careful attention in future work.

Further, we do not investigate whether providing CFEs did also improve user's trust in the system. Trust is an important factor in XAI, and prominently studied in various designs (Lim et al., 2009; Ribeiro et al., 2016; Davis et al., 2020). The current work, however, exclusively focuses on the aspect of usability. Extending the current set up to include evaluation of trust can be easily realized, for instance by extending the survey by corresponding items.

Finally, a further insight gained from this study is the critical impact of task difficulty on user performance and judgements. While not directly at the center of the current work, we shed a first light on these effects by observing differences between Experiment 1 and 2. Future research should look into the effects of data complexity on usability of CFEs.

4.2. Conclusions of the empirical study

The main contributions of the empirical proof of concept study are two-fold. First, we provide long-awaited empirical evidence for the claim that CFEs are indeed more beneficial for users than providing no explanations, at least in abstract setting, when tasked to gain new knowledge. Importantly, this advantage becomes apparent both in terms of objective performance measures and subjective user judgements. Second, we demonstrate the basic efficacy of the Alien Zoo framework for studying the usability of CFEs in XAI. Thus, within the limits of this game-type scenario in its low-knowledge domain, it promises to be an effective way for testing specific aspects of CFEs in XAI.

4.3. Future perspectives and general conclusions

The current paper introduces the Alien Zoo framework, developed to assess the usability of CFEs in XAI in a game-type scenario set in a low-knowledge domain. In a proof of concept study, we demonstrate its efficacy by examining the added benefit of providing CFEs over no explanations using an iterative learning task in the abstract Alien Zoo setting.

We believe that the Alien Zoo enables researchers to investigate a wide variety of different questions related to XAI strategies and specific aspects of CFEs for XAI. For instance, in a separate study, we use the Alien Zoo to investigate potential advantages of CFEs restricted to plausible regions of the data space compared to classical CFEs remaining as close to the original input as possible (Kuhl et al., 2022). Surprisingly, this investigation reveals that novice users in the current task do not benefit from an additional plausibility constraint.

Another issue for future research may be to examine usability of different types of CFEs. Importantly, CFEs may vary in terms of framing the respective result. Upward counterfactuals highlight how the current situation would be improved, while downward counterfactuals emphasize changes leading to a less desirable outcome (Epstude and Roese, 2008). The impact of such a framing in XAI is yet to be shown.

Moreover, further research may uncover potential differences in usability for CFEs generated for different models. While the way CFEs are presented in the Alien Zoo is always the same, the underlying models may be fundamentally different. Thus, if human users pick up on model differences solely based on their respective explanations, it may have critical implications for their usability. A particularly intriguing question to be addressed is whether users are able to identify a model that is objectively worse.

As a final suggestion of this by no means exhaustive list, we propose studying potentially negative effects of CFEs: In the field of XAIs, it is universally assumed that CFEs are intuitive and human-friendly. Thus, it will be extremely informative to investigate and identify cases where these types of explanation do more harm than good, e.g., when users come to trust ML models even if they are biased and unfair.

It is natural for people to interact with each other by explaining their behaviors to one another. The key to building a stable mental model for prediction and control of the world is to explain in a way that is understandable and usable (Heider, 1958). However, in the absence of a universally applicable definition of what constitutes a good explanation, a lack of user-based evaluations affects the assessment of automatically generated CFEs for ML. The lack of user-based research does not only bear upon assessments of CFEs as such, but also limits the overall evaluation of different conceptualizations for this kind of explanations. Consequently, with the Alien Zoo framework, we offer a flexible, easily adaptable design, applicable for various purposes and research questions. This approach in its implementation may be freely used by researchers and practitioners to further advance the field of XAI.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://github.com/ukuhl/IntroAlienZoo.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee, Bielefeld University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

UK, AA, and BH contributed to conception of the study. UK designed the study and the experimental framework, implemented the web interface, supervised data acquisition on AMT, performed the statistical analysis, and wrote the first draft of the manuscript. AA implemented the back end. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This research was supported by 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, by the European Research Council (ERC) under the ERC Synergy Grant Water-Futures (Grant agreement No. 951424), and project IMPACT funded in the frame of the funding line *AI and its Implications for Future Society* by the VW-Foundation. We acknowledge the financial support of the German Research Foundation (DFG) and the Open Access Publication Fund of Bielefeld University for the article processing charge.

Acknowledgments

The authors would like to thank Johannes Kummert for support with server set up and maintenance.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1087929/full#supplementary-material

References

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

Akula, A., Wang, S., and Zhu, S.-C. (2020). "CoCoX: Generating conceptual and counterfactual explanations *via* fault-lines," in *Proceedings of the AAAI Conference on Artificial Intelligence* (New York, NY), 34, 2594–2601. doi: 10.1609/aaai.v34i03.5643

Arras, L., Osman, A., and Samek, W. (2022). CLEVR-XAI: a benchmark dataset for the ground truth evaluation of neural network explanations. *Inform. Fus.* 81, 14–40. doi: 10.1016/j.inffus.2021.11.008

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fus.* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Artelt, A. (2019). *CEML: Counterfactuals for Explaining Machine Learning Models - A Python Toolbox*. Available online at: https://www.github.com/andreArtelt/ceml

Artelt, A., and Hammer, B. (2019). On the computation of counterfactual explanations - A survey. *CoRR, abs/1911.07749*.

Artelt, A., and Hammer, B. (2020). "Convex density constraints for computing plausible counterfactual explanations," in *Artificial Neural Networks and Machine Learning–ICANN 2020, Vol. 12396*, eds I. Farkaš, P. Masulli, and S. Wermter (Cham: Springer International Publishing), 353–365. doi: 10.1007/978-3-030-61609-0_28

Artelt, A., Vaquet, V., Velioglu, R., Hinder, F., Brinkrolf, J., Schilling, M., et al. (2021). "Evaluating robustness of counterfactual explanations," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (Orlando, FL), 1–9. doi: 10.1109/SSCI50451.2021.9660058

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., and Horvitz, E. (2019). Updates in human-AI teams: understanding and addressing the performance/compatibility tradeoff. *Proc. AAAI Conf. Artif. Intell.* 33, 2429–2437. doi: 10.1609/aaai.v33i01.33012429

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Ben-Shachar, M., Lüdecke, D., and Makowski, D. (2020). Effectsize: estimation of effect size indices and standardized parameters. *J. Open Source Softw.* 5:2815. doi: 10.21105/joss.02815

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification And Regression Trees, 1st Edn*. London: Routledge.

Browne, J. T. (2019). "Wizard of Oz prototyping for machine learning experiences," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1–6. doi: 10.1145/3290607.3312877

Byrne, R. M. (2016). Counterfactual thought. *Annu. Rev. Psychol.* 67, 135–157. doi: 10.1146/annurev-psych-122414-033249

Byrne, R. M. J. (2019). "Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19* (Macao), 6276–6282. doi: 10.24963/ijcai.2019/876

Cahour, B., and Forzy, J.-F. (2009). Does projection into use improve trust and exploration? An example with a cruise control system. *Saf. Sci.* 47, 1260–1270. doi: 10.1016/j.ssci.2009.03.015

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chi, M. T. H., and Wylie, R. (2014). The ICAP framework: linking cognitive engagement to active learning outcomes. *Educ. Psychol.* 49, 219–243. doi: 10.1080/00461520.2014.965823

Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., and Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. *Inform. Fus.* 81, 59–83. doi: 10.1016/j.inffus.2021.11.003

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of oz studies–why and how. *Knowl. Based Syst.* 6, 258–266. doi: 10.1016/0950-7051(93)90017-N

Dai, X., Keane, M. T., Shalloo, L., Ruelle, E., and Byrne, R. M .J. (2022). "Counterfactual explanations for prediction and diagnosis in XAI," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY: Association for Computing Machinery), 215–226. doi: 10.1145/3514094.3534144

Dandl, S., Molnar, C., Binder, M., and 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* (Macao: Springer), 448–469. doi: 10.1007/978-3-030-58112-1_31

Davis, B., Glenski, M., Sealy, W., and Arendt, D. (2020). "Measure utility, gain trust: practical advice for XAI researchers," in *2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)* (Salt Lake City, UT), 1–8. doi: 10.1109/TREX51495.2020.00005

Detry, M. A., and Ma, Y. (2016). Analyzing repeated measurements using mixed models. *JAMA* 315:407. doi: 10.1001/jama.2015.19394

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*. doi: 10.48550/arXiv.1702.08608

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I., Muller, M., et al. (2021). The who in explainable AI: how AI background shapes perceptions of AI explanations. *arXiv preprint arXiv:2107.13509*. doi: 10.48550/arXiv.2107.13509

Epstude, K., and Roese, N. J. (2008). The functional theory of counterfactual thinking. *Pers. Soc. Psychol. Rev.* 12, 168–192. doi: 10.1177/1088868308316091

European Union (2016). Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation). *Off. J. Eur. Union L110* 59, 1–88.

Goldinger, S. D., Kleider, H. M., Azuma, T., and Beike, D. R. (2003). "Blaming the victim" under memory load. *Psychol. Sci.* 14, 81–85. doi: 10.1111/1467-9280.01423

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., and Giannotti, F. (2018). Local rule-based explanations of black box decision systems. *arXiv:1805.10820 [cs]*. doi: 10.48550/arXiv.1805.10820

Heerink, M., Kröse, B., Evers, V., and Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: the almere model. *Int. J. Soc. Robot.* 4, 361–375. doi: 10.1007/s12369-010-0068-5

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York, NY: John Wiley & Sons Ltd. doi: 10.1037/10628-000

Hilton, D. J., and Slugoski, B. R. (1986). Knowledge-based causal attribution: the abnormal conditions focus model. *Psychol. Rev.* 93, 75–88. doi: 10.1037/0033-295X.93.1.75

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

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

Jentzsch, S. F., Hohn, S., and Hochgeschwender, N. (2019). "Conversational interfaces for explainable AI: a human-centred approach," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: First International Workshop, EXTRAAMAS 2019* (Montreal, QC: Springer), 77–92.

Kahneman, D. (2011). *Thinking Fast and Slow, 1st Edn*. New York, NY: Farrar Straus and Giroux.

Keane, M. T., Kenny, E. M., Delaney, E., and Smyth, B. (2021). If only we had better counterfactual explanations: five key deficits to rectify in the evaluation of counterfactual XAI techniques. *arXiv:2103.01035 [cs]*. doi: 10.24963/ijcai.2021/609

Kuhl, U., Artelt, A., and 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 *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (Seoul), 2125–2137. doi: 10.1145/3531146.3534630

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., and Wong, W.-K. (2013). "Too much, too little, or just right? Ways explanations impact end users' mental models," in *2013 IEEE Symposium on Visual Languages and Human Centric Computing* (San Jose, CA: IEEE), 3–10. doi: 10.1109/VLHCC.2013.6645235

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., et al. (2019). "Human evaluation of models built for interpretability," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 7* (Washington, DC), 59–67. doi: 10.1609/hcomp.v7i1.5280

Le, T., Wang, S., and Lee, D. (2020). "Grace: generating concise and informative contrastive sample to explain neural network model's prediction," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, NY: Association for Computing Machinery), 238–248. doi: 10.1145/3394486.3403066

Lim, B. Y., Dey, A. K., and Avrahami, D. (2009). "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA: ACM), 2119–2128. doi: 10.1145/1518701.1519023

Lipton, P. (1990). Contrastive explanation. *R. Instit. Philos. Suppl.* 27, 247–266. doi: 10.1017/S1358246100005130

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: a test of the instance theory of automaticity. *J. Exp. Psychol.* 18, 883–914. doi: 10.1037/0278-7393.18.5.883

Lombrozo, T. (2012). "Explanation and abductive inference," in *The Oxford Handbook of Thinking and Reasoning*, eds K. J. Holyoak and R. G. Morrison (Oxford, UK: Oxford University Press), 260–276. doi: 10.1093/oxfordhb/9780199734689.013.0014

Markman, K. D., and McMullen, M. N. (2003). A reflection and evaluation model of comparative thinking. *Pers. Soc. Psychol. Rev.* 7, 244–267. doi: 10.1207/S15327957PSPR0703_04

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

Mohseni, S., Zarei, N., and Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Trans. Interact. Intell. Syst.* 11, 1–45. doi: 10.1145/3387166

Muth, C., Bales, K. L., Hinde, K., Maninger, N., Mendoza, S. P., and Ferrer, E. (2016). Alternative models for small samples in psychological research: applying linear mixed effects models and generalized estimating equations to repeated measures data. *Educ. Psychol. Measure.* 76, 64–87. doi: 10.1177/0013164415580432

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., and Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*. doi: 10.48550/arXiv.1802.00682

Offert, F. (2017). "I know it when I see it". Visualization and Intuitive Interpretability. *arXiv:1711.08042 [stat]*. doi: 10.48550/arXiv.1711.08042

Pawelczyk, M., Bielawski, S., Heuvel, J. v. d., Richter, T., and Kasneci, G. (2021). Carla: a python library to benchmark algorithmic recourse and counterfactual explanation algorithms. *arXiv preprint arXiv:2108.00783*. doi: 10.48550/arXiv.2108.00783

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://jmlr.org/papers/v12/pedregosa11a.html

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52. doi: 10.1145/3411764.3445315

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ""Why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: Association for Computing Machinery), 1135–1144. doi: 10.1145/2939672.2939778

Roese, N. J. (1997). Counterfactual thinking. *Psychol. Bull.* 121, 133–148. doi: 10.1037/0033-2909.121.1.133

Roese, N. J., and Epstude, K. (2017). "The functional theory of counterfactual thinking: new evidence, new challenges, new insights," in *Advances in Experimental Social Psychology, Vol. 56* (Amsterdam: Elsevier), 1–79. doi: 10.1016/bs.aesp.2017.02.001

Sanna, L. J., and Turley, K. J. (1996). Antecedents to spontaneous counterfactual thinking: effects of expectancy violation and outcome valence. *Pers. Soc. Psychol. Bull.* 22, 906–919. doi: 10.1177/0146167296229005

Sattarzadeh, S., Sudhakar, M., and Plataniotis, K. N. (2021). "SVEA: a small-scale benchmark for validating the usability of *post-hoc* explainable AI solutions in image and signal recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC), 4158–4167. doi: 10.1109/ICCVW54120.2021.00462

Shalev-Shwartz, S., and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9781107298019

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

Sokol, K., and Flach, P. (2020b). One explanation does not fit all. *Künstl. Intell.* 34, 235–250. doi: 10.1007/s13218-020-00637-y

Stepin, I., Catala, A., Pereira-Fari na, M., and Alonso, J. M. (2019). "Paving the way towards counterfactual generation in argumentative conversational agents," in *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)* (Tokyo: Association for Computational Linguistics), 20–25. doi: 10.18653/v1/W19-8405

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

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv. JL Tech.* 31:841. doi: 10.2139/ssrn.3063289

Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). "Designing theory-driven user-centric explainable AI," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1–15. doi: 10.1145/3290605.3300831

Warren, G., Keane, M. T., and Byrne, R. M. J. (2022). *Features of Explainability: How Users Understand Counterfactual and Causal Explanations for Categorical and Continuous Features in XAI*. Vienna.

White, A., and d'Avila Garcez, A. (2020). "Measurable counterfactual local explanations for any classifier," in *ECAI* (Santiago de Compostela), 7.