Let's Go to the Alien Zoo: Introducing an Experimental Framework to Study Usability of Counterfactual Explanations for Machine Learning

Preprint

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

[1]Bielefeld University, Germany

Abstract

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. Thus, 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. Our results suggest 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. 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`

1 Introduction

In a step towards accountable and transparent ML, the European Union mandates safeguards against automated decision-making with the General Data Protection Regulation (GDPR) 2016/679 [25]. Specifically, Recital 71 of the GDPR states that a person subjected to automated decision-making may obtain an explanation of the given decision. While it is debated whether this establishes a legally binding 'right to explanation' [64], it has certainly sparked lively discussion about scope, realization, and feasibility of explanations for ML in the domain of XAI. Consequently, there has been an upswing of technical XAI approaches on how to make ML explainable in recent years [4, 16].

Alongside novel explainability approaches, authors have proposed evaluation criteria and guidelines to systematically characterize and assess XAI approaches [4, 19, 22, 60]. For instance, Sokol and Flach suggest a taxonomy defining criteria an explanatory method has to satisfy to be considered usable, summarized in an *"Explainability Fact Sheet"*.

This theoretical groundwork sparked generation of several practical validation frameworks, focusing on function level validation of explanation approaches [3, 7, 49, 50, 58]. For instance, these frameworks evaluate explanations in terms of their accuracy and fidelity [3, 50, 58, 67], or robustness [7].

Importantly, however, these considerations pass over the role of the user as eventual target - a curious limitation, given that user studies are considered the gold standard in XAI [22, 60]. While XAI taxonomies repeatedly emphasize the need for human-level validation of explanation approaches, only few authors concern themselves with user-based evaluations, often with limitations concerning statistical power and reproducibility [33].

The repeated emergence of counter-intuitive findings from sparse user evaluations acts as a stark reminder why accounting for the human factor is vital when evaluating XAI approaches. For instance, Poursabzi-Sangdeh et al. show that participants can more 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. Moreover, instead of the expected advantage of clear models in terms of their interpretability, users are actually less able to detect when the model had made a mistake [52]. Similarly, employing the Alien Zoo framework introduced in this manuscript, we demonstrate that introducing a theoretically motivated plausibility constraint on generated explanations may be less useful for users in certain settings [34].

What inhibits systematic and controlled comparisons of XAI approaches from a usability perspective? While user evaluations are essential to evaluate the efficacy of explanation modes, designing an effective user study is no easy feat. A well-designed study needs to closely consider the respective explainees, and the reason for explaining [1, 60], while simultaneously taking into account confounding factors and available resources [22]. Further, it is challenging to ensure comparability of conditions

human participants face, while systematically varying XAI approaches, underlying ML models, or data distributions.

Thus, lack of openly accessible and engaging user study designs that enable direct comparisons between different explainability 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 [5], targeting novice users in a domain-general context.

The aim of this contribution is to equip research groups and practitioners to with an easily adaptable design, adjustable for various purposes and research questions. With this paradigm, we account for a series of challenges that are sometimes overlooked in previous XAI user studies, overcoming prominent shortcomings in the literature (see Section 3.1). 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.

The remainder of this paper is as follows: We will first provide a primer on CFEs for ML, and briefly review previous usability assessments and respective lessons learned (Section 2). Subsequently, we will detail our the conceptualization of the proposed Alien Zoo framework, including guiding design principles, constructs and measurements, and implementation specifics (Section 3). Section 4 describes the accompanying proof of concept usability study, demonstrating the efficacy and feasibility of the Alien Zoo approach to evaluate CFEs. We close this paper with an in-depth discussion of insights drawn from this study, including limitations and avenues for future work in Section 4.4.

2 CFEs FOR ML

The advent of novel XAI approaches also triggered a shift toward a user-centered focus on explainability [44]. In the wake of this change in focus, CFEs received special attention as a supposedly useful, human-friendly solution [33, 44].

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").

What makes CFEs so attractive as an XAI approach? Foremost, their contrastive nature, emphasizing why a specific outcome occurred instead of another, strongly resembles human cognitive reasoning [29, 40, 42, 44]. Humans naturally reflect on past events by generating possible alternatives, thus routinely engaging in counterfactual thinking [55]. Specifically, when humans generate counterfactuals, they take the relevant facts of events as input, and mentally change their mental representation of these facts in order to produce a counterfactual scenario, while maintaining the factual representation in parallel [13].

Empirical evidence shows that humans engage in counterfactual thinking automatically [26]. In a series of experiments, Sanna and Turley demonstrate that participants produce counterfactual thoughts spontaneously in various settings (e.g., when re-telling stories, evaluating their own performance on an exam, or describing their performance in a laboratory anagram task) [57]. Importantly, the number of spontaneously uttered counterfactuals increases when participants face negative outcomes or results were unexpected [57].

Based on these and similar insights, Epstude and Roese emphasize the beneficial role of this mode of thinking to regulate one's behavior in order to improve future performance [24]. In the updated version of their *Functional Theory of Counterfactural Thinking*, they posit that disparities between the current state and an ideal goal state triggers spontaneous counterfactual thought [56]. In the same vein, Markman and McMullen argue that comparative thought modes like generating counterfactuals may help to prepare for the future by guiding the formation of intentions and thus changing prospective behavior [43]. Taken together, the contrastive and spontaneously way humans generate counterfactuals that presumably guide future behavior leads many authors in XAI to treat explanations formulated as counterfactuals as intuitively useful, and readily human-usable [6, 18, 27, 62].

However, these encouraging insights from psychology seem to have created the erroneous impression that technical CFEs approaches providing explainability for ML models may take the quality of the suggested explanation modes at face value [22, 48]. 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 [33].

2.1 *Insights from Previous CFE Usability Studies*

While still a minority, studies that do examine CFE approaches from a user-perspective give cause for cautious confidence in their usability.

Waa et al. 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 [63] Additionally, their data suggest that CFEs have a positive effect on perceived system comprehensibility compared to no explanation. Focusing on the issue of fairness, Dodge et al. find that counterfactual explanations most prominently expose biased classifiers, compared to alternative approaches like showing feature relevance or merely describing the distribution of underlying data [21]. Finally, a comparison of different explanation approaches reveals that participants judge counterfactual style explanations to be subjectively more intuitive and useful than, e.g., visualizing feature importance scores [37].

This positive evidence is not unanimous, however. Users tasked to learn how an automatic system behaves indeed show some

[1] Available at `https://github.com/ukuhl/IntroAlienZoo`

understanding of the types of rules governing said system after receiving counterfactual-style explanations, as compared to receiving no-explanation [38]. 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. Lage et al. 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 [36].

On top of these inconsistent results, previous user evaluations often suffer from a series of limitations. While it is intuitively clear that one explanation mode fitting all scenarios is unlikely to exist [61], some authors neglect to clearly formulate the given purpose for explaining and target group in their study [37]. Without an explicit classification of the experimental context, research runs the risk to reach all-too-general conclusions like declaring one mode of explaining universally superior to another.

Many user evaluations suffer from methodological limitations like low participant numbers [2, 38]. For reasons of simplicity, some approaches provide participants with explanations that follow a certain XAI approach, but were actually designed by the researchers themselves [36, 47, 63].

Such a *Wizard of Oz* approach, with a human behind the scenes plays the role of an automatic system [17], allows perfect control over materials encountered by participants. However, it fails to account for potential variability in the results of ML algorithms. For instance, it is perfectly conceivable that an approach may produce unhelpful explanations in certain settings. If this may happen to users 'in the wild', we posit that assessment of these approaches in the lab needs to account for such contingencies as well.

Another prominent limitation affects evaluations that merely 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. However, it is unclear whether such subjective evaluations translate into tangible behavioral effects [30]. In fact, a recent study fails to show a correlation between perceived system understandability and any objective measure of performance [63].

Further, experimental designs in XAI are often limited in terms of engaging participants, especially in studies assessing the efficacy of explanations for understanding how a system works. For instance, participants typically study pre-selected examples of a system's input and output values, together with a corresponding explanation [37, 38, 63]. However, greater focus on user action may be advantageous. Evidence from educational science suggests that learner's level of commitment relates to final learning outcomes, with interactive activities granting deeper understanding [15].

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.

Thus, while more and more empirical evaluations assessing the usability of CFEs for ML take the scene, experimental shortcomings become apparent in a number of ways. Moreover, the diverse range of different set-ups and paradigms, testing different CFE methods in different scenarios and with diverse groups of users, impeding firm conclusions in terms of strengths and weaknesses of methods, as studies plainly lack comparability.

3 THE ALIEN ZOO FRAMEWORK

3.1 Guiding Design Principles

As more and more user studies in the domain of XAI emerge, more and more recommendations and guidelines concerning design principles that need to be taken into account enter the picture [19, 45, 63].

For instance, based on their experiences setting up an XAI usability study, Waa et al. formulate recommendations as a reference for future XAI research [63]. We closely followed these guidelines when constructing the Alien Zoo framework.

3.1.1 Use case and experimental context

The effectiveness of an explanation depends decisively on the reason reason for explaining, and the intended target audience [1, 4, 45]. Both aspects determine the choice of an appropriate use case, and thus the experimental context.

On the one hand, a user's background knowledge crucially impacts their judgement whether a piece of information is relevant: Users who already possess a lot of applicable domain knowledge may find more sophisticated explanations more useful than users that are novices [22]. Even more critically, prior domain knowledge and user beliefs may impact how and even if users meaningfully engage with provided explanations [38]. Moreover, explainees equipped with AI expertise perceive and evaluate provided explanations differently than users that lack this kind of knowledge [23].

On the other hand, the explanation's purpose profoundly affects requirements a given XAI approach ought to meet. Users tasked to compare different models likely have other explanation needs than those who want to gain new knowledge from a predictive model or the data used to build it [1].

Consequently, generalizability of conclusions beyond a given use case, context, and target group is limited and needs to be treated with upmost caution [22, 60].

Prominent XAI taxonomies take these aspects into consideration. For instance, the previously mentioned *"Explainability Fact Sheet"* includes characteristics of context (i.e., *Explanation audience* and *Function of the Explanation*) as part of the operational dimension of the XAI evaluation [60]. Doshi-Velez and Kim [22] provide a structured classification of evaluation approaches for interpretability, differentiating between application-grounded, human-grounded, and functionally-grounded approaches.

In the Alien Zoo framework, we focus on an abstract experimental context: Participants imagine themselves as zookeepers in a zoo for aliens, so-called shubs (Figure 1a). Participants' main task is to find how to best feed the shubs under their care. They may choose from different plants to feed the aliens (Figure 1b), but it is not clear what plants (or which plant combination)

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

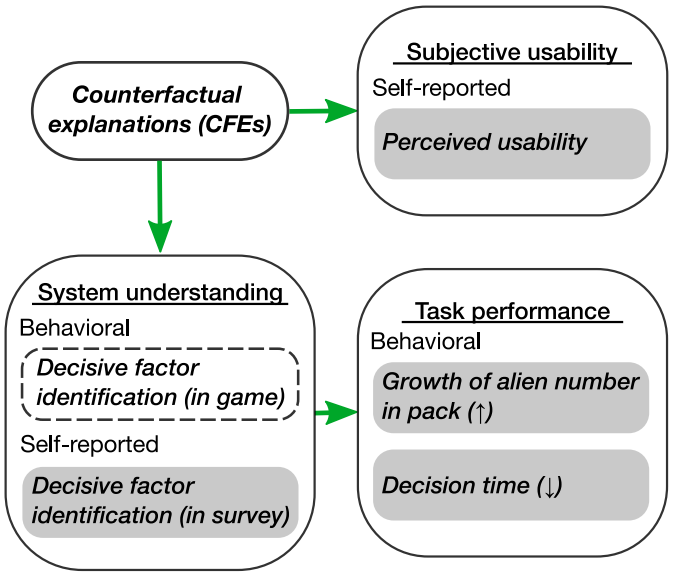


Figure 2: Causal diagram describing expected effects of counterfactual explanations on the constructs subjective usability, system understanding, and task performance investigated in the Alien Zoo framework. Green arrows depict expected positive effects. Opaque gray boxes show the measures for the respective 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.

makes up a nutritious diet, causing their pack to thrive. Feeding decisions have immediate consequences, leading to reductions or increases of the pack size. 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.

With the Alien Zoo framework, we provide a highly interactive, game-like scenario that triggers participant engagement over iterative rounds of user action and feedback. Evidence from educational science motivates this choice, demonstrating that learner's level of commitment affects learning outcome, and that interactive activities foster deeper understanding [15].

In the following we adhere to the advice by Waa et al. and first provide a structured account of the choice for a use case following well-defined taxonomies. Subsequently, we discuss the effectiveness of the use case domain for the intended purpose of the evaluation, and review our choice for running the proof of concept investigation as a web-based study [63].

With the given use case, we assess performance of real users in an abstract task setting. Thus, Alien Zoo user studies correspond to human grounded evaluations, with participants engaged in a variant of "counterfactual simulation" [22]. As described by Adadi and Berrada, 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 [1].

A particular advantage concerns the abstract nature of our task, ruling out any 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.

Last, we diverge from the final recommendation by Waa et al. regarding the experimental context: instead of using a controlled, lab-based setting, our framework is a web-based design. While we understand concerns regarding validity of data acquired this way [63], we demonstrate the feasibility of using the given online approach to obtain meaningful data if appropriate quality measures are in place (see Section 4.2.3).

3.1.2 Constructs and their relations

To underlay any XAI user evaluation with a solid basis for scientific theory, Waa et al. advocate clear definitions of utilized constructs and their interrelations.

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 4 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 [26, 55, 57], 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 [63].

3.1.3 Measurements

The Alien Zoo framework enables assessment of constructs subjective usability, system understanding, and task performance through different behavioral and self-report measures (Figure 2). Note that we rely on both objective behavioral variables and subjective self-reports. Given evidence of a disparity between perceived system understandability and objective measures of performance [63], we believe it crucial to address both aspects in order to provide a holistic assessment of usability of CFE for ML.

First, we hypothesize that providing CFEs to show participants alternative feeding choices leading to better outcomes triggers better understanding of the system. Specifically, 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. The first two survey items assess whether users can identify those plants that contribute to successful completion of a task, and those that do not matter to make the pack grow. Thus, we determine to what extent people have 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. Further, 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 [41].

Last, 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 Material A). For instance, users indicate whether they understood the explanations, in how far they find them useful, to which degree they can make use of them, and in how far they imagine the presented CFEs to be helpful for other users, too. These items assess user's subjective usability.

3.2 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. The web interface employs the JavaScript-based Phaser 3, an HTML5 game framework.[2] The back end of the system is Python3-based, with the sklearn package [51] 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 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.

The web-based nature of the infrastructure allows for prompt data collection of a large number of participants. In an associated study investigating the effects of plausibility constraints on generated CFEs, we acquired data from over 100 participants within four days via Amazon Mechanical Turk (AMT) [34]. Data acquisition from 90 participants in the current proof of concept study (Section 4) took five days, including the initial quality assessment.

Unlike previous designs that provide users with hand-crafted explanation examples [so-called *Wizard of Oz* designs, see 36, 47, 61, 63], the Alien Zoo equips participants with feedback from real XAI methods based on reproducible ML models. We agree that *Wizard of Oz* designs are the preferable option in terms of control over what participants experience and consistency of presented explanations [12, 32]. However, we believe that only a test of explanations that genuinely come from an ML model can show whether such explanations are useful, re-creating how users would interact with them 'in the wild'.

In the interest of reproducibility, we fully share data and code of the Alien Zoo framework, encouraging research groups and practitioners alike to adapt and utilize the implementation according to their own research needs.

4 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 4.2.5). 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] `https://phaser.io/`

[3] `https://github.com/andreArtelt/ceml`

4.1 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 [63].

Consequently, we address the following three hypotheses.

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.

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.

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.

4.2 Methods

4.2.1 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. 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 4.2.3) 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 [8]. The Ethics Committee of Bielefeld University, Germany, approved this study.

4.2.2 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 20s in an effort to prevent participants from skipping the instructions.

Game Phase Figure 3 visualizes the general flow of scenes displayed during the game phase.

The game phase begins with a padlock scene, where participants make their first feeding selection (left side in Figure 3, and Supplementary Material B). 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 10s 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.

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 [31].

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.

4.2.3 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 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 Material A). 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.

4.2.4 Statistical Analysis

We perform all statistical analyses using R-4.1.1 [53], 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 (lme4 v.4_1.1-27.1) [9]. 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 partici-

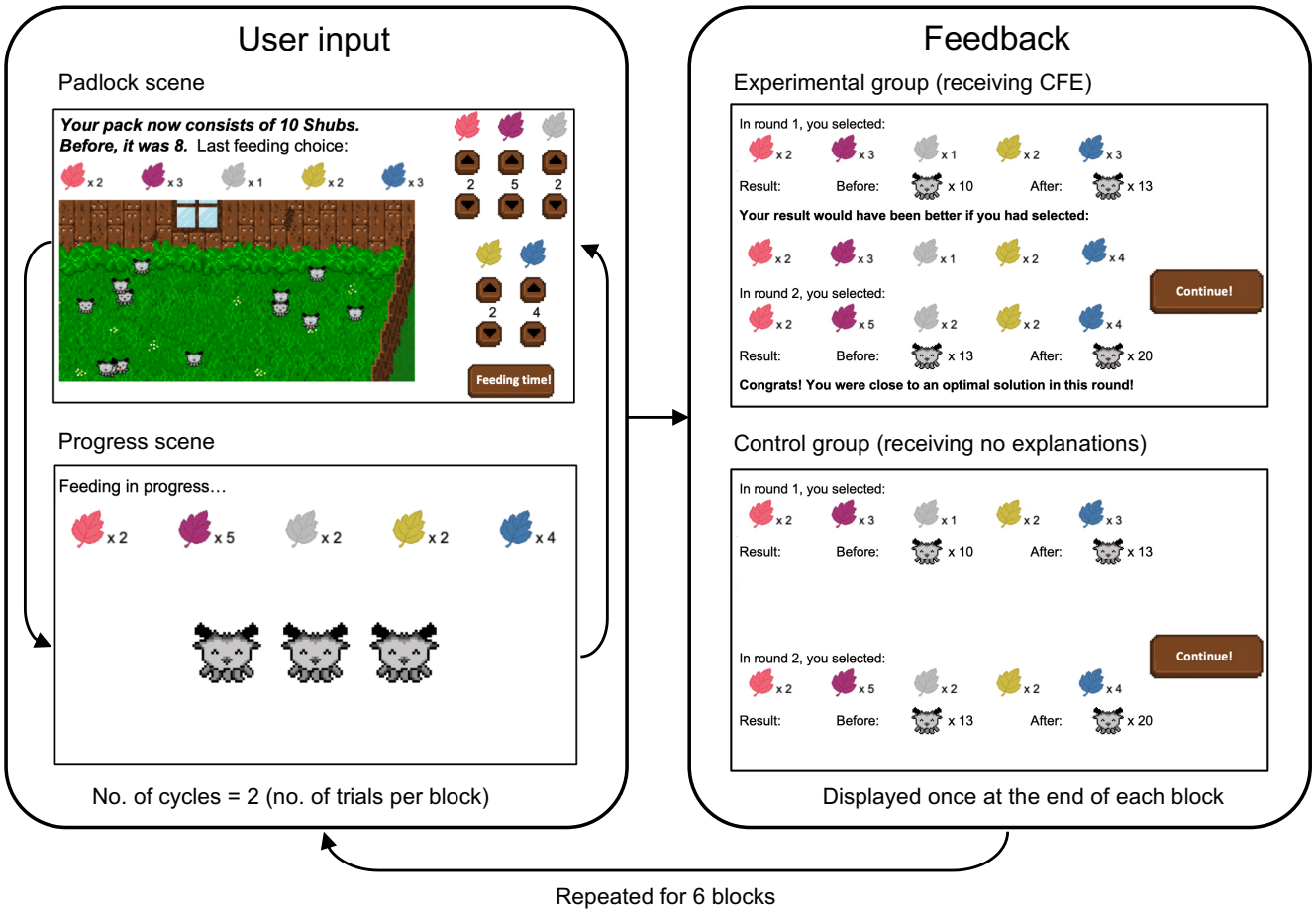


Figure 3: General flow of scenes displayed during the game phase. Note that this pattern was disrupted after trials 3 and 7 for an additional attention scene, asking participants to indicate the current number of shubs in their pack. A more detailed overview of scenes within a block can be found in Supplementary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

pant [20, 46]. 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) [10]. 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 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*.

4.2.5 *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 [59]. Decision trees are powerful enough to model our synthetic data set with sufficient accuracy, while allowing for efficient computation of CFE [5].[4] The current implementation uses the Gini splitting rule of CART [11]. To maintain comparable model outputs for all users within throughout one experiment, we use the same decision tree model once build in the beginning.

Hyperparameter tuning To ensure the models reliably present the respective underlying data structure without over-

[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.

fitting, 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).

Training Data The underlying data in Experiment 1 were generated according to the following scheme: The growth rate scales linearly with values 1 to 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 to 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 1680700 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) [14] using the bins as class labels. The final data set is obtained by removing the binning.

4.3 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.

4.3.1 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.

[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.

Table 1: Demographic information of participants in Experiment 1.

	Before quality assurance measures (N = 45)			
	control	*CFE*	U value[a]	p value
N	22	23	..	..
Gender[b]	5f/17m	5f/18m	255.5	.950
Age (*Mdn*)[c]	35–44y	35–44y	225.5	.516
	After quality assurance measures (N = 39)			
	control	*CFE*	U value[a]	p value
N	19	20	..	..
Gender[b]	4f/15m	5f/15m	182.5	.788
Age (*Mdn*)[c]	35–44y	35–44y	143	.168

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

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 17m:42s (± 01m:16s SEM) from accepting the task on AMTs to inserting their unique payment code.

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 <.001, η_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 <.001, η_p^2 = 0.299), but no significant main effect of group ($F(1,37)$ = 3.755, p = .060, η_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*

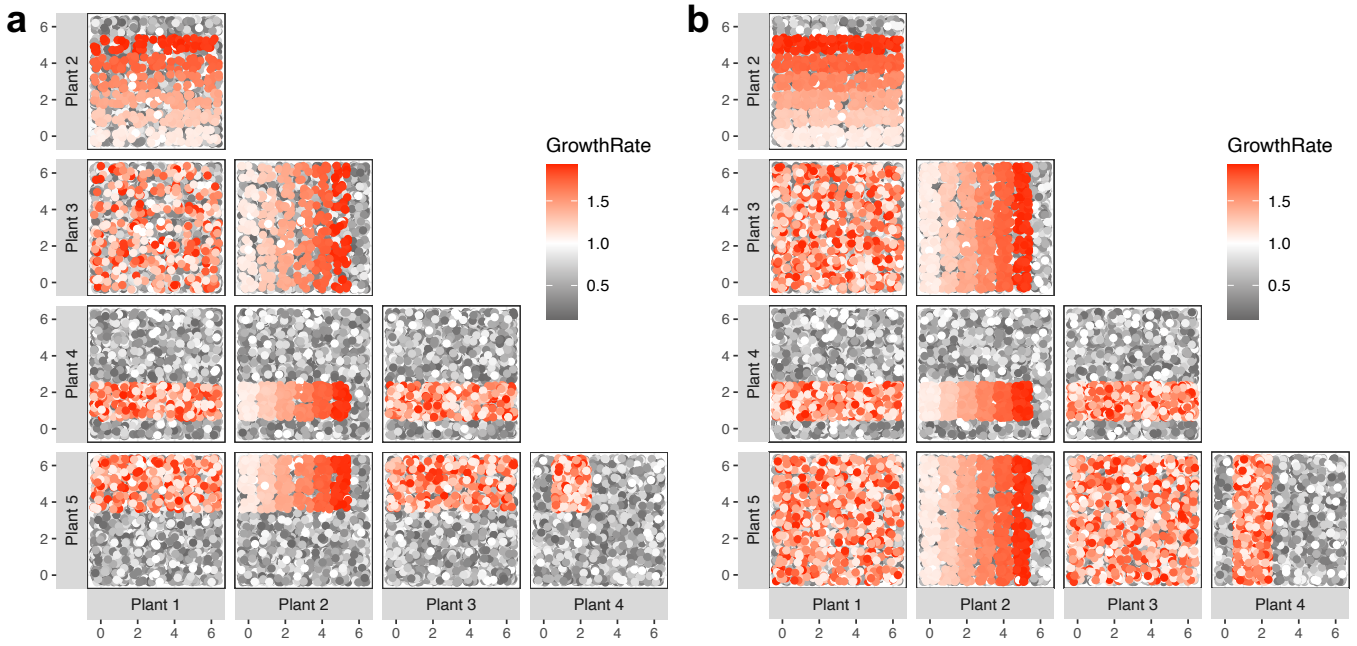


Figure 4: Distribution of synthetic data used for model training. Each point in each scatter plot represents the combination of two plant values, colored according to the corresponding growth rate of that point. Gray values indicate growth rate values below 1 (leading to pack size decreases), and red 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 are jittered around their true integer values.

number ($F(11,407) = 13.025$, $p < .001$, $\eta_p^2 = 0.260$) confirms this observation. Corresponding post-hoc analyses show significant differences between trial 1 and all other trials (all $t(407) \geq 5.189$, $p < .001$, $d > 1.175$). 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 ($t(407) = 3.755$, $p = .013$, $d = 0.850$), trial 7 ($t(407) = 4.020$, $p = .005$, $d = 0.911$), trial 10 ($t(407) = 3.397$, $p < .049$, $d = 0.769$), and trial 11 ($t(407) = 3.537$, $p < .030$, $d = 0.801$). Neither the main effect of factor *group* ($F(1,37) = 3.976$, $p = .054$, $\eta_p^2 = 0.097$), nor the interaction between factors *trial number* and *group* ($F(11,407) = 0.965$, $p = .477$, $\eta_p^2 = 0.025$) 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.

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*; $U = 281.5$, $p = .001$, $r = .517$) 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*; $U = 264.5$, $p = .009$, $r = .422$) more easily than users receiving no explanation.

Measures of Subjective Usability Hypothesis 2 posits that providing CFEs compared to no explanation increases user's subjective understanding. To assess this 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*; $U = 306.5$, $p < .001$, $r = .540$). 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*; $U = 351$, $p < .001$, $r = .759$). 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*; $U = 175$, $p = .674$, $r = .067$).

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

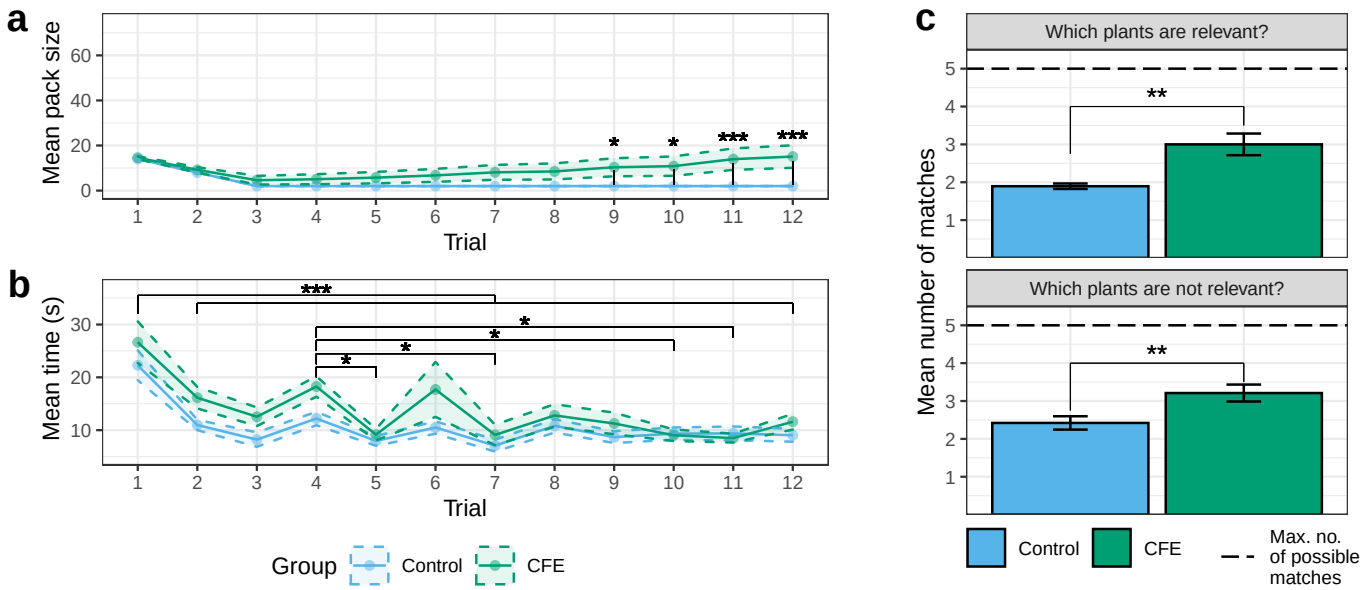


Figure 5: Experiment 1: Development of (**a**) mean pack size per group by trial, (**b**) mean decision time per group by trial, and (**c**) mean number of matches between user judgments and ground truth for survey items assessing relevant plants and irrelevant plants, respectively. Shaded areas in (**a**) and (**b**), and error bars in (**c**) denote the standard error of the mean. Asterisks denote statistical significance ($p < .05$ (*), and $p < .001$ (***), respectively.

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 ± 0.252 *SE*; *CFE* condition: M = 3.7 ± 0.309 *SE*; U = 312.5, p <.001, r = .567). When indicating their need for support for understanding, both groups reply with a comparable, more balanced response pattern (*control* condition: M = 2.684 ± 0.351 *SE*; *CFE* condition: M = 2.900 ± 0.383 *SE*; U = 205, p = .674 r = .067). User judgements on timing and efficacy of presented feedback is consistently high across groups (*control* condition: M = 4.316± 0.188 *SE*; *CFE* condition: M = 3.950±0.246 *SE*; U = 154.5, p = .289 r = .170).

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 ± 0.265 *SE*; *CFE* condition: M = 2.95 ± 0.352 *SE*; U = 232, p = .232, r = .192).

4.3.2 *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.

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).

Table 2: Demographic information of participants in Experiment 2.

	Before quality assurance measures (N = 45)			
	control	*CFE*	U value[a]	p value
N	21	24	..	..
Gender[b]	9f/11m/1nb	11f/13m	238	.725
Age (*Mdn*)[c]	35–44y	35–44y	280	.497
	After quality assurance measures (N = 43)			
	control	*CFE*	U value[a]	p value
N	21	22	..	..
Gender[b]	9f/11m/1nb	9f/13m	229	.967
Age (*Mdn*)[c]	35–44y	35–44y	253	.571

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

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

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 5a; significant interaction of factors trial number and group; F(11,451) = 32.748, p <.001, η^2_p = 0.444). In contrast to Experiment 1, some users in

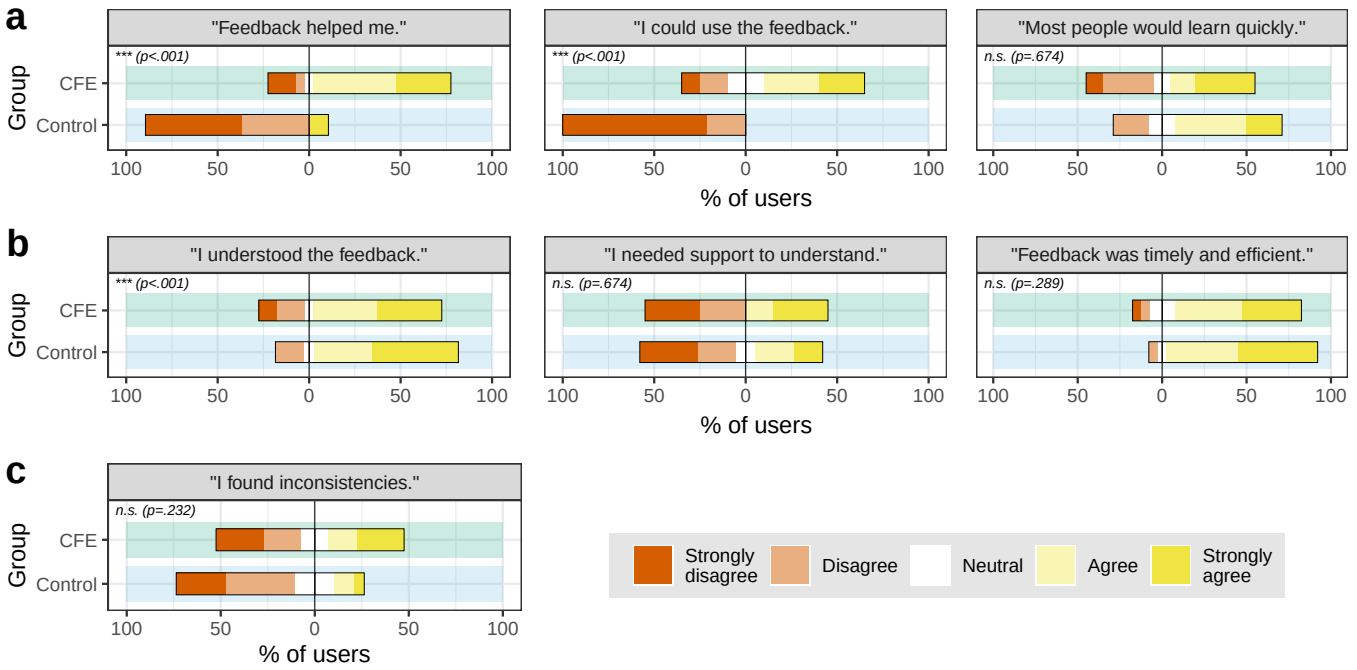


Figure 6: Experiment 1: Overview of user responses in post-game survey (adapted from [31]) per group. (**a**) depicts user replies in survey items relevant for hypothesis 2, (**b**) depicts user replies in survey items relevant for hypothesis 3, and (**c**) depicts replies relevant for our last exploratory analysis. Statistical information including the respective p-value is given within each item's box (n.s. = not significant).

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 ($F(1,451) = 62.556$, $p < .001$, $\eta_p^2 = 0.604$), as well as a significant main effect of group ($F(1,41) = 16.909$, $p < .001$, $\eta_p^2 = 0.292$).

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* ($F(11,451) = 4.991$, $p < .001$, $\eta_p^2 = 0.109$) confirms this observation. Corresponding post-hoc analyses show significant differences between trial 1 and all other trials (all $t(451) \geq 3.432$, $p \leq .043$, $d \geq 1.740$), except for trials 2 and 10. Neither the main effect of factor *group* ($F(1,41) = 2.758$, $p = .104$, $\eta_p^2 = 0.063$), nor the interaction between factors *trial number* and *group* ($F(11,451) = 1.439$, $p = .152$, $\eta_p^2 = 0.034$) 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.

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*; $U = 255$, $p = .554$, $r = .090$) 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*; $U = 223.5$, $p = .860$, $r = .221$), indicating greater success in building up explicit knowledge even without explanations, given the simpler data set.

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

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 (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*; $U = 351$, $p = .001$, $r = .489$) and subjective usability (*control* condition: $M = 4.048 \pm 0.189$ *SE*; *CFE* condition: $M = 4.591 \pm 0.157$ *SE*; $U = 325$, $p = .012$, $r = .385$). 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*; $U = 294.5$, $p = .091$, $r = .258$).

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: $M = 4.409 \pm 0.157$

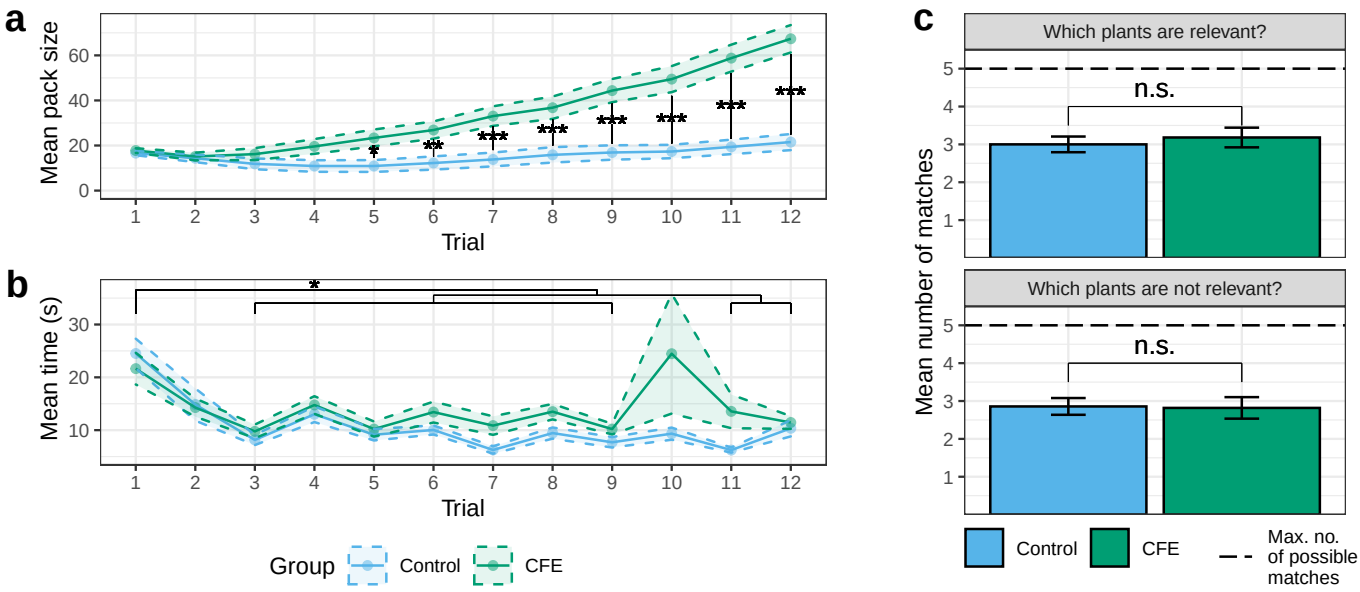


Figure 7: Experiment 2: Development of (**a**) mean pack size per group by trial, (**b**) mean decision time per group by trial, and (**c**) mean number of matches between user judgments and ground truth for survey items assessing relevant plants and irrelevant plants, respectively. Shaded areas in (**a**) and (**b**), and error bars in (**c**) denote the standard error of the mean. Asterisks denote statistical significance ($p < .05$ (*), $p < .01$ (**), and $p < .001$ (***), respectively. n.s. = not statistically significant (i.e., $p > .05$).

SE; $U = 306$, $p = .05$, $r = .610$). 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*; $U = 284$, $p \leq .177$ $r = .206$). 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*; $U = 266.5$, $p = .334$ $r = .147$).

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*; $U = 241.5$, $p = .786$, $r = .041$).

4.4 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 [65], 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 [39, 63]. For instance, Lim et al. review various explanation approaches in the domain of context-aware systems [39]. Their evidence suggests that users receiving counterfactual style *what-if*-explanations have no advantage over control users when manipulating abstract features (labelled *A*, *B* and *C*) to explore their influence on abstract predictions (labelled *a* or *b*).

In an attempt to explain this stark contrast to our results, we may turn to the details of both experimental tasks. 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. 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. 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

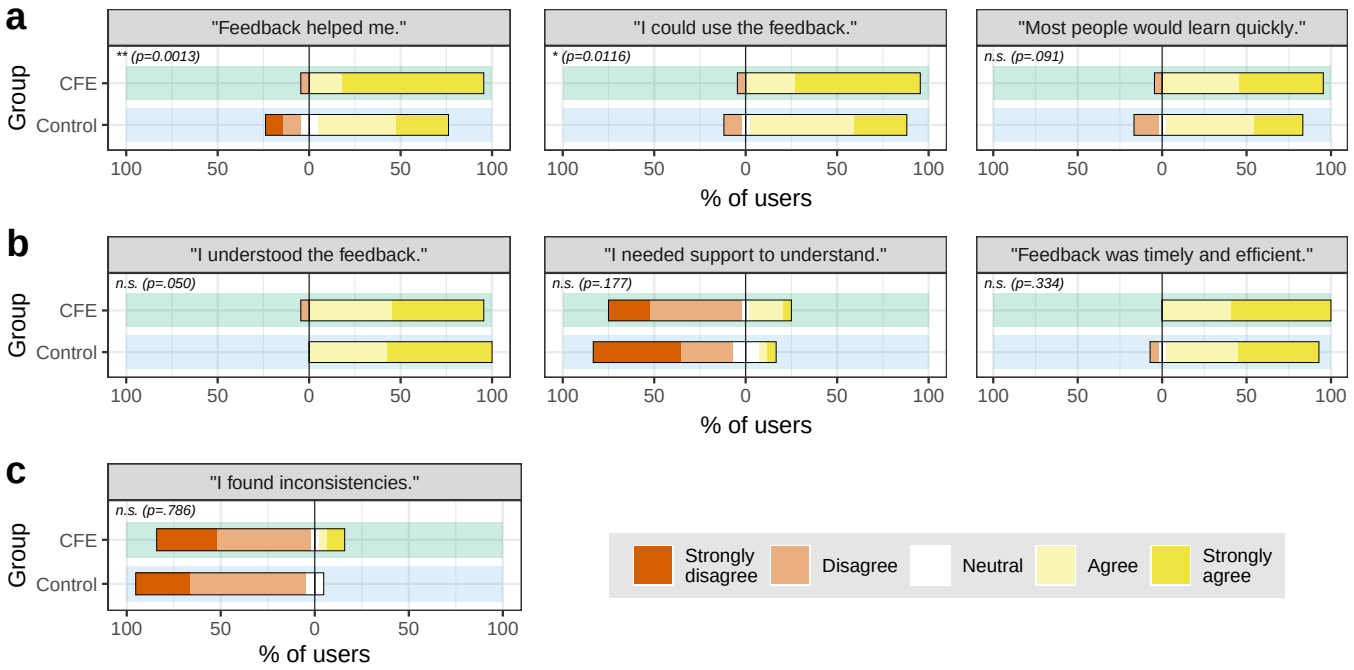


Figure 8: Experiment 2: Overview of user responses in post-game survey (adapted from [31]) per group. (**a**) depicts user replies in survey items relevant for hypothesis 2, (**b**) depicts user replies in survey items relevant for hypothesis 3, and (**c**) depicts replies relevant for our last exploratory analysis. Statistical information including the respective *p*-value is given within each item's box (n.s. = not significant).

educational science [15], 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 the time needed to reach a decision. This decrease becomes apparent already after the very first trial, most likely reflecting how participants initially take some moments to 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 focus their concentration again after this surprising 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 more automatic, and thus faster, decision-making.

In the more complex Experiment 1, users in the experimental group can more correctly state which plants are relevant for the task, 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 build up mental models of the underlying system that are indeed comparable, considering 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 Material C). 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 (Figures 6a and 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 that 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 (Figures 8b and 8b). In Experiment 1, a considerable proportion of both groups responds positively about understanding the feedback. However, the *control* group leans significantly more towards agreement. This might reflect higher cognitive load the CFE group, as they receive a more crowded, information-heavy screen. While in line with findings suggest-

ing that counterfactual style questions impose a larger cognitive load on participants [36], 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 Figures 8b and 8b are 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 [31]. While a very good starting point, it may be a worthwhile endeavor to perfect this measure in future user validations.

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 [36, 47, 61, 63], 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.

4.4.1 Limitations

Several limitations to this proof of concept study deserve discussion. While we clearly demonstrate the benefit of providing CFEs as feedback in an iterative learning design targeting an abstract domain for novice users, generalization of this observation to other tasks, domains and target groups is extremely limited[22, 60].

A cautionary note regards the efficacy of CFEs for human users more generally. 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 [35]. Thus, 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 [66]. 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.

Finally, 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 [19, 39, 54]. 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.

4.4.2 Conclusions

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.

5 FUTURE PERSPECTIVES AND CONCLUSIONS

The current paper introduces the Alien Zoo framework, developed to assess the usability of CFEs in XAI. 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.

User evaluations of XAI approaches are still in their infancy, leaving abundant room for studying various aspects. We believe that the Alien Zoo enables researchers to investigate a wide variety of different questions.

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 [34]. 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 [24]. The impact of such a framing in XAI is yet to be shown.

Moreover, further research should be done to 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 [28]. 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.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6 (2018), pp. 52138–52160 (cit. on pp. 1, 3, 4).

[2] Arjun Akula, Shuai Wang, and Song-Chun Zhu. "Cocox: Generating conceptual and counterfactual explanations via fault-lines". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 03. 2020, pp. 2594–2601 (cit. on p. 3).

[3] Leila Arras, Ahmed Osman, and Wojciech Samek. "CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations". In: *Information Fusion* 81 (2022), pp. 14–40 (cit. on p. 1).

[4] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115 (cit. on pp. 1, 3).

[5] André Artelt and Barbara Hammer. "On the computation of counterfactual explanations - A survey". In: *CoRR* abs/1911.07749 (2019). _eprint: 1911.07749. URL: http://arxiv.org/abs/1911.07749 (cit. on pp. 2, 8).

[6] André Artelt and Barbara Hammer. "Convex Density Constraints for Computing Plausible Counterfactual Explanations". en. In: *Artificial Neural Networks and Machine Learning – ICANN 2020*. Ed. by Igor Farkaš, Paolo Masulli, and Stefan Wermter. Vol. 12396. Cham: Springer International Publishing, 2020, pp. 353–365. ISBN: 978-3-030-61608-3. DOI: 10.1007/978-3-030-61609-0_28. URL: https://link.springer.com/10.1007/978-3-030-61609-0_28 (cit. on p. 2).

[7] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. "Evaluating Robustness of Counterfactual Explanations". In: (2021), pp. 01–09. DOI: 10.1109/SSCI50451.2021.9660058 (cit. on p. 1).

[8] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. "Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 2429–2437. ISSN: 2374-3468, 2159-5399. DOI: 10.1609/aaai.v33i01.33012429. URL: https://www.aaai.org/ojs/index.php/AAAI/article/view/4087 (visited on 11/16/2021) (cit. on p. 6).

[9] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. "Fitting Linear Mixed-Effects Models Using **lme4**". en. In: *Journal of Statistical Software* 67.1 (2015). ISSN: 1548-7660. DOI: 10.18637/jss.v067.i01. URL: http://www.jstatsoft.org/v67/i01/ (visited on 11/16/2021) (cit. on p. 7).

[10] Mattan Ben-Shachar, Daniel Lüdecke, and Dominique Makowski. "effectsize: Estimation of Effect Size Indices and Standardized Parameters". In: *Journal of Open Source Software* 5.56 (Dec. 2020), p. 2815. ISSN: 2475-9066. DOI: 10.21105/joss.02815. URL: https://joss.theoj.org/papers/10.21105/joss.02815 (visited on 11/16/2021) (cit. on p. 8).

[11] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification And Regression Trees*. en. 1st ed. London, UK: Routledge, 1984. ISBN: 978-1-315-13947-0. DOI: 10.1201/9781315139470. URL: https:

//www.taylorfrancis.com/books/9781351460491 (visited on 01/05/2022) (cit. on p. 8).

[12] Jacob T Browne. "Wizard of OZ prototyping for machine learning experiences". In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–6 (cit. on p. 5).

[13] Ruth M.J. Byrne. "Counterfactual Thought". en. In: *Annual Review of Psychology* 67.1 (Jan. 2016), pp. 135–157. ISSN: 0066-4308, 1545-2085. DOI: 10.1146/annurev-psych-122414-033249. URL: http://www.annualreviews.org/doi/10.1146/annurev-psych-122414-033249 (visited on 08/02/2019) (cit. on p. 2).

[14] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357 (cit. on p. 9).

[15] Michelene T. H. Chi and Ruth Wylie. "The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes". en. In: *Educational Psychologist* 49.4 (Oct. 2014), pp. 219–243. ISSN: 0046-1520, 1532-6985. DOI: 10.1080/00461520.2014.965823. URL: http://www.tandfonline.com/doi/abs/10.1080/00461520.2014.965823 (visited on 08/25/2021) (cit. on pp. 3, 4, 14).

[16] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications". In: *Information Fusion* 81 (2022), pp. 59–83 (cit. on p. 1).

[17] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. "Wizard of Oz studies—why and how". In: *Knowledge-based systems* 6.4 (1993), pp. 258–266 (cit. on p. 3).

[18] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. "Multi-objective counterfactual explanations". In: *International Conference on Parallel Problem Solving from Nature*. Springer. 2020, pp. 448–469 (cit. on p. 2).

[19] Brittany Davis, Maria Glenski, William Sealy, and Dustin Arendt. "Measure utility, gain trust: practical advice for XAI researchers". In: *2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. IEEE. 2020, pp. 1–8 (cit. on pp. 1, 3, 15).

[20] Michelle A. Detry and Yan Ma. "Analyzing Repeated Measurements Using Mixed Models". en. In: *JAMA* 315.4 (Jan. 2016), p. 407. ISSN: 0098-7484. DOI: 10.1001/jama.2015.19394. URL: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2015.19394 (visited on 10/14/2021) (cit. on p. 8).

[21] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. "Explaining models: an empirical study of how explanations impact fairness judgment". In: *Proceedings of the 24th international conference on intelligent user interfaces*. 2019, pp. 275–285 (cit. on p. 2).

[22] Finale Doshi-Velez and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning". en. In: *arXiv:1702.08608* (Feb. 2017). URL: http://arxiv.org/abs/1702.08608 (cit. on pp. 1–4, 15).

[23] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. "The who in explainable AI: how AI background shapes perceptions of AI explanations". In: *arXiv preprint arXiv:2107.13509* (2021) (cit. on p. 3).

[24] Kai Epstude and Neal J. Roese. "The Functional Theory of Counterfactual Thinking". In: *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc* 12.2 (May 2008), pp. 168–192. ISSN: 1088-8683. DOI: 10.1177/1088868308316091. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2408534/ (cit. on pp. 2, 16).

[25] European Union. "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)". In: *Official Journal of the European Union L110* 59 (2016), pp. 1–88 (cit. on p. 1).

[26] Stephen D. Goldinger, Heather M. Kleider, Tamiko Azuma, and Denise R. Beike. ""Blaming The Victim" Under Memory Load". en. In: *Psychological Science* 14.1 (Jan. 2003), pp. 81–85. ISSN: 0956-7976, 1467-9280. DOI: 10.1111/1467-9280.01423. URL: http://journals.sagepub.com/doi/10.1111/1467-9280.01423 (visited on 02/24/2020) (cit. on pp. 2, 4).

[27] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. "Local Rule-Based Explanations of Black Box Decision Systems". In: *arXiv:1805.10820 [cs]* (May 2018). URL: http://arxiv.org/abs/1805.10820 (visited on 08/16/2021) (cit. on p. 2).

[28] Fritz Heider. *The psychology of interpersonal relations*. New York, NY, US: John Wiley & Sons Ltd., 1958 (cit. on p. 16).

[29] Denis J. Hilton and Ben R. Slugoski. "Knowledge-based causal attribution: The abnormal conditions focus model." en. In: *Psychological Review* 93.1 (1986), pp. 75–88. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/0033-295X.93.1.75. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.93.1.75 (visited on 08/13/2021) (cit. on p. 2).

[30] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. "Metrics for explainable AI: Challenges and prospects". In: *arXiv preprint arXiv:1812.04608* (2018) (cit. on p. 3).

[31] Andreas Holzinger, André Carrington, and Heimo Müller. "Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations". en. In: *KI - Künstliche Intelligenz* 34.2 (Jan. 2020), pp. 193–198. ISSN: 0933-1875, 1610-1987. DOI: 10.1007/s13218-020-00636-z. URL: http://link.springer.com/10.1007/s13218-020-00636-z (visited on 05/25/2020) (cit. on pp. 7, 12, 14, 15).

[32] Sophie F Jentzsch, Sviatlana Höhn, and Nico Hochgeschwender. "Conversational interfaces for explainable AI: a human-centred approach". In: *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer. 2019, pp. 77–92 (cit. on p. 5).

[33] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. "If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques". In: *arXiv:2103.01035 [cs]* (Feb. 2021). URL: http://arxiv.org/abs/2103.01035 (cit. on pp. 1, 2).

[34] Ulrike Kuhl, André Artelt, and Barbara Hammer. "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), June 21–24, 2022, Seoul, Republic of Korea. 2022. DOI: 10.1145/3531146.3534630 (cit. on pp. 1, 5, 16).

[35] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. "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*. IEEE. 2013, pp. 3–10 (cit. on p. 15).

[36] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. "Human evaluation of models built for interpretability". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7. 2019, pp. 59–67 (cit. on pp. 3, 5, 15).

[37] Thai Le, Suhang Wang, and Dongwon Lee. "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*. 2020, pp. 238–248 (cit. on pp. 2, 3).

[38] Brian Y Lim, Anind K Dey, and Daniel Avrahami. "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*. 2009, pp. 2119–2128 (cit. on p. 3).

[39] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. "*Why* and why not explanations improve the intelligibility of context-aware intelligent systems". en. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Boston MA USA: ACM, Apr. 2009, pp. 2119–2128. ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1519023. URL: https://dl.acm.org/doi/10.1145/1518701.1519023 (cit. on pp. 13, 15).

[40] Peter Lipton. "Contrastive Explanation". en. In: *Royal Institute of Philosophy Supplement* 27 (Mar. 1990), pp. 247–266. ISSN: 1358-2461, 1755-3555. DOI: 10.1017/S1358246100005130. URL: https://www.cambridge.org/core/product/identifier/S1358246100005130/type/journal_article (visited on 08/13/2021) (cit. on p. 2).

[41] Gordon D Logan. "Shapes of Reaction-Time Distributions and Shapes of Learning Curves: A Test of the Instance Theory of Automaticity". en. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18.5 (1992), pp. 883–914 (cit. on p. 5).

[42] Tania Lombrozo. "Explanation and Abductive Inference". In: *The Oxford Handbook of Thinking and Reasoning*. Ed. by Keith J. Holyoak and Robert G. Morrison. Oxford, UK: Oxford University Press, Mar. 2012, pp. 260–276. DOI: 10.1093/oxfordhb/9780199734689.013.0014. URL: http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780199734689.001.0001/oxfordhb-9780199734689-e-14 (visited on 08/13/2021) (cit. on p. 2).

[43] Keith D. Markman and Matthew N. McMullen. "A Reflection and Evaluation Model of Comparative Thinking". en. In: *Personality and Social Psychology Review* 7.3 (Aug. 2003), pp. 244–267. ISSN: 1088-8683, 1532-7957. DOI: 10.1207/S15327957PSPR0703_04. URL: http://journals.sagepub.com/doi/10.1207/S15327957PSPR0703_04 (visited on 08/16/2021) (cit. on p. 2).

[44] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". en. In: *Artificial Intelligence* 267 (Feb. 2019), pp. 1–38. ISSN: 00043702. DOI: 10.1016/j.artint.2018.07.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0004370218305988 (visited on 11/30/2021) (cit. on p. 2).

[45] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. "A multidisciplinary survey and framework for design and evaluation of explainable AI systems". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11.3-4 (2021), pp. 1–45 (cit. on p. 3).

[46] Chelsea Muth, Karen L. Bales, Katie Hinde, Nicole Maninger, Sally P. Mendoza, and Emilio Ferrer. "Alternative Models for Small Samples in Psychological Research: Applying Linear Mixed Effects Models and Generalized Estimating Equations to Repeated Measures Data". en. In: *Educational and Psychological Measurement* 76.1 (Feb. 2016), pp. 64–87. ISSN: 0013-1644, 1552-3888. DOI: 10.1177/0013164415580432. URL: http://journals.sagepub.com/doi/10.1177/0013164415580432 (visited on 11/18/2021) (cit. on p. 8).

[47] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation". In: *arXiv preprint arXiv:1802.00682* (2018) (cit. on pp. 3, 5, 15).

[48] Fabian Offert. ""I know it when I see it". Visualization and Intuitive Interpretability". In: *arXiv:1711.08042 [stat]* (Dec. 2017). URL: http://arxiv.org/abs/1711.08042 (cit. on p. 2).

[49] Raphael Mazzine Barbosa de Oliveira and David Martens. "A Framework and Benchmarking Study for Counterfactual Generating Methods on Tabular Data". en. In: *Applied Sciences* 11.16 (Aug. 2021), p. 7274. ISSN: 2076-3417. DOI: 10.3390/app11167274. URL: https://www.mdpi.com/2076-3417/11/16/7274 (visited on 01/05/2022) (cit. on p. 1).

[50] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. "Carla: a python library to benchmark algorithmic recourse and counterfactual explanation algorithms". In: *arXiv preprint arXiv:2108.00783* (2021) (cit. on p. 1).

[51] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 5).

[52] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. "Manipulating and measuring model interpretability". In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–52 (cit. on p. 1).

[53] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021. URL: https://www.R-project.org/ (cit. on p. 7).

[54] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" 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*. 2016, pp. 1135–1144 (cit. on p. 15).

[55] Neal J. Roese. "Counterfactual thinking". en. In: *Psychological Bulletin* 121.1 (1997), pp. 133–148. ISSN: 1939-1455, 0033-2909. DOI: 10.1037/0033-2909.121.1.133. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.121.1.133 (visited on 08/13/2021) (cit. on pp. 2, 4).

[56] Neal J. Roese and Kai Epstude. "The Functional Theory of Counterfactual Thinking: New Evidence, New Challenges, New Insights". en. In: *Advances in Experimental Social Psychology*. Vol. 56. Amsterdam, Netherlands: Elsevier, 2017, pp. 1–79. ISBN: 978-0-12-812120-7. DOI:

10.1016/bs.aesp.2017.02.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0065260117300187 (visited on 08/16/2021) (cit. on p. 2).

[57] Lawrence J. Sanna and Kandi Jo Turley. "Antecedents to Spontaneous Counterfactual Thinking: Effects of Expectancy Violation and Outcome Valence". In: *Personality and Social Psychology Bulletin* 22.9 (Sept. 1996), pp. 906–919. ISSN: 0146-1672. DOI: 10.1177/0146167296229005. URL: https://doi.org/10.1177/0146167296229005 (visited on 02/24/2020) (cit. on pp. 2, 4).

[58] Sam Sattarzadeh, Mahesh Sudhakar, and Konstantinos N Plataniotis. "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*. 2021, pp. 4158–4167 (cit. on p. 1).

[59] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: from theory to algorithms*. New York, NY, USA: Cambridge University Press, 2014. ISBN: 978-1-107-05713-5 (cit. on p. 8).

[60] Kacper Sokol and Peter Flach. "Explainability fact sheets: a framework for systematic assessment of explainable approaches". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 56–67 (cit. on pp. 1, 3, 15).

[61] Kacper Sokol and Peter Flach. "One explanation does not fit all". In: *KI-Künstliche Intelligenz* 34.2 (2020), pp. 235–250 (cit. on pp. 3, 5, 15).

[62] Ilia Stepin, Alejandro Catala, Martin Pereira-Fariña, and Jose M. Alonso. "Paving the way towards counterfactual generation in argumentative conversational agents". en. In: *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*. Tokyo, Japan: Association for Computational Linguistics, 2019, pp. 20–25. DOI: 10.18653/v1/W19-8405. URL: https://www.aclweb.org/anthology/W19-8405 (visited on 01/21/2020) (cit. on p. 2).

[63] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. "Evaluating XAI: A comparison of rule-based and example-based explanations". en. In: *Artificial Intelligence* 291 (Feb. 2021), p. 103404. ISSN: 00043702. DOI: 10.1016/j.artint.2020.103404. URL: https://linkinghub.elsevier.com/retrieve/pii/S0004370220301533 (visited on 08/10/2021) (cit. on pp. 2–6, 13, 15).

[64] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. "Why a right to explanation of automated decision-making does not exist in the general data protection regulation". In: *International Data Privacy Law* 7.2 (2017), pp. 76–99 (cit. on p. 1).

[65] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31 (2017), p. 841 (cit. on p. 13).

[66] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. "Designing theory-driven user-centric explainable AI". In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–15 (cit. on p. 15).

[67] Adam White and Artur d'Avila Garcez. "Measurable Counterfactual Local Explanations for Any Classifier". en. In: *ECAI*. Santiago de Compostela, Spain, 2020, p. 7 (cit. on p. 1).

SUPPLEMENTARY MATERIAL

A FULL LIST OF SURVEY ITEMS

* marks the catch item in place to evaluate if users are still paying attention.

Item No.	Control group	CFE group	Response options
1	What do you think: Which plants were relevant to increase the number of Shubs in your pack? Please select ALL that you think were relevant.		5 checkboxes, together with icons of the available plants + option "I do not know."
2	What do you think: Which plants were not relevant to increase the number of Shubs in your pack? Please select ALL that you think were not relevant.		
3	I understood the overview of my past choices.	I understood the feedback on what choice would have led to a better result.	5 point Likert-scale, checkboxes with options: Strongly disagree - disagree - neutral - agree - strongly agree + option "I prefer not to answer."
4	I needed support to understand the overview of my past choices.	I needed support to understand the feedback on what choice would have led to a better result.	
5	I found that the overview of my past choices helped me to increase the number of Shubs.	I found that the feedback on what choice would have led to a better result helped me to increase the number of Shubs.	
6	I was able to use the overview of my past choices to increase the number of Shubs.	I was able to use the feedback on what choice would have led to a better result to increase the number of Shubs.	
7*	Are you still paying attention? If so, please select 'I prefer not to answer' for this question.		
8	I found inconsistencies in the overview of my past choices.	I found inconsistencies in the feedback on what choice would have led to a better result.	
9	I think most people would learn to work with the overview of their past choices very quickly.	I think most people would learn to work with the feedback on what choice would have led to a better result very quickly.	
10	I received the overview of my past choices in a timely and efficient manner.	I received the feedback on what choice would have led to a better result in a timely and efficient manner.	
Age	Please indicate your age.		Checkboxes with options: 18-24y, 25-34y, 35-44y, 45-54y, 55-64y, 65y and over
Gender	Which term most accurately describes your gender?		Checkboxes with options: Female, Male, Transgender female, Transgender male, Non-binary / gender non-conforming, Not listed, I prefer not to answer

B EXEMPLARY USER JOURNEY THROUGH THE FIRST BLOCK OF THE GAME PHASE

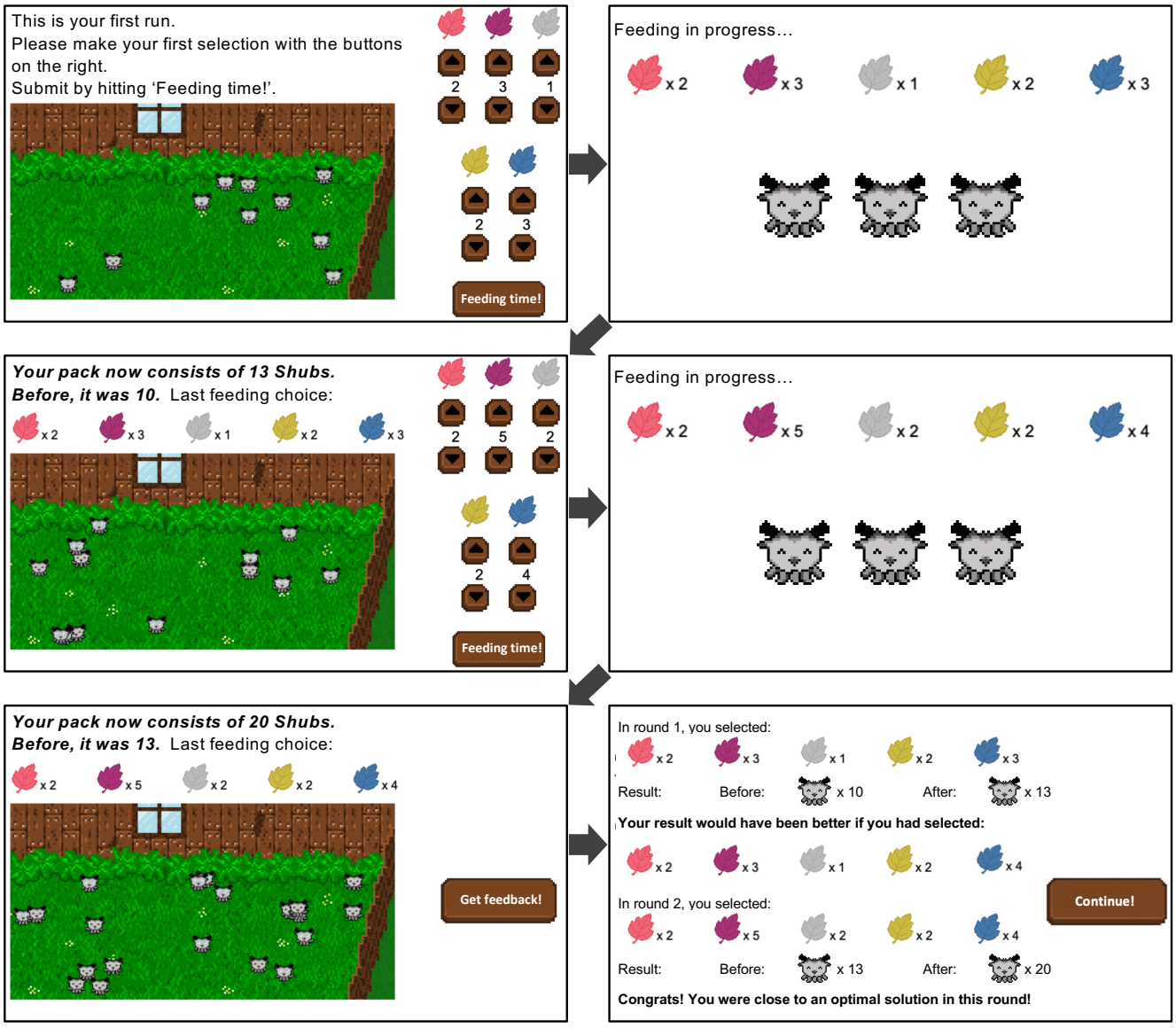
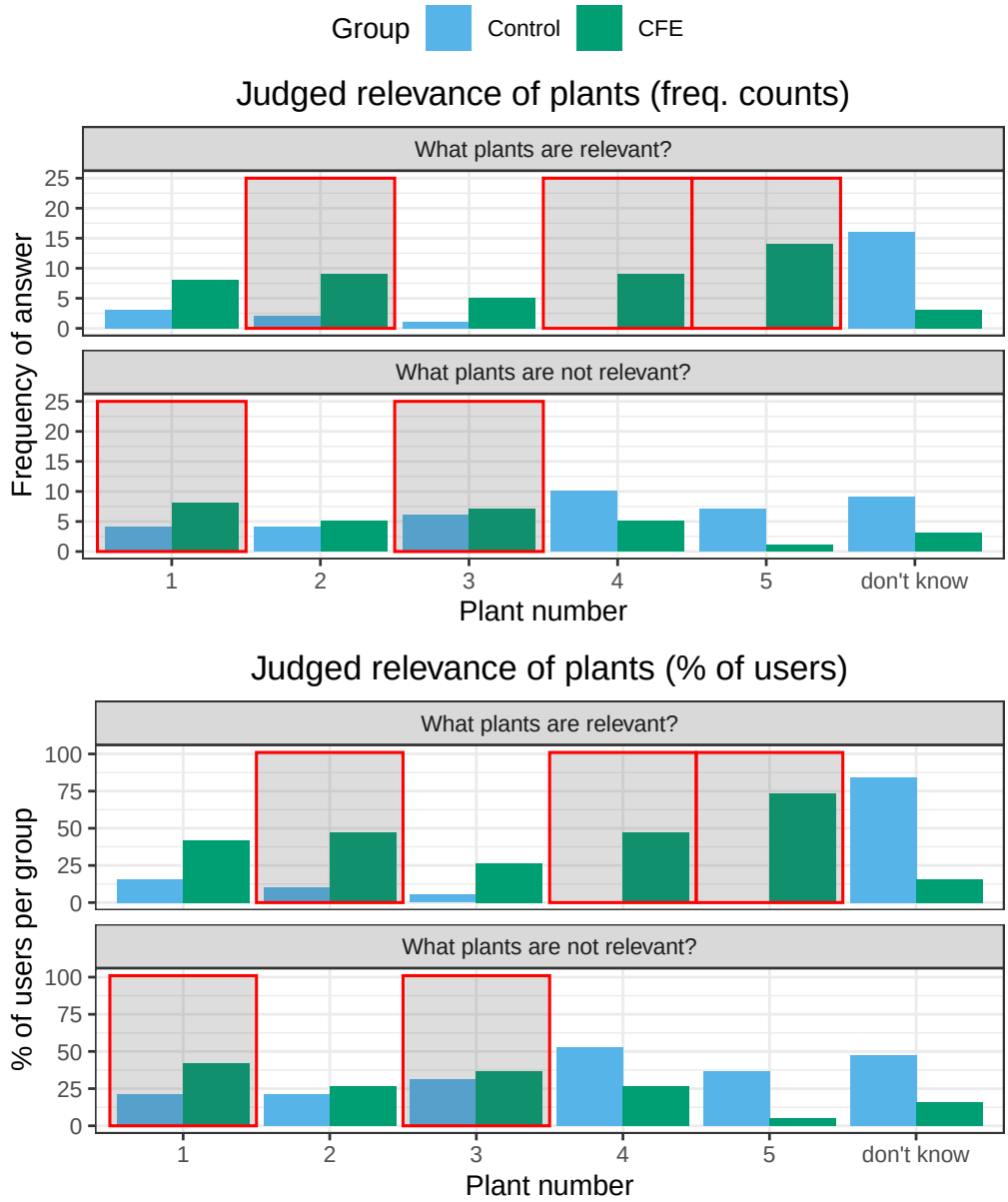
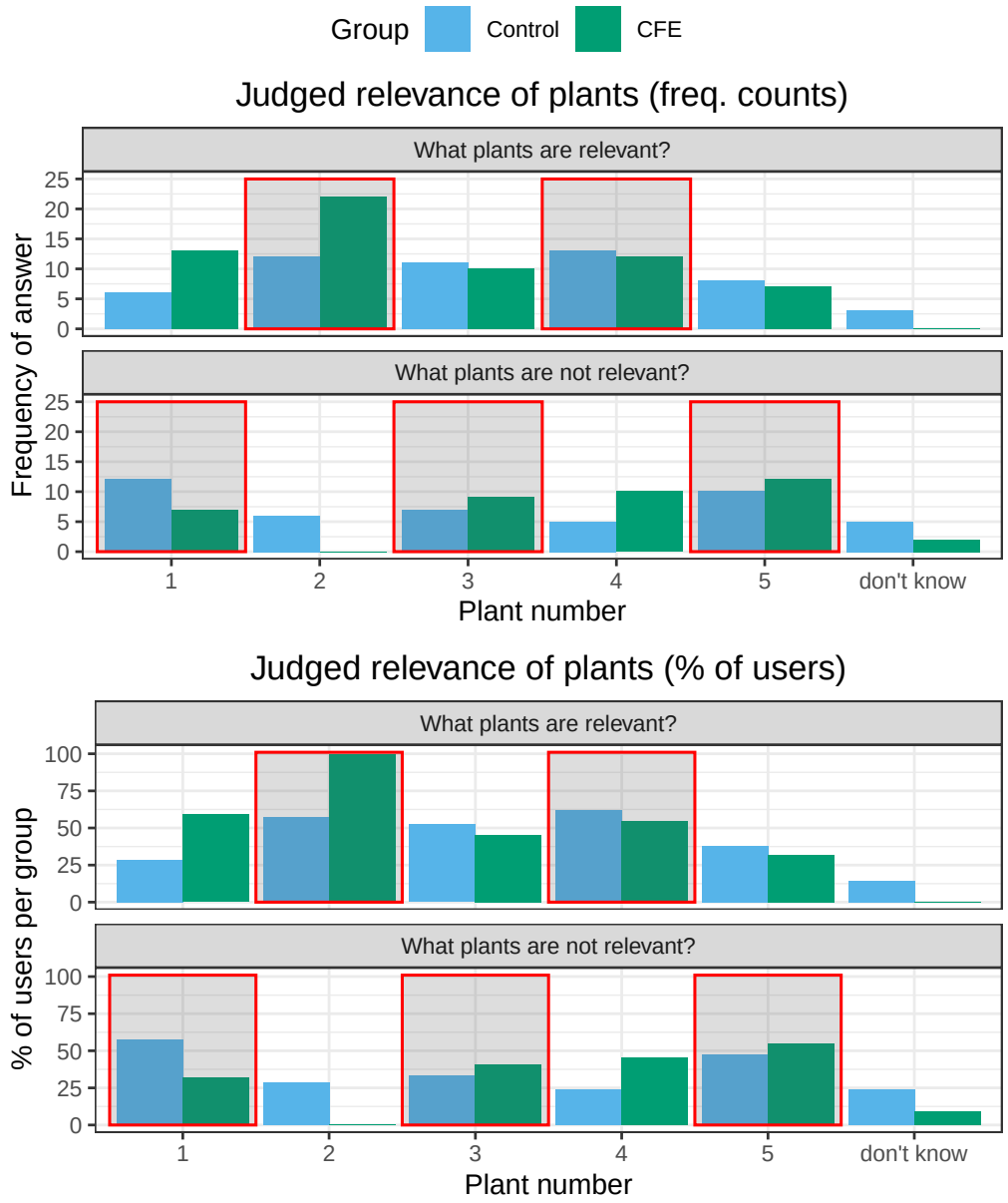


Figure 9: Exemplary user journey through the first block of the Alien Zoo game. Bold arrows indicate temporal succession of respective scenes. The figure highlights the iterative nature of the game with repeated user input and end-of-block presentation of CFEs (experimental group), or overview of past choices (control group). Note that plant counters are set to 0 at the beginning of each padlock scene. The figure displays the state after the exemplary user inserted their current choice. For this manuscript, font size in images of scenes was increased to improve visibility.

C DETAILED FEATURE RELEVANCE JUDGEMENTS



Experiment 1: Detailed relevance judgements per plant, as frequency of users (top), and percentage of users (bottom), respectively. Red boxes indicate relevant and non-relevant features.



Experiment 2: Detailed relevance judgements per plant, as frequency of users (top), and percentage of users (bottom), respectively. Red boxes indicate relevant and non-relevant features.